Topics in Current Chemistry 345

Şule Atahan-Evrenk
Alán Aspuru-Guzik *Editors*

Prediction and Calculation of Crystal Structures

Methods and Applications

Springer

345
Topics in Current Chemistry

Editorial Board:

K.N. Houk, Los Angeles, CA, USA
C.A. Hunter, Sheffield, UK
M.J. Krische, Austin, TX, USA
J.-M. Lehn, Strasbourg, France
S.V. Ley, Cambridge, UK
M. Olivucci, Siena, Italy
J. Thiem, Hamburg, Germany
M. Venturi, Bologna, Italy
C.-H. Wong, Taipei, Taiwan
H.N.C. Wong, Shatin, Hong Kong

For further volumes:
http://www.springer.com/series/128

Aims and Scope

The series Topics in Current Chemistry presents critical reviews of the present and future trends in modern chemical research. The scope of coverage includes all areas of chemical science including the interfaces with related disciplines such as biology, medicine and materials science.

The goal of each thematic volume is to give the non-specialist reader, whether at the university or in industry, a comprehensive overview of an area where new insights are emerging that are of interest to larger scientific audience.

Thus each review within the volume critically surveys one aspect of that topic and places it within the context of the volume as a whole. The most significant developments of the last 5 to 10 years should be presented. A description of the laboratory procedures involved is often useful to the reader. The coverage should not be exhaustive in data, but should rather be conceptual, concentrating on the methodological thinking that will allow the non-specialist reader to understand the information presented.

Discussion of possible future research directions in the area is welcome.

Review articles for the individual volumes are invited by the volume editors.

Readership: research chemists at universities or in industry, graduate students.

Şule Atahan-Evrenk • Alán Aspuru-Guzik
Editors

Prediction and Calculation of Crystal Structures

Methods and Applications

With contributions by

C.S. Adjiman · A. Aspuru-Guzik · S. Atahan-Evrenk · G.J.O. Beran · J.G. Brandenburg · S. Grimme · G. Hautier · Y. Heit · R.G. Hennig · Y. Huang · A.V. Kazantsev · K. Nanda · A.R. Oganov · C.C. Pantelides · B.C. Revard · R.Q. Snurr · W.W. Tipton · S. Wen · C.E. Wilmer · X.-F. Zhou · Q. Zhu

Springer

Editors
Şule Atahan-Evrenk
Alán Aspuru-Guzik
Dept. of Chemistry and Chemical Biology
Harvard University
Cambridge
Massachusetts
USA

ISSN 0340-1022 ISSN 1436-5049 (electronic)
ISBN 978-3-319-05773-6 ISBN 978-3-319-05774-3 (eBook)
DOI 10.1007/978-3-319-05774-3
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014938743

© Springer International Publishing Switzerland 2014
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The prediction of crystal structure for a chemical compound is still a challenge. It requires advanced algorithms for exhaustive searches of the possible packing forms and highly accurate computational methodologies to rank the possible crystal structures. This book presents some of the important developments in crystal structure prediction in recent years. The chapters do not cover every area but rather present a wide range of methodologies with applications in organic, inorganic, and hybrid compounds.

The blind tests organized by the Cambridge Crystallographic Data Center (CCDC) showed a notable improvement for the crystal structure prediction of organic compounds over recent years. The first two chapters of this book present two of the methodologies contributed to the success in recent blind tests. The chapter "Dispersion Corrected Hartree–Fock and Density Functional Theory for Organic Crystal Structure Prediction" by Brandenburg and Grimme is dedicated to recent advances in the dispersion-corrected Hartree–Fock and density functional theory. Another important area showing remarkable progress is the efficient treatment of the internal flexibility of molecules with many rotatable bonds. The chapter "General Computational Algorithms for Ab Initio Crystal Structure Prediction for Organic Molecules" by Pantelides et al. summarizes some of the algorithms that have contributed to this success. In addition, the chapter "Accurate and Robust Molecular Crystal Predictions Using Fragment-Based Electronic Structure Methods" by Beran et al. illustrates how fragment-based electronic structure methods can provide accurate prediction of the lattice energy differences of polymorphs of organic compounds.

One research area that would benefit tremendously from the crystal structure prediction of organic compounds is the design of organic semiconductors. In the chapter "Prediction and Theoretical Characterization of Organic Semiconductor Crystals for Field-Effect Transistor Applications" by Şule Atahan-Evrenk and Alán Aspuru-Guzik, discuss some aspects of theoretical characterization and prediction of crystal structures of p-type organic semiconductors for organic transistor applications. The chapter also provides information about the structure–property relationships in organic semiconductors.

In organic systems, thanks to the internal constraints of molecular structures, random sampling methods can be used successfully. In inorganic crystals, however, there are no constraints other than the chemical compositions. Therefore, the challenge in the crystal structure prediction of inorganic compounds is the search problem, and the methodologies that span the search space effectively are crucial. The chapters by Hautier, by Revard et al., and by Zhu et al. are dedicated to cover recent advances towards achieving inorganic crystal prediction. The chapter "Data Mining Approaches to High-Throughput Crystal Structure and Compound Prediction" by Hautier discusses data mining approaches and the chapters by Revard et al. and by Zhu et al. cover evolutionary algorithms for compound prediction. In particular, the chapter "Structure and Stability Prediction of Compounds with Evolutionary Algorithms" by Revard et al. presents different methodologies adapted for the evolutionary algorithms approaches and the chapter "Crystal Structure Prediction and Its Application in Earth and Materials Sciences" by Zhua et al. focuses on the state of the art of the USPEX methodology.

The prediction of hybrid materials such as metal-organic frameworks posits a specific set of challenges for structure prediction. The chapter "Large-Scale Generation and Screening of Hypothetical Metal-Organic Frameworks for Applications in Gas Storage and Separation" by Wilmer and Snurr discusses the large-scale generation and screening of metal-organic frameworks. With possible applications in storage, catalysis, pharmaceuticals, and electrochemistry, these methodologies show great potential for development of hybrid systems.

We believe crystal structure prediction will be one of the most important tools in solid-state chemistry in the near future. Applications ranging from pharmaceuticals to energy technologies would benefit tremendously from computational prediction of the solid forms of materials. We believe this book provides up-to-date, concise, and accessible coverage of the subject for a wide audience in academia and industry and we hope that it will be useful for chemists and materials scientists who want to learn more about the state-of-the-art in crystal structure prediction methods and applications.

We would like to thank Springer editors Birke Dalia and Elizabeth Hawkins for inviting us to edit this volume and all the authors for their contributions. Lastly, we would like to thank all the members of the Aspuru-Guzik Group for their support and camaraderie.

Cambridge, MA, USA
December 2013

Şule Atahan-Evrenk and Alán Aspuru-Guzik

Contents

Top Curr Chem (2014) 345: 1–24
DOI: 10.1007/128_2013_488
© Springer-Verlag Berlin Heidelberg 2013
Published online: 13 November 2013

Dispersion Corrected Hartree–Fock and Density Functional Theory for Organic Crystal Structure Prediction

Jan Gerit Brandenburg and Stefan Grimme

Abstract We present and evaluate dispersion corrected Hartree–Fock (HF) and Density Functional Theory (DFT) based quantum chemical methods for organic crystal structure prediction. The necessity of correcting for missing long-range electron correlation, also known as van der Waals (vdW) interaction, is pointed out and some methodological issues such as inclusion of three-body dispersion terms are discussed. One of the most efficient and widely used methods is the semi-classical dispersion correction D3. Its applicability for the calculation of sublimation energies is investigated for the benchmark set X23 consisting of 23 small organic crystals. For PBE-D3 the mean absolute deviation (MAD) is below the estimated experimental uncertainty of 1.3 kcal/mol. For two larger π-systems, the equilibrium crystal geometry is investigated and very good agreement with experimental data is found. Since these calculations are carried out with huge plane-wave basis sets they are rather time consuming and routinely applicable only to systems with less than about 200 atoms in the unit cell. Aiming at crystal structure prediction, which involves screening of many structures, a pre-sorting with faster methods is mandatory. Small, atom-centered basis sets can speed up the computation significantly but they suffer greatly from basis set errors. We present the recently developed geometrical counterpoise correction gCP. It is a fast semi-empirical method which corrects for most of the inter- and intramolecular basis set superposition error. For HF calculations with nearly minimal basis sets, we additionally correct for short-range basis incompleteness. We combine all three terms in the HF-3c denoted scheme which performs very well for the X23 sublimation energies with an MAD of only 1.5 kcal/mol, which is close to the huge basis set DFT-D3 result.

Keywords Counterpoise correction · Crystal structure prediction · Density Functional Theory · Dispersion correction · Hartree–Fock

J.G. Brandenburg and S. Grimme (✉)
Mulliken Center for Theoretical Chemistry, Institut für Physikalische und Theoretische Chemie der Universität Bonn, Beringstraße 4, 53115 Bonn, Germany
e-mail: gerit.brandenburg@thch.unibonn.de; grimme@thch.uni-bonn.de

Contents

Abbreviations

ANCOPT	Approximate normal coordinate rational function optimization program
AO	Gaussian atomic orbitals
B3LYP	Combination of Becke's three-parameter hybrid functional B3 and the correlation functional LYP of Lee, Yang, and Parr
BSE	Basis set error
BSIE	Basis set incompleteness error
BSSE	Basis set superposition error
CN	Coordination number
CRYSTAL09	Crystalline orbital program
D3	Third version of a semi-classical *first-principles* dispersion correction
DF	Density functional
DFT	Density Functional Theory
DFT-D3	Density Functional Theory with atom-pairwise and three-body dispersion correction
gCP	Geometrical counterpoise correction
GGA	Generalized gradient approximation
HF	Hartree–Fock
HF-3c	Dispersion corrected Hartree–Fock with semi-empirical basis set corrections
MAD	Mean absolute deviation
MBD	Many-body dispersion interaction by Tkatchenko and Scheffler
MD	Mean deviation
Me-TBTQ	Centro-methyl tribenzotriquinazene
MINIX	Combination of polarized minimal basis and SVP basis
PAW	Projector augmented plane-wave
PBE	Generalized gradient-approximated functional of Perdew, Burke, and Ernzerhof
RMSD	Root mean square deviation
RPA	Random phase approximation

RPBE	Revised version of the PBE functional
SAPT	Symmetry Adapted Perturbation Theory
SCF	Self-consistent field
SD	Standard deviation
SIE	Self interaction error
SRB	Short-range basis incompleteness correction
SVP	Polarized split-valence basis set of Ahlrichs
TBTQ	Tribenzotriquinazene
TS	Tkatchenko and Scheffler dispersion correction
VASP	Vienna ab initio simulation package
vdW	Van der Waals
VV10	Vydrov and van Voorhis non-local correlation functional
X23	Benchmark set of 23 small organic crystals
XDM	Exchange-dipole model of Becke and Johnson
ZPV	Zero point vibrational energy

1 Introduction

Aiming at organic crystal structure prediction, two competing requirements for the utilized theoretical method exist. On the one hand, the calculation of crystal energies has to be accurate enough to distinguish between different polymorphs. This involves an accurate account of inter- as well as intramolecular interactions in various geometrical situations. On the other hand, each single computation (energy including the corresponding derivatives for geometry optimization or frequency calculation) has to be fast enough to sample all space groups under consideration (and possibly different molecular conformations) in a reasonable time [1–5]. Typically, one presorts the systems with a fast method and investigates the energetically lowest ones with a more accurate (but more costly) method. For the inclusion of zero point vibrational energy (ZPVE) contributions a medium quality level is often sufficient. A corresponding algorithm is sketched in Fig. 1. The generation of the initial structure (denoted as sample space groups) is an important issue, but will not be discussed in this chapter. Here we focus on the different electronic structure calculations, denoted by the quadratic framed steps in Fig. 1. We present dispersion corrected Density Functional Theory (DFT-D3) as a possible high-quality method with medium computational cost and dispersion corrected Hartree–Fock (HF) with semi-empirical basis error corrections (HF-3c) as a faster method with medium quality.

Density Functional Theory (DFT) is the "work horse" for many applications in chemistry and physics and still an active research field of general interest [6–9]. In many covalently bound (periodic and non-periodic) systems, DFT provides a very good compromise between accuracy and computational cost. However, common generalized gradient approximated (GGA) functionals are not capable of describing long-range electron correlation, a.k.a. the London dispersion interaction [10–13]. This dispersion term can be empirically defined as the attractive part of

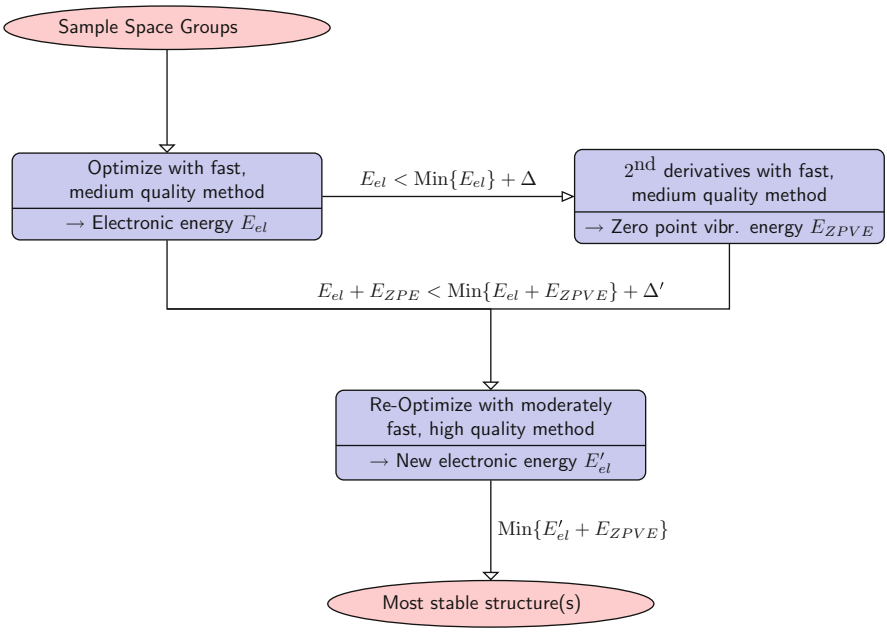


Fig. 1 A typical crystal structure prediction algorithm [1]. First, the optimum electronic crystal energy E_{el} is calculated with a fast, medium quality method. Second, the more costly second derivatives for the electronically lowest structures in a certain energy interval (Δ) are calculated to get the zero point vibrational energy E_{ZPVE}. Finally, the electronic energy E'_{el} is re-calculated for the energetically lowest structures in a (different) energy interval (Δ') with a more accurate method. The data from step two can be finally used also to estimate thermal and entropic corrections

the van der Waals-type interaction between atoms and molecules that are not directly bonded to each other. For the physically correct description of molecular crystals, dispersion interactions are crucial [14, 15]. In the last decade, several well-established methods for including dispersion interactions into DFT were developed. For an overview and reviews of the different approaches, see, e.g., [16–25] and references therein. Virtual orbital dependent (e.g., random phase approximation, RPA [26]) and fragment based (e.g., symmetry adapted perturbation theory, SAPT [27]) methods are not discussed further here because they are currently not routinely applicable to larger molecular crystals. For the alternative combination of accurate molecular quantum chemistry calculations for crystal fragments with force-fields and subsequent periodic extension see, e.g., [28, 29].

Here we focus on the atom-pairwise dispersion correction D3 [30, 31] coupled with periodic electronic structure theory. The D3 scheme incorporates non-empirical, chemical environment-dependent dispersion coefficients, and for dense systems a non-additive Axilrod–Teller–Muto three-body dispersion term. We present the details of this method in Sect. 2.1. Compared to the self-consistent

solution of the Kohn–Sham (KS) or HF equations, the calculation of the D3 dispersion energy requires practically no additional computation time. Although it does not include information about the electron density, it provides good accuracy with typical deviations for the asymptotic dispersion energy of only 5% [19]. The accuracy for non-covalent interaction energies with current standard functionals and D3 is about 5–10%, which is also true for small relative energies [32]. Therefore, it is an ideal tool to fulfill fundamental requirements of crystal structure prediction. We evaluate the DFT-D3 scheme with huge plane-wave basis sets in Sect. 2.2 and compare it to competing pairwise-additive methods, which partially employ electron density information.

Because the calculation of the DFT or HF energy is the computational bottleneck, a speed-up of these calculations without losing too much accuracy is highly desirable. The computational costs mainly depend on the number of utilized single particle basis functions N with a typical scaling behavior from N^2 to N^4. The choice of the type of basis functions is also an important issue. Bulk metals have a strongly delocalized valence electron density and plane-wave based basis sets are probably the best choice [33]. In molecular crystals, however, the charge density is more localized and a typical molecular crystal involves a lot of "vacuum." For plane-wave based methods this can result in large and inefficient basis sets. In a recently studied typical organic system (tribenzotriquinacene, $C_{22}H_{16}$), up to 1.5×10^5 projector augmented plane-wave (PAW) basis functions must be considered for reasonable basis set convergence [34]. For this kind of system, atom-centered Gaussian basis functions as usually employed in molecular quantum chemistry could be more efficient. However, small atom-centered basis sets strongly suffer from basis set errors (BSE), especially the basis set superposition error (BSSE) which leads to overbinding and too high computed weight densities (too small crystal volumes) in unconstrained optimizations. Because different polymorphs often show various packings with different densities, correcting for BSSE is mandatory in our context. In order to get reasonable absolute sublimation energies and good crystal geometries, these basis set errors must be corrected. A further problem compared to plane-wave basis sets is the non-orthogonality of atom-centered basis functions which can lead to near-linear dependencies and bad self-consistent field (SCF) convergence. We have recently mapped the standard Boys and Bernardi correction [35], which corrects for the BSSE, onto an atom-pairwise repulsive potential. It was fitted for a number of typical Gaussian basis sets and depends otherwise only on the system geometry and is therefore denoted gCP [36]. Analytic gradients are problematic in nearly all other counterpoise schemes, but are easily obtained for gCP. For the calculation of second derivatives, analytic first derivatives are particularly crucial. Periodic boundary conditions are included and the implementation has been tested in [37].We present the gCP scheme here together with an additional short-range basis (SRB) incompleteness correction in Sect. 3.1. In Sect. 3.2 the combination of small (almost minimal) basis set DFT and HF, dispersion correction D3, geometrical counterpoise correction gCP, and short-range incompleteness correction SRB is evaluated for typical molecular crystals. The plane-wave, large basis PBE-D3 results are briefly discussed and used for comparison.

2 Dispersion Corrected Density Functional Theory

2.1 London Dispersion Correction

At short inter-atomic distances, standard density functionals (DF) describe the effective electron interaction rather well because of their deep relation to the corresponding electron density changes. Long-range electron correlation cannot be accurately described by the local (or semi-local) DFs in inhomogeneous materials. To describe this van der Waals (vdW)-type interaction, one can include non-local kernels in the vdW-DFs as pioneered by Langreth and Lundquist [38, 39] and later improved by Vydrov and van Voorhis (VV10 [25]). For the total exchange-correlation energy E_{xc} of a system, the following approximation is employed in all vdW-DF schemes:

$$E_{\mathrm{xc}}=E_{\mathrm{X}}^{\mathrm{GGA}}+E_{\mathrm{C}}^{\mathrm{GGA}}+E_{\mathrm{c}}^{\mathrm{NL}}, \tag{1}$$

where standard exchange (X) and correlation (C) components (in the semi-local generalized gradient approximation GGA) are used for the short-range parts and $E_{\mathrm{c}}^{\mathrm{NL}}$ represents the non-local correlation term describing the dispersion energy. In the vdW-DF framework it takes the form of a double-space integral:

$$E_{\mathrm{c}}^{\mathrm{NL}}=\frac{1}{2}\int\int\rho(\mathbf{r})\Phi^{\mathrm{NL}}\left(\mathbf{r},\mathbf{r}'\right)\rho\left(\mathbf{r}'\right)d^3rd^3r'. \tag{2}$$

The electron density ρ at positions $\mathbf{r}$ and $\mathbf{r}'$ is correlated via the integration kernel $\Phi^{\mathrm{NL}}(\mathbf{r},\mathbf{r}')$. It is physically approximated by local approximations to the frequency dependent dipole polarizability $\boldsymbol{\alpha}(\mathbf{r},\boldsymbol{\omega})$. The VV10 kernel has been successfully used in various molecular applications [40–43] by us but is not discussed further in this work.

The famous Casimir–Polder relationship [44] connects the polarizability with the long-range dispersion energy, which scales as $C_6 = R^6$ where R is the distance between two atoms or molecules. The corresponding dispersion coefficient $\mathrm{C}_6^{\mathrm{AB}}$ for interacting fragments A and B is given by

$$\mathrm{C}_6^{\mathrm{AB}}=\frac{3}{\pi}\int_0^{\infty}\boldsymbol{\alpha}^{\mathrm{A}}(i\boldsymbol{\omega})\boldsymbol{\alpha}^{\mathrm{B}}(i\boldsymbol{\omega})d\boldsymbol{\omega}, \tag{3}$$

where $\boldsymbol{\alpha}^{A}(i\boldsymbol{\omega})$ is the averaged dipole polarizability at imaginary frequency $\boldsymbol{\omega}$. In vdW-DF (but not in DFT-D3) dispersion can be calculated self-consistently and changes the density in turn. Because this change is normally insignificant [25, 38, 40], $E_{\mathrm{c}}^{\mathrm{NL}}$ is typically added non-self-consistently to the SCF-GGA energy. The main advantage of vdW-DF methods is that dispersion effects are naturally included via the system electron density. Therefore, they implicitly account for changes in the

dispersion coefficients due to different "atoms-in-molecules" oxidation states in a physically sound manner. The disadvantage is the raised computational cost compared to pure (semi-)local DFs.

By treating the short-range part with DFs and the dispersion interaction with a semi-classical atom-pairwise correction, one can combine the advantages of both worlds. Semi-classical models for the dispersion interaction like D3 show very good accuracy compared to, e.g., the VV10 functional [43, 45] for very little computational overheads, particularly when analytical gradients are required.

The total energy E_{tot} of a system can be decomposed into the standard, dispersion-uncorrected DFT/HF electronic energy $E_{\mathrm{DFT/HF}}$ and the dispersion energy E_{disp}:

$$E_{\mathrm{tot}} = E_{\mathrm{DFT/HF}} + E_{\mathrm{disp}}. \tag{4}$$

We use our latest first-principles type dispersion correction DFT-D3, where the dispersion coefficients are non-empirically obtained from a time-dependent, linear response DFT calculation of $\boldsymbol{\alpha}^A(i\boldsymbol{\omega})$. The dispersion energy can be split into two- and three-body contributions $E_{\mathrm{disp}} = E^{(2)} + E^{(3)}$:

$$E^{(2)} = -\frac{1}{2}\sum_{n=6,8}\sum_{\mathrm{A}\neq\mathrm{B}}^{\text{atom pairs}}\sum_{\mathbf{T}} s_n \frac{C_{\mathrm{n}}^{\mathrm{AB}}}{\|\mathbf{r}_{\mathrm{B}} - \mathbf{r}_{\mathrm{A}} + \mathbf{T}\| + f\left(R_0^{\mathrm{AB}}\right)^n} \tag{5}$$

$$E^{(3)} = \frac{1}{6}\sum_{\mathrm{A}\neq\mathrm{B}}^{\text{atom pairs}}\sum_{\mathbf{T}} \frac{C_9^{\mathrm{ABC}}(3\cos\boldsymbol{\theta}_{\mathrm{a}}\cos\boldsymbol{\theta}_{\mathrm{b}}\cos\boldsymbol{\theta}_{\mathrm{c}} + 1)}{r_{\mathrm{ABC}}^9 \cdot (1 + 6(r_{\mathrm{ABC}}/R_0) - \alpha)}. \tag{6}$$

Here, $C_{\mathrm{n}}^{\mathrm{AB}}$ denotes the averaged (isotropic) nth-order dispersion coefficient for atom pair AB, and $\mathbf{R}_{\mathrm{A/B}}$ are their Cartesian positions. The real-space summation over all unit cells is done by considering all translation invariant vectors $\mathbf{T}$ inside a cut-off sphere. The scaling parameter s6 equals unity for the DFs employed here and ensures the correct limit for large interatomic distances, and s8 is a functional-dependent scaling factor. The rational Becke and Johnson damping function $f(R_0^{\mathrm{ab}})$ is [46]

$$f\left(R_0^{\mathrm{ab}}\right) = a_1 R_0^{\mathrm{ab}} + a_2,\ R_0^{\mathrm{ab}} = \sqrt{\frac{C_8^{\mathrm{ab}}}{C_6^{\mathrm{ab}}}}. \tag{7}$$

The dispersion coefficients C_6^{AB} are computed for molecular systems with the Casimir–Polder relation (3).We use the concept of fractional coordination numbers (CN) to distinguish the different hybridization states of atoms in molecules in a differentiable way. The CN is computed from the coordinates and does not use information from the electronic wavefunction or density but recovers basic information about the bonding situation of an atom in a molecule, which has a dominant influence on the C_6^{AB} coefficients [30]. The higher order C_8 coefficients are obtained from the well-known relation [47]

$$C_8 = \frac{3}{2} C_6 \frac{\langle r^4 \rangle}{\langle r^2 \rangle}. \tag{8}$$

With the recursion relation $C_{i+4} = C_{i\text{-}2}\left(\frac{C_{i+2}}{C_i}\right)$ and $C_{10} = \frac{49}{40}\frac{C_8^2}{C_6}$, one can in principle also generate higher orders, but terms above C_{10} do not improve the performance of the D3 method. The three parameters s_8, a_1, and a_2 are fitted for each DF on a benchmark set of small, non-covalently bound complexes. This fitting is necessary to prevent double counting of dispersion interactions at short range and to interpolate smoothly between short- and long-range regimes. These parameters are successfully applied to large molecular complexes and to periodic systems [45, 48]. In the non-additive Axilrod–Teller–Muto three-body contribution (6) [30, 49], r_{ABC} is an average distance in the atom-triples and $\boldsymbol{\theta}_{\mathrm{a/b/c}}$ are the corresponding angles. The dispersion coefficient C_9^{ABC} describes the interaction between three virtually interacting dipoles and is approximated from the pairwise coefficients as

$$C_9^{\mathrm{ABC}} = -\sqrt{C_6^{\mathrm{AB}} C_6^{\mathrm{AC}} C_6^{\mathrm{BC}}}. \tag{9}$$

The applicability of this atom-pairwise dispersion correction with three-body corrections in dense molecular systems was shown in a number of recent publications [16, 50, 51].

For early precursors of DFT-D3 also in the framework of HF theory, see [52–56]. Related to the D3 scheme are approaches that also compute the C_6 coefficients specific for each atom (or atom pair) and use a functional form similar to (5). A system dependency of the dispersion coefficients is employed by all modern DFT-D variants. We explicitly mention the works of Tkatchenko and Scheffler [57, 58] (TS. "atom-in-molecules" C_6 from scaled atomic volumes), Sato et al. [59] (use of a local atomic response function), and Becke and Johnson [46, 60, 61] (XDM utilizes a dipole-exchange hole model). The TS and XDM methods are used routinely in solid-state applications [62–65].

2.2 *Evaluation of Dispersion Corrected DFT*

2.2.1 X23 Benchmark Set

A benchmark set for non-covalent interactions in solids consisting of 21 molecular crystals (dubbed C21) was compiled by Johnson [24]. Two properties for benchmarking are provided: (1) thermodynamically back-corrected experimental sublimation energies and (2) geometries from low-temperature X-ray diffraction. The error of the experimental sublimation energies was estimated to be 1.2 kcal/mol [66]. Recently, the C21 set was extended and refined by Tkatchenko et al. [67]. The X23 benchmark set (16 systems from [67] and data for 7 additional systems were

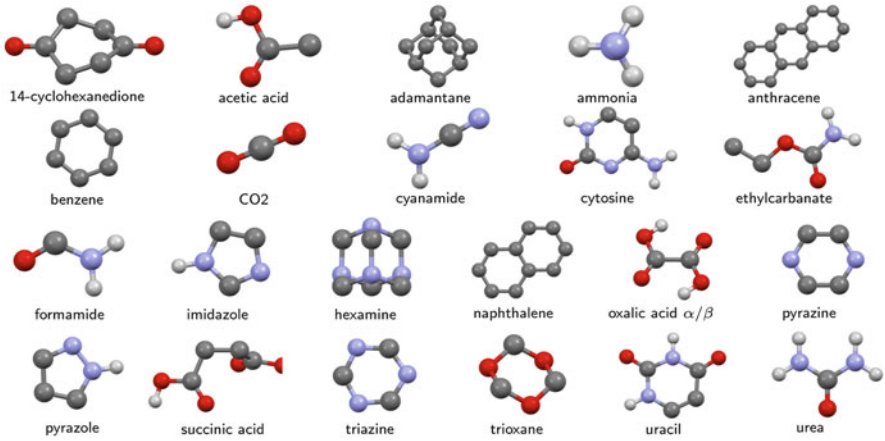


Fig. 2 Geometries of the 23 small organic molecules in the X23 benchmark set for non-covalent interactions in solids. Hydrogen atoms at carbons are omitted for clarity. Carbons are denoted by *dark gray balls*, hydrogens are *light gray*, oxygens are *red*, and nitrogens are *light blue*

obtained from these authors) includes two additional molecular crystals, namely hexamine and succinic acid. The molecular geometries of the X23 set are shown in Fig. 2. The thermodynamic back-correction was consistently done at the PBE-TS level. Semi-anharmonic frequency corrections were estimated by solid state heat capacity data. Further details of the back-correction scheme are summarized in [67] The mean absolute deviation (MAD) between both data sets is 0.55 kcal/mol. Because the X23 data seem to be more consistent, we use these as a reference. If we take the standard deviation (SD) between both thermodynamic corrections as statistical error measure, the total uncertainty of the reference values is about 1.3 kcal/mol. In the following, all sublimation energies and their deviations consistently refer to one molecule (and not the unit cell).

The calculations are carried out with the Vienna Ab-initio Simulation Package VASP 5.3 [68, 69]. We utilize the GGA functional PBE [70] in combination with a projector-augmented plane-wave basis set (PAW) [71, 72] with a huge energy cut-off of 1,000 eV. This corresponds to 200% of the recommended high-precision cut-off. We sample the Brillouin zone with a Γ-centered k-point grid with four k-points in each direction, generated via the Monkhorst–Pack scheme [73]. To simulate isolated molecules in the gas phase, we compute the Γ-point energy of a single molecule in a large unit cell (minimum distance between separate molecules of 16 Å, e.g., adamantine is calculated inside a $19 \times 19 \times 19 \times$ Å^3 unit cell). In order to calculate the sublimation energy, we optimize the single molecule and the corresponding molecular crystal. The unit cells are kept fixed at the experimental values. The atomic coordinates are optimized with an extended version of the approximate normal coordinate rational function optimization program (ANCOPT) [74] until all forces are below 10^{-4} Hartree/Bohr. We compute the D3 dispersion

Table 1 Mean absolute deviation (MAD), mean deviation (MD), and standard deviation (SD) of the calculated, zero-point exclusive sublimation energy from reference values for the X23 test set. The energies and geometries refer to the PBE/1,000 eV, PBE-D3/1,000 eV, PBE-D3/1,000 eV $+E^{(3)}$ levels. Values for the XDM and TS method are taken from [24] and the data for 16 systems on the PBE-MBD level from [67]. Negative MD values indicate systematic underbinding

	X23 sublimation energy		
Method	MAD	MD	SD
PBE/1,000 eV	11.55	−11.55	6.20
PBE-D3/1,000 eV	1.07	0.43	1.34
PBE-D3/1,000 eV+$E^{(3)}$	1.21	−0.49	1.65
PBE-XDM/1,088 eV	1.50	−0.45	2.12
B86b-XDM/1,088 eV	1.37	−0.33	1.91
PBE-TS/1,088 eV	1.53	3.50	2.32
PBE-MBD/1,000 eV	1.53	1.53	0.95

All energies are in kcal/mol per molecule

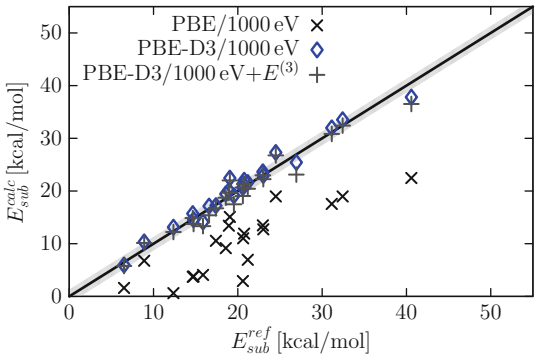


Fig. 3 Correlation between experimental and PBE computed sublimation energy with and without dispersion correction. The *gray shading* along the *diagonal line* denotes the experimental error interval. All energies are calculated on optimized structures but with experimental lattice constants

energy in the Becke–Johnson damping scheme with a conservative distance cut-off of 100 Bohr. The three-body dispersion energy is always calculated as a single-point on the optimized PBED3/1,000 eV structure. The results for X23 are summarized in Table 1. Figure 3 shows the correlation between experimental sublimation energies and the calculated values on the PBE/1,000 eV, PBE-D3/1,000 eV, and PBE-D3/1,000 eV+$E^{(3)}$ levels. The uncorrected functional yields unreasonable results. Because of the missing dispersion interactions, the attraction between the molecules is significantly underestimated, which results in too small sublimation energies. Some systems are not bound at all on the PBE/1,000 eV level. For PBE-D3 all results are significantly improved. The MAD is exceptionally low and drops below the estimated experimental error of 1.3 kcal/mol. The mean deviation of +0.4 kcal/mol indicates a slight overbinding on the PBE-D3/

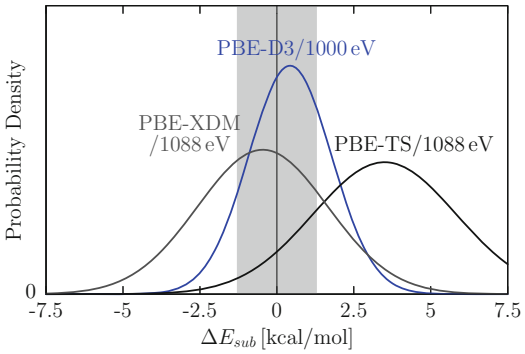


Fig. 4 Deviations between experimental and theoretical sublimation energies for the X23 set. We convert the statistical data into standard normal error distributions for visualization. The *gray shading* denotes the experimental error interval. The quality of the theoretical methods decreases in the following order: PBE-D3/1,000 eV, PBE-XDM/1,088 eV, and PBE-TS/1,088 eV

1,000 eV level. The three-body dispersion correction is always repulsive and therefore decreases the sublimation energy. At the PBE-D3/1,000 eV+$E^{(3)}$ level the MAD and SD is slightly raised but these changes are within the uncertainty of the reference data and hence we cannot draw definite conclusions about the importance of three-body dispersion effects from this comparison. Because inclusion of three-body dispersion has been shown to improve the description of binding in large supramolecular structures [45] and is not spoiling the results here, we recommend that the term is always included. However, the many-body effect (i.e., adding $E^{(3)}$ to the PBE-D3 data) is smaller than found in recent studies by another group [58, 75] employing a general many-body dispersion scheme. We compare our results to the pairwise dispersion corrections XDM and TS and show the normal error distributions in Fig. 4. The XDM model works reasonably well with an MAD of 1.5 kcal/mol, while the TS scheme is significantly overbinding with an MAD of 3.5 kcal/mol. The overbinding of the TS model is partially compensated by large many-body contributions and the MAD on the PBE-MBD level drops to 1.5 kcal/mol. A remarkable accuracy with an MAD of 0.9 kcal/mol was reported with the hybrid functional PBE0-MBD [67, 76]. The XDM model works slightly better in combination with the more repulsive B86b functional. However, the mean deviation of –0.5 kcal/mol and –0.3 kcal/mol reveals a systematic underbinding of the XDM method consistent with results for supramolecular systems (ER Johnson (2013), Personal Communication). This will lead to a worse result when a three-body term is included.

As a further test we investigate the unit cell volume for the same systems. We perform a full geometry optimization and compare with the experimental low-temperature X-ray structures. The unit cell optimization is done with the VASP quasi-Newton optimizer with a force convergence threshold of 0.005 eV/A°. Without dispersion correction, too large unit cells are obtained. On the PBE/1,000 eV level, the volumes of the orthorhombic systems are overestimated by 9.7%. We compare the

theoretical zero Kelvin geometries with low-temperature X-ray diffraction data at approximately 100 K. Therefore, the calculated values should always be smaller than the measured ones due to thermal expansion effects. After applying the D3 correction, the unit cells are systematically too small by 0.8% which is reasonable considering typical thermal volume expansions assumed to be approximately 3%. In passing it is noted that the geometries of isolated organic molecules are systematically too large in volume by about 2% with PBE-D3 [77], which is consistent with the above findings. In summary, PBE-D3 or PBE-D3 + $E^{(3)}$ provide a consistent treatment of interaction energies and structures in organic solids. Screening effects on the dispersion interaction as discussed in [58, 75] seem to be unimportant in the D3 model.

2.2.2 Structure of Tribenzotriquinazene (TBTQ)

As an example for a larger system where London dispersion is even more important, we re-investigate the recently studied tribenzotriquinacene (TBTQ) compound [34] which involves π-stacked aromatic units. We utilized the GGA functionals PBE [70] and RPBE [78], a PAW basis set [71, 72] with huge energy cut-off of 1,000 eV within the VASP program package. The crystal structures of TBTQ and its centro-methyl derivate (Me-TBTQ) was measured and a space group R3m was found for both TBTQ and Me-TBTQ. However, a refined analysis revealed the true space group of TBTQ to be R3c (an additional *c*-glide plane), while the space group of Me-TBTQ is confirmed. The structure in Fig. 5 shows the tilting between neighboring TBTQ layers. With dispersion corrected DFT (PBE-D3/1,000 eV), we were able to obtain all subtle details of the structures as summarized in Table 2. The unusual packing induced torsion between vertically stacked molecules was computed correctly as well as an accurate stacking distance. The deviations from experimental unit cell volumes of 1.4% for TBTQ and 1.5% for Me-TBTQ are within typical thermal volume expansions. The agreement between theory and experiment is excellent but necessitated a huge basis set with 1.46×10^5 plane-wave basis functions. A calculation of the crystal structure of Me-TBTQ on the same theoretical level confirms the measured untilted stacking geometry.

The dispersion correction is also crucial for the correct description of the sublimation energy. For PBE negative values (no net bindings) are obtained. On the PBE-D3 level reasonable ZPVE-exclusive sublimation energies of 35 and 29 kcal/mol are calculated, which fit the expectations for molecules of this size. In Fig. 6 we show the potential energy surface (PES) with respect to the vertical stacking distance for Me-TBTQ. In addition to the PBE functional, we applied the Hammer et al. modified version, dubbed RPBE [78], to investigate the effect of the short-range correlation kernel. For each point, we perform a full geometry optimization with a fixed unit cell geometry. The curves for both uncorrected

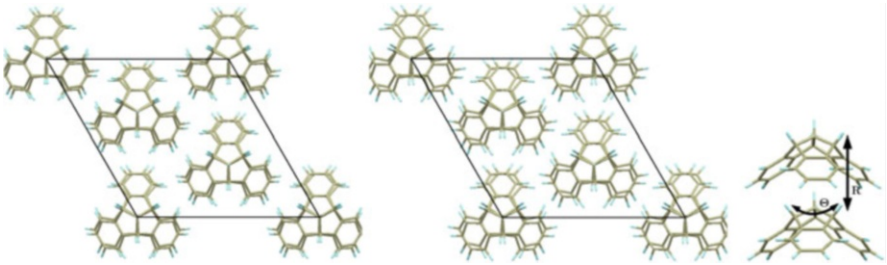


Fig. 5 X-Ray (*left*) and PBE-D3/1,000 eV (*middle*) crystal structure of TBTQ. The computed structure was obtained by an unconstrained geometry optimization [34]. The *right* figure highlights the analyzed geometry descriptors

Table 2 Comparison of experimental X-ray and computed PBE-D3/1,000 eV structures. The first block corresponds to the TBTQ crystal, the second to the Me-TBTQ crystal. As important geometrical descriptors the vertical stacking distance R, the tilting angle Θ, and the unit cell volume Ω are highlighted

	X-Ray	PBE-D3/1,000 eV
R	4.75	4.67
$\boldsymbol{\Theta}$	6.2°	9.8°
$\boldsymbol{\Omega}$	2,075	2,046
a, b, c	15.96, 15.96, 9.48	15.92, 15.92, 9.32
α, β, γ	90.0, 90.0, 120.0	90.0, 90.0, 120.0
R	5.95	5.91
$\boldsymbol{\Theta}$	0.0°	0.0°
$\boldsymbol{\Omega}$	2,306	2,272
a, b, c	14.96, 14.96, 11.90	14.90, 14.90, 11.82
α, β, γ	90.0, 90.0, 120.0	90.0, 90.0, 120.0

All lengths are given in Å

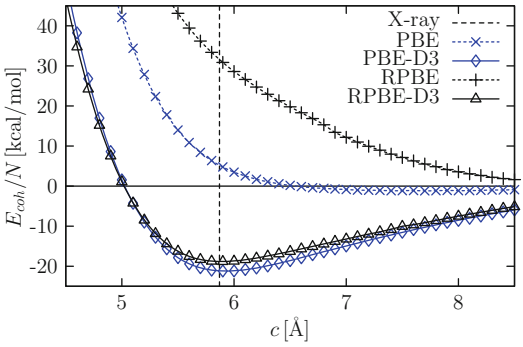


Fig. 6 Dependence of the cohesive energy E_{coh} per molecule on the vertical cell parameter c (the *dashed line* denotes the experimental value). The results refer to the PBE and RPBE functional with a PAW basis set and an energy cut-off of 1,000 eV. The cell parameters a and b are fixed to their experimental value. For each point we perform a full geometry optimization with a fixed unit cell geometry. The asymptotic energy limit $c \to \infty$ corresponds to the interaction in one Me-TBTQ layer, approximated by a large distance of 15 Å

functionals show no significant minimum in agreement with the wrong sign of the sublimation energy. Furthermore, we see significant deviation between the two functionals, i.e., PBE is much less repulsive than RPBE. With the inclusion of the D3 correction the differences between both functionals diminishes nicely and the PES are nearly identical. This is a strong indication that the D3 correction provides a physically sound description of long- and medium-range correlation effects. In fact, RPBE-D3 reproduces the equilibrium structure even slightly better than PBED3. This confirms previous observations from different groups that dispersion corrections are ideally coupled to inherently more repulsive (semi-local) functionals [19, 79, 80].

3 Dispersion Corrected Hartree–Fock with Basis Set Error Corrections

3.1 Basis Set Error Corrections

The previously presented results were obtained with huge plane-wave basis sets and these DFT calculations are rather costly. It seems hardly possible to use fewer plane-wave functions, because the stronger oscillating functions are necessary to describe the relatively localized electron density in molecular crystals. A significant reduction of basis functions seems only possible with atom centered functions, i.e., Gaussian atomic orbitals (AO). In contrast to plane-waves, however, small AO basis sets suffer greatly from basis set incompleteness errors, especially the BSSE. Semi-diffuse AOs can exhibit near linear dependencies in periodic calculations and the reduction of the BSSE by systematic improvement of the basis is often not possible. A general tool to correct for the BSSE efficiently in a semi-empirical way was developed in 2012 by us [36]. Recently, we extended the gCP denoted scheme to periodic systems and tested its applicability for molecular crystals [37].

Additionally, the basis set incompleteness error (BSIE) becomes crucial when near minimal basis sets are used. For a combination of Hartree–Fock with a MINIX basis (combination of valence scaled minimal basis set MINIS and split valence basis sets SV, SVP as defined in [81]), dispersion correction D3, and geometric counterpoise correction gCP, we developed a short-ranged basis set incompleteness correction dubbed SRB. The SRB correction compensates for too long covalent bonds. These are significant in an HF calculation with very small basis sets, especially when electronegative elements are present. The HF-D3-gCP-SRB/MINIX method will be abbreviated HF-3c in the following. The HF method has the advantage over current GGA functionals that it is (one-electron) self interaction error (SIE) free [82, 83]. Further, it is purely analytic and no grid error can occur. The numerical noise-free derivatives are important for accurate frequency calculations. In contrast to many semi-empirical methods, HF-3c can be applied to almost all elements of the periodic table without any further parameterization and the

physically important Pauli-exchange repulsion is naturally included. Here, we extend the HF-3c scheme to periodic systems and propose its use as a cheap DFT-D3 alternative or for crosschecking of DFT-D3 results.

The corrected total energy $E_{\mathrm{tot}}^{\mathrm{HF-3c}}$ is given by the sum of the HF energy $E^{\mathrm{HF/MINIX}}$, dispersion energy $E_{\mathrm{disp}}^{\mathrm{D3}}$, BSSE correction $E_{\mathrm{BSSE}}^{\mathrm{gCP}}$, and short-ranged basis incompleteness correction E_{SRB}:

$$E_{\mathrm{tot}}^{\mathrm{HF-3c}} = E^{\mathrm{HF/MINIX}} + E_{\mathrm{disp}}^{\mathrm{D3}} + E_{\mathrm{BSSE}}^{\mathrm{gCP}} + E_{\mathrm{SRB}}. \tag{10}$$

The form of the first term $E_{\mathrm{disp}}^{\mathrm{D3}}$ is already described in Sect. 2.1. For the HF-3c method the three parameters of the damping function s_8, a_1, and a_2 were refitted in the MINIX basis (while applying gCP) against reference interaction energies [84] and this is denoted D3(refit). The second correction, namely the geometrical counterpoise correction gCP [36, 37], depends only on the atomic coordinates and the unit cell of the crystal. The difference in atomic energy $e_{\mathrm{A}}^{\mathrm{miss}}$ between a large basis (def2-QZVPD [85]) and the target basis set (e.g., the MINIX basis) inside a weak electric field is computed for free atoms A. The $e_{\mathrm{A}}^{\mathrm{miss}}$ term measures the basis incompleteness and is used to generate an exponentially decaying, atom-pairwise repulsive potential. The BSSE energy correction $E_{\mathrm{BSSE}}^{\mathrm{gCP}}$ EgCP BSSE reads

$$E_{\mathrm{BSSE}}^{\mathrm{gCP}} = \frac{\boldsymbol{\sigma}}{2} \sum_{\mathrm{A}\neq\mathrm{B}}^{\text{atom pairs}} \sum_{\mathbf{T}} e_{\mathrm{A}}^{\mathrm{miss}} \frac{\exp(-\boldsymbol{\alpha} \cdot \|\mathbf{r}_{\mathrm{B}} - \mathbf{r}_{\mathrm{A}} + \mathbf{T}\|\boldsymbol{\beta})}{\sqrt{S_{\mathrm{AB}} \cdot N_{\mathrm{B}}^{\mathrm{virt}}}}, \tag{11}$$

with Slater-type overlap integral S_{AB}, number of virtual orbitals on atom B in the target basis set $N_{\mathrm{B}}^{\mathrm{virt}}$, and basis set dependent fit parameters $\boldsymbol{\sigma}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$. The Slater exponents of s- and p-valence orbitals are averaged and scaled by a fourth fit parameter $\boldsymbol{\eta}$ to get a single s-function exponent. For each combination of Hamiltonian (DFT or HF) and basis set, the four parameters were fitted in a least-squares sense against counterpoise correction data obtained by the Boys–Bernardi scheme [35].

Systematically overestimated covalent bond lengths for electronegative elements

are corrected by the third term E_{SRB}:

$$E_{\mathrm{SRB}} = -\frac{s}{2} \sum_{\mathrm{A}\neq\mathrm{B}}^{\text{atom pairs}} \sum_{\mathbf{T}} (Z_{\mathrm{A}} Z_{\mathrm{B}})^{3/2} \exp\left(-\gamma \left(R_{\mathrm{AB}}^{0,\mathrm{D3}}\right)^{3/4} \|\mathbf{r}_{\mathrm{B}} - \mathbf{r}_{\mathrm{A}} + \mathbf{T}\|\right). \tag{12}$$

We use the default cut-off radii $R_{\mathrm{AB}}^{0,\mathrm{D3}}$ as determined ab initio for the D3 dispersion correction and $Z_{\mathrm{A/B}}$ are the nuclear charges. The parameters s and γ were determined by fitting the HF-3c total forces against B3LYP-D3/def2-TZVPP [86] equilibrium structures of 107 small organic molecules. Altogether, the HF-3c method consists of nine empirically determined parameters, three for the D3

dispersion, four in the gCP scheme, and two for the SRB correction. The HF-3c method was recently tested for geometries of small organic molecules, interaction energies and geometries of non-covalently bound complexes, for supramolecular systems, and protein structures [81], and good results superior to traditional semi-empirical methods were obtained. In particular the accurate non-covalent HF-3c interactions energies for a standard benchmark [84] (i.e., better than with the "costly" MP2/CBS method and close to the accuracy of DFT-D3/"large basis") are encouraging for application to molecular crystals.

3.2 Evaluation of Dispersion and Basis Set Corrected DFT and HF

We evaluate the basis corrections gCP and SRB by comparison with reference sublimation energies for the X23 benchmark set, introduced in Sect. 2.2. We calculate the HF and DFT energies with the widely used crystalline orbital program CRYSTAL09 [87, 88]. In the CRYSTAL code, the Bloch functions are obtained by a direct product of a superposition of atom-centered Gaussian functions and a **k** dependent phase factor. We use raw HF, the GGA functional PBE [70], and the hybrid GGA functional B3LYP [89, 90]. The Γ-centered k-point grid is generated via the Monkhorst–Pack scheme [73] with four k-points in each direction. The large integration grid (LGRID) and tight tolerances for Coulomb and exchange sums (input settings. TOLINTEG 8 8 8 8 16) are used. The SCF energy convergence threshold is set to 10^{-8} Hartree. We exploit the polarized split-valence basis set SVP [91] and the near minimal basis set MINIX. The atomic coordinates are optimized with the extended version of the approximate normal coordinate rational function optimization program (ANCOPT) [74].

Mean absolute deviation (MAD), mean deviation (MD), and standard deviation (SD) of the sublimation energy for the X23 test set and for the subset X12/Hydrogen (systems dominated by hydrogen bonds) are presented in Table 3. The dispersion and BSSE corrected PBE-D3-gCP/SVP and B3LYP-D3-gCP/SVP methods yield good sublimation energies with MADs of 2.5 and 2.0 kcal/mol, respectively. The artificial overbinding of the gCP-uncorrected DFT-D3/SVP methods is demonstrated by the huge MD of 8.5 kcal/mol for PBE and 10.1 kcal/mol for B3LYP. Adding the three-body dispersion energy changes the MADs for D3-gCP to 2.9 and 1.7 kcal/mol, respectively. As noted before [37], the PBE functional with small basis sets underbinds hydrogen bonded systems systematically. The HF-3c calculated sublimation energies are of very good quality with an MAD of 1.7 and 1.5 kcal/mol without and with three-body dispersion energy, respectively, which is similar to the previous PBE-D3/1,000 eV results. Considering the simplicity of this approach, this result is remarkable. The MD is with 0.6 and –0.2 kcal/mol, respectively, also very close to zero. This indicates that, with the three correction terms, most of the systematic errors of pure HF are eliminated. For hydrogen bonded systems the MAD is only slightly

Table 3 Mean absolute deviation (MAD), mean deviation (MD), and standard deviation (SD) of the computed sublimation energy with respect to experimental reference data for the X23 test set and for the subset X12/Hydrogen dominated by hydrogen bonds. We compare the HF-3c method with gCP corrected PBE-D3/SVP and B3LYP-D3/SVP methods. For PBE/SVP level, we also give deviations to the corresponding large plane-wave basis set values in parentheses

	X23			X12/Hydrogen	
Method	MAD	MD	SD	MAD	MD
PBE-D3/SVP	8.5 (8.1)	8.5 (8.1)	3.5 (3.4)	10.5 (9.7)	10.5 (9.7)
PBE-D3-gCP/SVP	2.5 (2.1)	−1.1 (1.5)	3.0 (2.6)	2.8 (2.5)	−1.4 (–2.3)
PBE-D3-gCP/SVP+$E^{(3)a}$	2.9 (2.0)	−2.0 (–1.5)	3.2 (2.5)	3.1 (2.4)	−2.2 (–2.2)
B3LYP-D3/SVP	10.1	10.1	4.1	12.0	12.0
B3LYP-D3-gCP/SVP	2.0	0.5	2.3	1.7	−0.1
B3LYP-D3-gCP/SVP+$E^{(3)a}$	1.7	−0.4	2.2	1.8	−0.8
HF/MINIX[b]	11.3	−11.3	6.1	10.7	−10.7
HF-D3(refit)/MINIX[b]	6.3	6.3	3.6	7.5	7.5
HF-D3(refit)-gCP/MINIX[b]	1.6	0.5	1.9	1.8	−0.0
HF-3c	1.7	0.6	2.0	1.8	0.0
HF-3c+$E^{(3)a}$	1.5	−0.2	2.0	2.0	−0.7

[a]Three-body dispersion $E^{(3)}$ as single-point energy on optimized structures
[b]Single-point energies on HF-3c optimized structures
All values are in kcal/mol per molecule

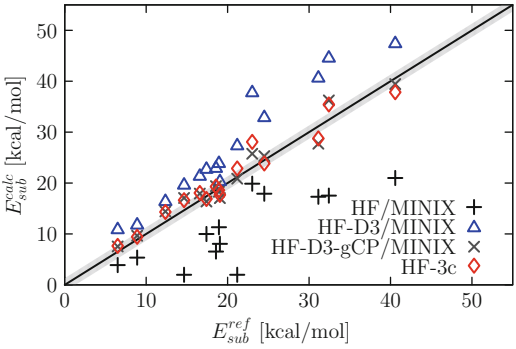


Fig. 7 Correlation between experimental sublimation energy and HF results with subsequent addition of the three corrections. All sublimation energies are calculated on optimized HF-3c structures for experimental lattice constants. The *gray shading* along the *diagonal line* denotes the experimental error interval

higher, which indicates an overall consistent treatment. To analyze the HF-3c method in more detail, we investigate the different energy contribution to the sublimation energy on the optimized HF-3c structures as shown in Fig. 7.

Plain HF is not capable of describing the intermolecular attraction in the crystals and has the largest MAD of 11.3 kcal/mol. The only significant physical attraction between the molecules arises in hydrogen bonded systems which are dominated by

electrostatics which is properly described by HF. By inclusion of dispersion, the MAD drops to 6.3 kcal/mol on the HF-D3(refit)/MINIX level, but the sublimation energy is significantly overestimated. This too strong attraction can be efficiently and accurately corrected with the gCP scheme. The MAD on the HF-D3(refit)-gCP/MINIX level is 1.6 kcal/mol and very similar to the MAD of the full HF-3c method. This demonstrates that the SRB correction mainly affects geometries as intended. Because the energy decomposition analysis is done for fixed geometries, we cannot investigate the importance of the E_{SRB} contribution in more detail. In conclusion, the computationally very cheap HF-3c method provides encouraging energies. However, for a few systems we encounter convergence problems of the SCF procedure with the CRYSTAL09 code. This can be sometimes avoided with tighter tolerances for Coulomb and exchange integral sums with the side effect of increased computational cost. Zero point vibrational energies are not analyzed here, but numerically stable second energy derivatives of HF-3c were reported in [81].

4 Conclusions

We have presented and evaluated dispersion corrected Hartree–Fock and Density Functional Theories for their potential application to computed organic crystals and their properties. For a correct description of molecular crystals, semi-local (hybrid) density functionals have to be corrected for London dispersion interactions. A variety of modern DFT-D methods, namely D3, TS/MBD, and XDM, can calculate sublimation energies of small organic crystals with errors close to the experimental uncertainty. For the X23 test set we found that the D3 scheme gives the best performance of the tested additive dispersion corrections with an MAD of 1.1 kcal/mol, which is well below the estimated error range of 1.3 kcal/mol. In the DFT-D3 scheme the three-body dispersion energy corrections are approximately 5% of the sublimation energy. The finding that the method, which has been developed originally for molecules and molecular complexes, can be applied without further, solid-state specific modifications is encouraging. It was furthermore shown that DFT-D3 can calculate the π-stacking of tribenzotriquinacene and its centro-methyl derivative with all subtle geometry details. This example demonstrates that larger molecules routinely considered in organic chemistry can also be treated accurately in their solid state by DFT based methods.

In addition to these calculations with huge plane-wave based basis sets, we exploited Gaussian atom-centered orbitals. We demonstrated the large basis set errors on the DFT-D3/SVP and HF-D3/MINIX levels and presented and evaluated two semi-empirical basis set corrections. The resulting DFT-D3-gCP/SVP and HF-3c methods perform well and the MAD of 1.5 kcal/mol (with three-body dispersion) for HF-3c is especially remarkable. However, the SCF convergence with unscreened Fock-exchange is sometimes problematic and, despite a larger basis being used, the PBE-D3-gCP/SVP calculations converge faster and yield an acceptable MAD of 2.5 kcal/mol for the X23 sublimation energies.

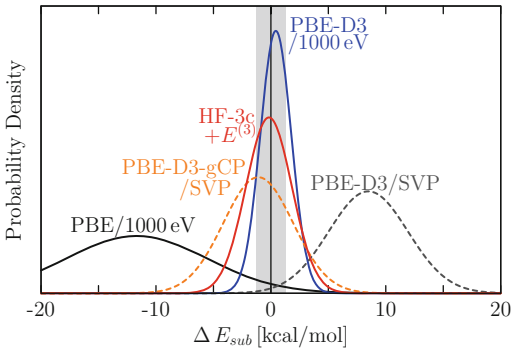


Fig. 8 Deviations between experimental and theoretical sublimation energies for the X23 set. We convert the statistics into standard normal error distributions for visualization. The *gray shading* denotes the experimental error interval. The quality of the theoretical methods decreases in the following order. PBED3/1,000 eV, HF-3c+$E^{(3)}$, PBE-D3-gCP/SVP, PBED3/ SVP, and PBE/1,000 eV

In Fig. 8 we summarize the results of the various theoretical methods for the X23 benchmark set by converting the statistical data into standard normal distributions. The best results are calculated with the D3 dispersion corrected PBE functional in a huge PAW basis set. HF-3c + $E^{(3)}$ and PBE-D3-gCP/SVP can also be recommended.

In future work the description of energy rankings of polymorphs on the different theoretical levels has to be investigated systematically. Furthermore, coupling of the D3 dispersion correction to different GGA, meta-GGA, and hybrid GGA functionals might provide even better performance. In any case, the future for fully quantum chemical based first principles crystal structure prediction seems bright.

References

1. Neumann MA, Leusen FJJ, Kendrick J (2008) A major advance in crystal structure prediction. Angew Chem Int Ed 47:2427–2430
2. Oganov AR (2010) Modern methods of crystal structure prediction. Wiley-VCH, Berlin
3. Oganov AR, Glass CW (2006) Crystal structure prediction using ab initio evolutionary techniques. Principles and applications. J Chem Phys 124:244–704
4. Sanderson K (2007) Model predicts structure of crystals. Nature 450:771–771
5. Woodley SM, Catlow R (2008) Crystal structure prediction from first principles. Nat Mater 7:937–964
6. Dreizler J, Gross EKU (1990) Density functional theory, an approach to the quantum many-body problem. Springer, Berlin
7. Koch W, Holthausen MC (2001) A chemist's guide to Density Functional Theory. Wiley-VCH, New York

8. Parr RG, Yang W (1989) Density-functional theory of atoms and molecules. Oxford University Press, Oxford
9. Paverati R, Truhlar DG (2013) The quest for a universal density functional. The accuracy of density functionals across a braod spectrum of databases in chemistry and physics. Phil Trans R Soc A, in press. http://arxiv.org/abs/1212.0944
10. Allen M, Tozer DJ (2002) Helium dimer dispersion forces and correlation potentials in density functional theory. J Chem Phys 117:11113
11. Hobza P, Sponer J, Reschel T (1995) Density functional theory and molecular clusters. J Comput Chem 16:1315–1325
12. Kristy´an S, Pulay P (1994) Can (semi)local density functional theory account for the London dispersion forces? Chem Phys Lett 229:175–180
13. Pérez-Jord´a JM, Becke AD (1995) A density-functional study of van der Waals forces. Rare gas diatomics. Chem Phys Lett 233:134–137
14. Kaplan IG (2006) Intermolecular interactions. Wiley, Chichester
15. Stone AJ (1997) The theory of intermolecular forces. Oxford University Press, Oxford
16. Burns LA, Vazquez-Mayagoitia A, Sumpter BG, Sherrill CD (2011) A comparison of dispersion corrections (DFT-D), exchange-hole dipole moment (XDM) theory, and specialized functionals. J Chem Phys 134:084,107
17. Civalleri B, Zicovich-Wilson CM, Valenzano L, Ugliengo P (2008) B3LYP augmented with an empirical dispersion term (B3LYP-D*) as applied to molecular crystals. CrystEngComm 10:405–410
18. Gräfenstein J, Cremer D (2009) An efficient algorithm for the density functional theory treatment of dispersion interactions. J Chem Phys 130:124,105
19. Grimme S (2011) Density functional theory with London dispersion corrections. WIREs Comput Mol Sci 1:211–228
20. Grimme S, Antony J, Schwabe T, Mück-Lichtenfeld C (2007) Density functional theory with dispersion corrections for supramolecular structures, aggregates, and complexes of (bio) organic molecules. Org Biomol Chem 5:741–758
21. Jacobsen H, Cavallo L (2012) On the accuracy of DFT methods in reproducing ligand substitution energies for transition metal complexes in solution. The role of dispersive interactions. ChemPhysChem 13:562–569
22. Johnson ER, Mackie ID, DiLabio GA (2009) Dispersion interactions in density-functional theory. J Phys Org Chem 22:1127–1135
23. Klimes J, Michaelides A (2012) Perspective. Advances and challenges in treating van der Waals dispersion forces in density functional theory. J Chem Phys 137:120901
24. de-la Roza AO, Johnson ER (2012) A benchmark for non-covalent interactions in solids. J Chem Phys 137:054103
25. Vydrov OA, Van Voorhis T (2010) Nonlocal van derWaals density functional. The simpler the better. J Chem Phys 133:244103
26. Eshuis H, Bates JE, Furche F (2012) Electron correlation methods based on the random phase approximation. Theor Chem Acc 131:1084
27. Heßelmann A, Jansen G, Schütz M (2005) Density-functional theorysymmetry-adapted intermolecular perturbation theory with density fitting. A new efficient method to study intermolecular interaction energies. J Chem Phys 122:014103
28. Nanda K, Beran G (2012) Prediction of organic molecular crystal geometries from MP2-level fragment quantum mechanical/molecular mechanical calculations. J Chem Phys 138:174106
29. Wen S, Nanda K, Huang Y, Beran G (2012) Practical quantum mechanics based fragment methods for predicting molecular crystal properties. Phys Chem Chem Phys 14:7578–7590
30. Grimme S, Antony J, Ehrlich S, Krieg H (2010) A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. J Chem Phys 132:154104
31. Grimme S, Ehrlich S, Goerigk L (2011) Effect of the damping function in dispersion corrected density functional theory. J Comput Chem 32:1456–1465

32. Goerigk L, Grimme S (2011) A thorough benchmark of density functional methods for general main group thermochemistry, kinetics, and noncovalent interactions. Phys Chem Chem Phys 13:6670–6688
33. Krukau AV, Vydrov OA, Izmaylov AF, Scuseria GE (2006) Influence of the exchange screening parameter on the performance of screened hybrid functionals. J Chem Phys 125:224106
34. Brandenburg JG, Grimme S, Jones PG, Markopoulos G, Hopf H, Cyranski MK, Kuck D (2013) Unidirectional molecular stacking of tribenzotriquinacenes in the solid state – a combined X-ray and theoretical study. Chem Eur J 19:9930–9938
35. Boys SF, Bernardi F (1970) The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors. Mol Phys 19:553–566
36. Kruse H, Grimme S (2012) A geometrical correction for the inter- and intramolecular basis set superposition error in Hartree–Fock and density functional theory calculations for large systems. J Chem Phys 136:154101
37. Brandenburg JG, Alessio M, Civalleri B, Peintinger MF, Bredow T, Grimme S (2013) Geometrical correction for the inter- and intramolecular basis set superposition error in periodic density functional theory calculations. J Phys Chem A 117:9282–9292
38. Dion M, Rydberg H, Schr¨oder E, Langreth DC, Lundqvist BI (2004) Van der Waals density functional for general geometries. Phys Rev Lett 92:246401
39. Lee K, Murray ED, Kong L, Lundqvist BI, Langreth DC (2010) Higher accuracy van der Waals density functional. Phys Rev B 82:081101
40. Grimme S, Hujo W, Kirchner B (2012) Performance of dispersion-corrected density functional theory for the interactions in ionic liquids. Phys Chem Chem Phys 14:4875–4883
41. Hujo W, Grimme S (2011) Comparison of the performance of dispersion corrected density functional theory for weak hydrogen bonds. Phys Chem Chem Phys 13:13942–13950
42. Hujo W, Grimme S (2011) Performance of the van der Waals density functional VV10 and (hybrid) GGA variants for thermochemistry and noncovalent interactions. J Chem Theory Comput 7:3866–3871
43. HujoW GS (2013) Performance of non-local and atom-pairwise dispersion corrections to DFT for structural parameters of molecules with noncovalent interactions. J Chem Theory Comput 9:308–315
44. Casimir HBG, Polder D (1948) The influence of retardation on the London van der Waals forces. Phys Rev 73:360–372
45. Grimme S (2012) Supramolecular binding thermodynamics by dispersion-corrected density functional theory. Chem Eur J 18:9955–9964
46. Becke AD, Johnson ER (2005) A density-functional model of the dispersion interaction. J Chem Phys 123:154101
47. Starkschall G, Gordon RG (1972) Calculation of coefficients in the power series expansion of the long range dispersion force between atoms. J Chem Phys 56:2801
48. Moellmann J, Grimme S (2010) Importance of London dispersion effects for the packing of molecular crystals: a case study for intramolecular stacking in a bis-thiophene derivative. Phys Chem Chem Phys 12:8500–8504
49. Axilrod BM, Teller E (1943) Interaction of the van der Waals type between three atoms. J Chem Phys 11:299
50. Ehrlich S, Moellmann J, Grimme S (2013) Dispersion-corrected density functional theory for aromatic interactions in complex systems. Acc Chem Res 46:916–926
51. Reckien W, Janetzko F, Peintinger MF, Bredow T (2012) Implementation of empirical dispersion corrections to density functional theory for periodic systems. J Comput Chem 33:2023–2031
52. Ahlrichs R, Penco R, Scoles G (1977) Intermolecular forces in simple systems. Chem Phys 19:119–130
53. Cohen JS, Pack RT (1974) Modified statistical method for intermolecular potentials. Combining rules for higher van der Waals coefficients. J Chem Phys 61:2372–2382

54. Gianturco FA, Paesani F, Laranjeira MF, Vassilenko V, Cunha MA (1999) Intermolecular forces from density functional theory. III. A multiproperty analysis for the Ar(^{1}S)-CO($^{1\Sigma}$) interaction. J Chem Phys 110:7832
55. Wu Q, Yang W (2002) Empirical correction to density functional theory for van der Waals interactions. J Chem Phys 116:515–524
56. Wu X, Vargas MC, Nayak S, Lotrich V, Scoles G (2001) Towards extending the applicability of density functional theory to weakly bound systems. J Chem Phys 115:8748
57. Tkatchenko A, Scheffler M (2009) Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data. Phys Rev Lett 102:073005
58. Tkatchenko A, DiStasio RA, Car R, Scheffler M (2012) Accurate and efficient method for many-body van der Waals interactions. Phys Rev Lett 108:236402
59. Sato T, Nakai H (2009) Density functional method including weak interactions. Dispersion coefficients based on the local response approximation. J Chem Phys 131:224104
60. Becke AD, Johnson ER (2007) A unified density-functional treatment of dynamical, nondynamical, and dispersion correlations. J Chem Phys 127:124108
61. Johnson ER, Becke AD (2006) A post-Hartree–Fock model of intermolecular interactions. Inclusion of higher-order corrections. J Chem Phys 124:174,104
62. Johnson ER, de-la Roza AO (2012) Adsorption of organic molecules on kaolinite from the exchange-hole dipole moment dispersion model. J Chem Theory Comput 8:5124–5131
63. Kim HJ, Tkatchenko A, Cho JH, Scheffler M (2012) Benzene adsorbed on Si(001). The role of electron correlation and finite temperature. Phys Rev B 85:041403
64. de-la Roza AO, Johnson ER, Contreras-García J (2012) Revealing non-covalent interactions in solids. NCI plots revisited. Phys Chem Chem Phys 14:12165–12172
65. Ruiz VG, Liu W, Zojer E, Scheffler M, Tkatchenko A (2012) Density-functional theory with screened van der Waals interactions for the modeling of hybrid inorganic–organic systems. Phys Rev Lett 108:146103
66. Chickos JS (2003) Enthalpies of sublimation after a century of measurement. A view as seen through the eyes of a collector. Netsu Sokutei 30:116–124
67. Reilly AM, Tkatchenko A (2013) Seamless and accurate modeling of organic molecular materials. J Phys Chem Lett 4:1028–1033
68. Bucko T, Hafner J, Lebegue S, Angyan JG (2010) Improved description of the structure of molecular and layered crystals. Ab initio DFT calculations with van der Waals corrections. J Phys Chem A 114(11):814–11824
69. Kresse G, Furthmüller J (1996) Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. Comput Mat Sci 6:15–50
70. Perdew JP, Burke K, Ernzerhof M (1996) Generalized gradient approximation made simple. Phys Rev Lett 77:3865, erratum: Phys Rev Lett 78:1369 (1997)
71. Bl¨ochl PE (1994) Projector augmented-wave method. Phys Rev B 50:17953
72. Kresse G, Joubert D (1999) From ultrasoft pseudopotentials to the projector augmented-wave method. Phys Rev B 59:1758
73. Monkhorst HJ, Pack JD (1976) Special points for Brillouin-zone integrations. Phys Rev B 13:5188–5192
74. Grimme S (2013) ANCOPT. Approximate normal coordinate rational function optimization program. Universität Bonn, Bonn
75. de-la Roza AO, Johnson ER (2013) Many-body dispersion interactions from the exchange-hole dipole moment model. J Chem Phys 138:054103
76. Reilly AM, Tkatchenko A (2013) Understanding the role of vibrations, exact exchange, and many-body van der Waals interactions in the cohesive properties of molecular crystals. J Chem Phys 139:024705
77. Grimme S, Steinmetz M (2013) Effects of London dispersion correction in density functional theory on the structures of organic molecules in the gas phase. Phys Chem Chem Phys 15:16031–16042

78. Hammer B, Hansen LB, Norskov JK (1999) Improved adsorption energetics within density-functional theory using revised Perdew–Burke–Ernzerhof functionals. Phys Rev B 59:7413
79. Kannemann FO, Becke AD (2010) van der Waals interactions in density-functional theory. Intermolecular complexes. J Chem Theory Comput 6:1081–1088
80. de-la Roza AO, Johnson ER (2012) Van der Waals interactions in solids using the exchange-hole dipole moment model. J Chem Phys 136:174109
81. Sure R, Grimme S (2013) Corrected small basis set Hartree–Fock method for large systems. J Comput Chem 34:1672–1685
82. Gritsenko O, Ensing B, Schipper PRT, Baerends EJ (2000) Comparison of the accurate Kohn–Sham solution with the generalized gradient approximations (GGAs) for the SN2 reaction $F^- + CH_3F \rightarrow FCH_3 + F^-$: a qualitative rule to predict success or failure of GGAs. J Phys Chem A 104:8558–8565
83. Zhang Y, Yang W (1998) A challenge for density functionals: self-interaction error increases for systems with a noninteger number of electrons. J Chem Phys 109:2604–2608
84. Řezáč J, Riley KE, Hobza P (2011) S66: a well-balanced database of benchmark interaction energies relevant to biomolecular structures. J Chem Theory Comput 7:2427–2438
85. Weigend F, Furche F, Ahlrichs R (2003) Gaussian basis sets of quadruple zeta valence quality for atoms HKr. J Chem Phys 119:12753
86. Weigend F, Ahlrichs R (2005) Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn. Design and assessment of accuracy. Phys Chem Chem Phys 7:3297–3305
87. Dovesi R, Orlando R, Civalleri B, Roetti C, Saunders VR, Zicovich-Wilson CM (2005) CRYSTAL: a computational tool for the ab initio study of the electronic properties of crystals. Z Kristallogr 220:571–573
88. Dovesi R, Saunders VR, Roetti C, Orlando R, Zicovich-Wilson CM, Pascale F, Civalleri B, Doll K, Harrison NM, Bush IJ, DArco P, Llunell M (2009) CRYSTAL09 user's manual. University of Torino, Torino
89. Becke AD (1993) Density-functional thermochemistry. III. The role of exact exchange. J Chem Phys 98:5648
90. Stephens PJ, Devlin FJ, Chabalowski CF, Frisch MJ (1994) Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. J Phys Chem 98:11623–11627
91. Schäfer A, Horn H, Ahlrichs R (1992) Fully optimized contracted Gaussian basis sets for atoms Li to Kr. J Chem Phys 97:2571–2577

Top Curr Chem (2014) 345: 25–58
DOI: 10.1007/128_2013_497
© Springer-Verlag Berlin Heidelberg 2014
Published online: 6 February 2014

General Computational Algorithms for Ab Initio Crystal Structure Prediction for Organic Molecules

Constantinos C. Pantelides, Claire S. Adjiman, and Andrei V. Kazantsev

Abstract The prediction of the possible crystal structure(s) of organic molecules is an important activity for the pharmaceutical and agrochemical industries, among others, due to the prevalence of crystalline products. This chapter considers the general requirements that crystal structure prediction (CSP) methodologies need to fulfil in order to be able to achieve reliable predictions over a wide range of organic systems. It also reviews the current status of a multistage CSP methodology that has recently proved successful for a number of systems of practical interest. Emphasis is placed on recent developments that allow a reconciliation of conflicting needs for, on the one hand, accurate evaluation of the energy of a proposed crystal structure and on the other hand, comprehensive search of the energy landscape for the reliable identification of all low-energy minima. Finally, based on the experience gained from this work, current limitations and opportunities for further research in this area are identified. We also consider issues relating to the use of empirical models derived from experimental data in conjunction with ab initio CSP.

Keywords CrystalOptimizer · CrystalPredictor · Lattice energy · Local approximate model · Polymorph

Contents

C.C. Pantelides (✉), C.S. Adjiman, and A.V. Kazantsev
Department of Chemical Engineering Centre for Process Systems Engineering,
Imperial College London, London SW7 2AZ, UK
e-mail: c.pantelides@imperial.ac.uk

Abbreviations

API	Active pharmaceutical ingredient
CCDC	Cambridge Crystallographic Data Centre
CDF	Conformational degree of freedom
CSD	Cambridge Structural Database
CSP	Crystal structure prediction
DFT	Density functional theory
DFT+D	Dispersion corrected density functional theory
LAM	Local approximate model
QM	Quantum mechanical
$rmsd_{15}$	Root mean square deviation of the 15-molecule coordination sphere

1 Introduction

Crystalline organic materials play an important role in many high-value manufacturing sectors such as the pharmaceutical, agrochemical and fine chemicals industries. For instance, the majority of active pharmaceutical ingredients (APIs) are produced and delivered as solids [1]. The propensity of medium-size organic molecules to crystallize in multiple forms ("polymorphs") leads to significant challenges for the industry as differences in crystal structure can lead to large changes in physical properties such as solubility, dissolution rate and mechanical strength. These variations affect both manufacturing process and product effectiveness, and the appearance of a new, more stable, crystal structure of a given API can have wide-ranging effects on the availability and economic value of a drug [2]. As a result, the crystalline structure of an API has become a key element of patent protection and regulatory approval.

Given the practical importance of polymorphism and its intrinsic scientific interest, much research effort has been devoted towards increased understanding of this phenomenon and converting this understanding into methodologies for crystal structure prediction (CSP). Five blind tests for CSP have been organised by the Cambridge Crystallographic Data Centre (CCDC) since 1999 [3], providing useful benchmarks and helping to identify areas where improvements and further research are needed. While the blind tests are based on a relatively small set of compounds, the publications summarising their results [3–7] provide some evidence of progress in the development of increasingly reliable methodologies. Of particular note is the growing ability to predict the solid state behaviour of molecules of size, complexity and characteristics that are relevant to the pharmaceutical industry [8–10].

1.1 Definition and Scope of the CSP Problem

The central problem of CSP can be summarised as follows:

Given the molecular diagrams for all chemical species (*neutral molecule*(*s*) *or ions*) *in the crystal*, *identify the thermodynamically most stable crystal structure at a given temperature and pressure*, *and also*, *in correct order of decreasing stability*, *other* (*metastable*) *crystal structures that are likely to occur in nature*.

From a thermodynamic point of view, the most stable crystal structure is that with the lowest Gibbs free energy at the given temperature and pressure and, where relevant, at the given composition (crystal stoichiometry). The other structures of interest are normally metastable structures with relatively low free energy values. Mathematically, all these structures correspond to local minima of the Gibbs free energy surface, with the global (i.e. lowest) minimum determining the most stable structure.

The scope of the CSP methodology presented in this chapter includes both single-component crystals and co-crystals, hydrates, solvates and salts. It is applicable to flexible molecules of a size typical of "small molecule" pharmaceuticals (i.e. up to several hundred daltons) and to crystals in all space groups, without restriction on the number of molecules in the asymmetric unit (any $Z' > 0$). Examples of such systems are presented in Fig. 1.

1.2 Requirements for General CSP Methodologies

In this chapter we are interested in CSP methodologies that can be applied reliably in a systematic and standardised manner across the wide range of systems defined above. Based on the experience of the last two decades of activity in CSP, but also from other areas of model-based science and engineering, this translates into certain key requirements:

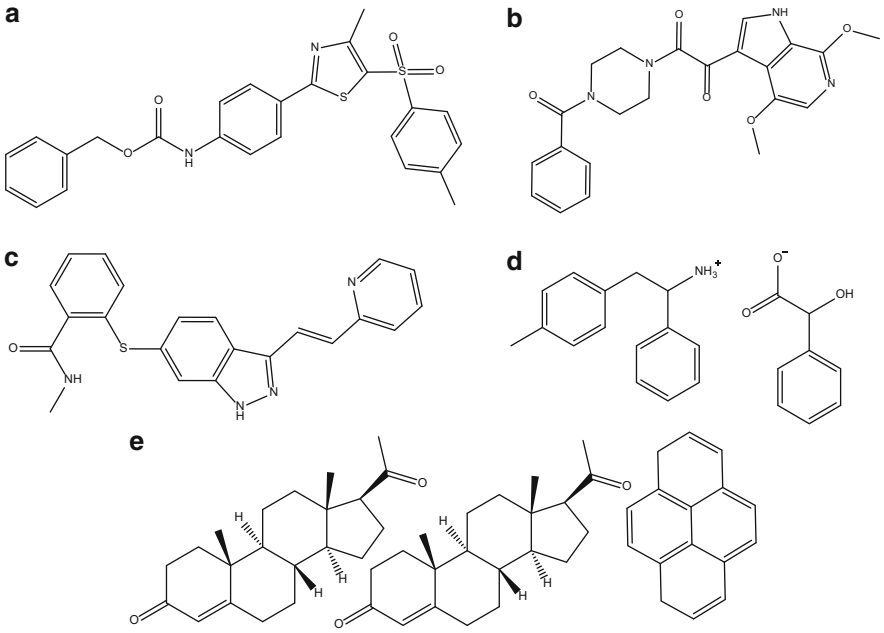


Fig. 1 Examples of systems of interest to current CSP methodologies. (**a**) "Molecule XX", fifth CCDC blind test target [4] (benzyl-(4-(4-methyl-5-(*p*-tolylsulfonyl)-1,3-thiazol-2-yl)phenyl)-carbamate). (**b**) Bristol-Myers Squibb's BMS-488043 [11] (1-[4-(benzoyl)piperazin-1-yl]-2-(4,7-dimethoxy-1*H*-pyrrolo[5,4-*c*]pyridin-3-yl)ethane-1,2-dione). (**c**) Pfizer's Axitinib anti-cancer drug [12] (*N*-methyl-2-[[3-[(*E*)-2-pyridin-2-ylethenyl]-1*H*-indazol-6-yl]sulfanyl]benzamide). (**d**) (*R*)-1-phenyl-2-(4-methylphenyl)ethylammonium-(*S*)-mandelate salt [13]. (**e**) Progesterone-pyrene (2:1) co-crystal [14]

- A reliable CSP methodology must be based on automated algorithms, with minimal need for user intervention beyond the specification of the problem to be tackled. This in turn limits the scope for reliance on previous experience and/or similarities with other systems, which in any case can lead to erroneous results as small changes in molecular structure can result in significant changes in the crystal energy landscape [15], including the number of local minima and the detailed geometry of the crystal packing. Statistical analysis of experimental evidence, such as that contained in the Cambridge Structural Database (CSD), does not always provide reliable guidance and sometimes leads to potentially relevant stable/metastable crystal structures being missed. In past blind tests [7], this was one of the stated reasons for failing to produce successful matches to experimental crystal structures.
- It must have a consistent, fundamental physical basis that can be applied uniformly to wide classes of systems. In our experience, "special tricks" (e.g. case-by-case adjustments of intermolecular interactions), whilst sometimes successful at reproducing known experimental structures for specific molecules, lead to limited predictive capability. They also sometimes obscure the real issues

that need to be addressed, acting as an obstacle to gaining the understanding that is necessary for the advancement of the field.
- It must produce consistently reliable solutions, e.g. as judged in terms of its ability to reproduce experimental evidence for different systems, predicting all known polymorphs with low energy ranking. However, such an assessment is complicated by the practical unfeasibility of conducting exhaustive experimental "polymorph screening" programs. While it is always possible to recognise that a CSP approach has failed to identify an experimental structure or to find its correct stability rank, it is harder to draw conclusions when it predicts structures that have *not* been observed experimentally [16, 17].
- It must take advantage of current state-of-the-art computer hardware and software within practicable cost. There is little benefit in a computationally efficient CSP methodology that is capable of producing results within minutes on a desktop computer if it fails to identify significant low-energy structures. While there is certainly a higher cost in securing access to advanced distributed computing hardware, this is usually negligible compared to the cost of a missed polymorph.

Current methodologies for crystal structure prediction pay varying degrees of attention to the above requirements. In any case, the blind test papers and several recent reviews provide a good overview of current thinking and of the tools that have been developed [18–25].

1.3 The CrystalPredictor and CrystalOptimizer Algorithms

As much of the relevant background is readily available elsewhere, our focus in this chapter is to provide a coherent overview of a CSP methodology that we have been developing over the past 15 years in the Centre for Process Systems Engineering at Imperial College London. Consistent with the principles outlined above, our methodology, algorithms and workflow have been heavily influenced by a systems engineering background and have drawn on experience in developing algorithms and implementing them in large software codes in other areas. We aim to provide a CSP algorithm designer's perspective, setting out the general considerations that need to be taken into account in a manner that can hopefully be of value to designers of future algorithms. The approach presented is one concrete example of what can be achieved given current constraints on underlying software infrastructure (e.g. for quantum-mechanical (QM) calculations) and on computing hardware.

Our work has focused on two general-purpose algorithms and codes, namely *CrystalPredictor* [26, 27] which performs a global search of the crystal energy landscape, and *CrystalOptimizer* [28] which performs a local energy minimisation starting from a given structure. Over the last few years, these algorithms have been applied both by us and more extensively by others to a relatively wide variety of systems including single compound crystals [15, 29–36], co-crystals [14, 37–39],

including chiral co-crystals [40, 41], hydrates and solvates [42, 43]. The codes have also been used separately, e.g. Gelbrich et al. [44] report a recent application of *CrystalOptimizer* to the study of four polymorphs of methyl paraben.

The *CrystalPredictor* algorithm has been in use since the third blind test [4–6], while *CrystalOptimizer* has been available only since the latest (fifth) blind test [4], where it was applied successfully to the prediction of the crystal structure of target molecule XX [9], one the largest and most flexible molecules considered in a blind test to date. Both codes have been evolving continually in terms both of the range of systems to which they are applicable and of their computational efficiency.

1.4 Structure of Chapter

Section 2 of this chapter reviews the main considerations that need to be taken into account in the design of CSP algorithms. Based on this background, Sect. 3 provides a description of the key elements of our methodology in its most recent form. Finally, Sect. 4 seeks to draw some general conclusions based on the experience gathered from a fairly consistent application of this methodology across a relatively wide range of systems over the last few years. In particular, we consider the limitations of our current approach and identify areas of further work that are needed to address them. We also consider issues relating to the use of empirical models derived from experimental data in conjunction with ab initio CSP.

2 Key Considerations in the Design of CSP Algorithms

2.1 Mathematical Formulation of the CSP Problem

A crystal formed from one or more chemical species is a periodic structure defined in terms of its space group, the size and shape of the unit cell, the numbers of molecules of each species within the unit cell and the positions of their atoms. For example, Fig. 2 shows the unit cell of crystalline Form II of piracetam ((2-oxo-1-pyrrolidinyl)acetamide). In this case, there is only one molecule per unit cell, and the crystal structure is also characterised by the Cartesian coordinates of the atoms within this cell. For the purposes of this chapter, we are interested in systems that extend practically infinitely in each direction and are free of all defects.

The crystal structures of practical interest are those which are stable or metastable at the given temperature, pressure and composition; as such they correspond to local minima in the free energy surface with relatively low values of the Gibbs free energy, G, which can be expressed as:

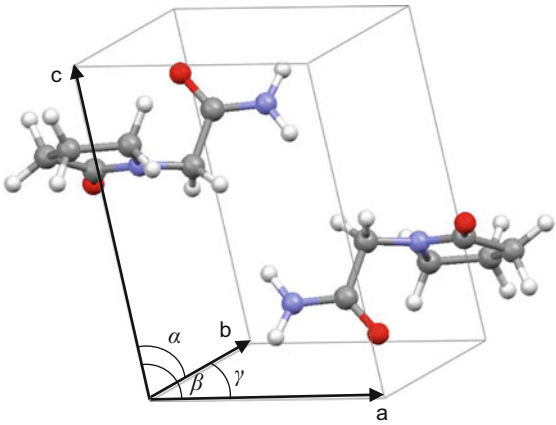


Fig. 2 Lattice vectors (**a**, **b**, **c**) and angles (α, β, γ) defining the unit cell in the Form II crystal of piracetam [45]

$$\min G = U + pV - TS \tag{1}$$

where U denotes the internal energy of the crystal, p the pressure, V the volume, T the temperature and S the entropy on a molar basis. The minimisation is carried out with respect to the variables defining the crystal structure as listed above.

The entropic contribution $-TS$ is typically omitted in the context of CSP as it is difficult to compute reliably and at low computational cost for systems of practical interest. The magnitude of this term is expected to be small compared to the enthalpic contribution at the relatively low temperatures of interest [46]; on the other hand, omission of the term is often cited as one of the possible reasons for failing to predict experimentally observed structures accurately. In any case, any predictions made by CSP methodologies making use of this simplification in principle relate to a temperature of 0 K.

The work term $+pV$ is also often omitted from the free energy expression. It is worth mentioning that, in contrast to the $-TS$ term, this term can be computed with negligible cost, and is sometimes important for predictive accuracy at high pressures.

Based on the above approximations, the energy function used to judge stability of a crystal structure is usually reduced to the lattice internal energy U, typically computed with reference to the gas-phase internal energy U_i^{gas} of the crystal's constituents i:

$$\min \Delta G \cong U - \sum_i x_i U_i^{\text{gas}} \tag{2}$$

where x_i is the molar fraction of chemical species i in the crystal structure. Posing the CSP problem in this manner reduces it to two important sub-problems, namely the accurate computation of this energy for a proposed crystal structure and the reliable identification of all local minima, or at least those with relatively low energy values. We consider these in more detail in the two sections below.

2.2 Accurate Computation of Lattice Energy

In principle the lattice energy can be computed through QM computations, as is the case in periodic solid-state density functional theory approaches, e.g. [47, 48]. However, such an approach is computationally very demanding, to an extent that may currently limit its applicability with respect to the size of the system to which it can be applied successfully; its theoretical rigour is also somewhat compromised by the need to use an empirical model of dispersion interactions. The alternative is the "classical" approach to computing lattice energy which distinguishes intramolecular and pair-wise intermolecular contributions, with the latter being further divided into repulsive, dispersive and electrostatic terms. Moreover, starting with a reference unit cell, one has to add up the interactions of its molecules with those in all other cells within an infinite periodic structure.

Most organic molecules of interest to CSP have a non-negligible degree of molecular flexibility which allows them to deform in the closely packed crystalline environment. In turn, the deformation induces changes to their intramolecular energy, but also to two other aspects that affect intermolecular interactions within the crystal, namely the relative positioning of the atoms in the molecule and their electronic density field. Overall, then, stable/metastable crystal structures represent a trade-off between the increase in intramolecular energy caused by deformations from in vacuo conformations and the overall energy decrease due to attractive and repulsive intermolecular interactions. This is illustrated in Fig. 3 for xylitol (1,2,3,4,5-pentapentanol) using a model that includes separate contributions to the lattice energy from the intra- and intermolecular interactions (cf. Sect. 3). Intramolecular forces tend to favour larger values of the torsions in the range considered (cf. Fig. 3d where the minimum energy point occurs at the top right corner). On the other hand, intermolecular forces drive torsion angle H1-O1-C1-C2 to a low value, and torsion angle O1-C1-C2-C3 towards an intermediate value of approximately 180° (cf. Fig. 3c where the minimum energy point is near the middle of the left vertical axis). These opposite effects are of similar magnitudes, resulting in the torsions adopting intermediate values in the experimentally observed conformation (cf. Fig. 3b).

The classical approach to lattice energy computation is common to most current CSP approaches. Notwithstanding the approximations that are already inherent in the classical calculations, what is not always appreciated is the very significant extent to which even relatively small inaccuracies in them affect the quality of crystal structure predictions, especially when considering relative stability rankings as a measure of success. Potential pitfalls include:

- Inaccuracies in Intramolecular Energy Calculation

These may arise either from failing to take account of all the conformational degrees of freedom that are substantially affected by the crystalline environment, or from approximations in the calculation of the intramolecular energy for a given conformation (e.g. via the use of inappropriate empirical force fields).

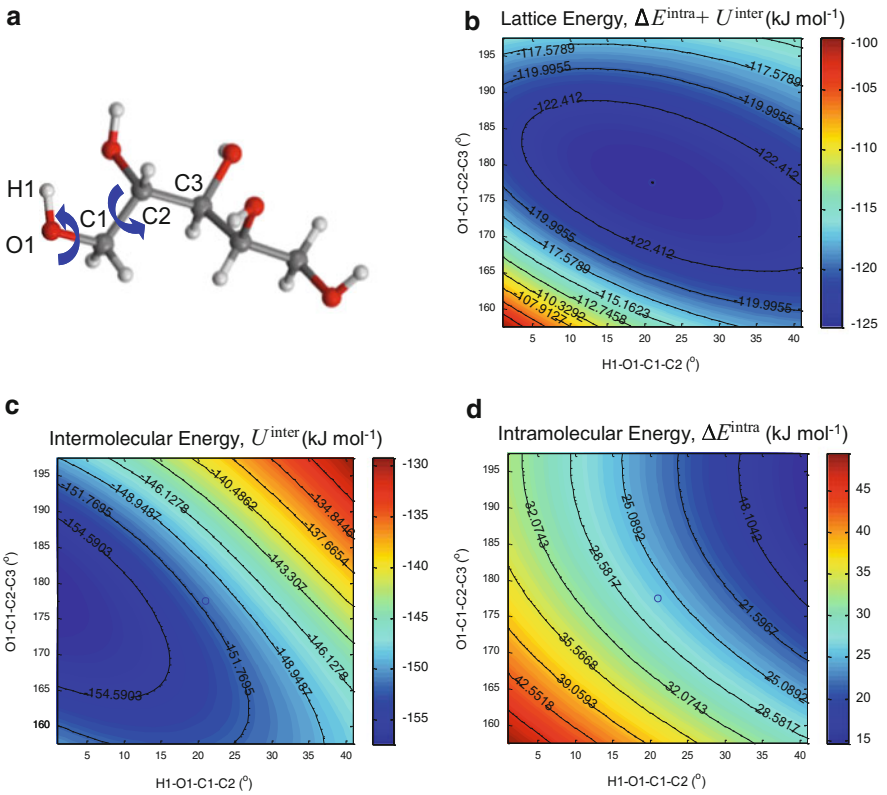


Fig. 3 Effect of conformational flexibility on the energetics of xylitol. (**a**) Molecular conformation of xylitol in the experimental crystal structure [49], with *blue arrows* denoting the two torsional angles being considered here. (**b**) Lattice energy map as a function of the two angles. (**c**) Intermolecular energy map. (**d**) Intramolecular energy map. The *open circle* on each map denotes the values of the torsions in the experimentally observed crystal

- Inaccuracies in Pairwise Intermolecular Interactions—Electrostatic Contributions

Analysis of blind test results (for example for Molecule VIII [19]) indicates that partial charges do not provide a sufficiently accurate representation of the electrostatic field, and one has to resort to more complex alternatives such as off-centre charges [50] or distributed multipoles [51, 52]. These classical electrostatic descriptions are often derived from gas-phase isolated-molecule QM calculations and therefore ignore the effects of polarisability, which can sometimes lead to inaccurate ranking, especially for polar crystals. Approaches aiming to address this issue include the use of gas-phase calculations on dimers [53], or of continuum polarisable models for the isolated-molecule calculations [54]. Developments in more accurate atom-atom potentials also hold promise in this area [55, 56].

- Inaccuracies in Pairwise Intermolecular Interactions—Dispersion/Repulsion Contributions

Given the difficulty in their ab initio computation, the contributions of repulsion/dispersion interactions are usually computed via empirical potentials fitted to experimental data [57–63]. A potential pitfall in this context is that the values of the repulsion/dispersion potential parameters derived from such an exercise depend on what *other* terms are included in the lattice energy (e.g. intramolecular and/or intermolecular electrostatic contributions) and on precisely how each such term is computed (e.g. whether electrostatic contributions are accounted for in terms of partial charges or distributed multipoles, and the level of theory employed in the QM isolated-molecule calculations used to derive these partial charges/multipoles). For example, the commonly used parameters from [60, 61, 63] were estimated assuming perfectly rigid molecules, with electrostatic interactions computed via atomic charges derived from HF/6-31G** QM calculations. Therefore, these parameter values are not necessarily consistent with more recent CSP techniques that take account of molecular flexibility and/or employ distributed multipoles derived from QM computations at much higher levels of theory.

- Errors in Summation of Intermolecular Interactions Over Infinite Periodic Structures

The importance of efficiently and accurately computing these summations is generally well understood, and techniques such as Ewald summations [64] are routinely used to calculate conditionally convergent electrostatic sums such as charge–charge interactions. However, the quality of practical implementations varies widely. For example, cut-off distances for determining which terms to include in these summations are often set to inappropriately low values, and/or are applied to distances between centres of mass (rather than individual atoms) of the molecules involved – even when the size of the molecule is a significant fraction of the cut-off distance itself; in the latter case, at least some of the terms omitted from the summation relate to pairs of atoms that are much closer to each other than centre-of-mass distances suggest.

2.3 Identification of Local Minima on the Lattice Energy Surface

Addressing the issues identified above is clearly important for ensuring an accurate calculation of the lattice energy. The next area of concern is ensuring that the crystal structures predicted are local minima on the energy surface. This may not be the case if the optimisation algorithm used for energy minimisation converges to points that are not true local minima. Such failures may be caused by using algorithms, such as simplex [65], which do not make use of the values of the partial derivatives of the energy with respect to the crystal structure decision variables, and

consequently exhibit slow convergence. More recent work had tended to avoid this problem by using gradient-based optimisation algorithms [66]; nevertheless, failure may still occur because of inaccurate values of these partial derivatives (e.g. when they are approximated via finite difference perturbations).

Ensuring that any crystal structures obtained are true local minima does not necessarily guarantee that *all* such structures of practical relevance are identified. The standard approach for identifying multiple local minima is based on generating a large number of structures which are used as initial points for local energy minimisation along the lines described above. Mathematically, it can be shown that such an approach is guaranteed to identify all local minima provided an infinite number of initial points are generated in a manner that sufficiently covers the space of decision variables. The more practical question is how many structures need to be generated in order to provide a reasonably high probability of identification of all structures of interest. Some relevant insight is provided in Fig. 4 which shows the local minima identified during the global search phase for the ROY molecule [36]. Even for such a relatively small molecule, there are several thousands of local minima, many hundreds of which would be of interest as potential starting points for a refinement using a more accurate lattice energy model. Given that there is currently no technique which can selectively and directly identify only relatively low-energy structures, it seems that the desired degree of reliability in CSP can be achieved only by generating very large numbers (in the order of tens or hundreds of thousands) of candidates.

Insufficient exploration of the space of possible crystal structures may also arise in more subtle ways as a result of the introduction of artificial constraints during the global search. A common pitfall is to base the search on a finite number of rigid molecular conformations generated a priori by fixing some of the key flexible degrees of freedom (e.g. torsion angles) to specific sets of values. This "*multiple rigid-body searches*" approach avoids the need to handle molecular flexibility during the global search. However, whilst this approach can be successful in specific cases (cf. the "RCM" algorithm reported in [9]), its outcome is highly dependent upon the specific choice and indeed the total number of rigid structures tested; for highly flexible molecules, comprehensive coverage of the crystal structure space may be achievable only via a very large number of global searches, each based on a different rigid conformation. Moreover, taken together, these rigid-body global searches may result in many more unique structures than a single flexible search: two or more neighbouring but ostensibly distinct local minima may relax into a single one if the molecules are allowed to deform continuously under the intermolecular forces exerted on them. Not taking advantage of this relaxation effect during the global search stage invariably results in a higher number of structures that need to be analysed at the refinement stage, and consequently a higher computational load.

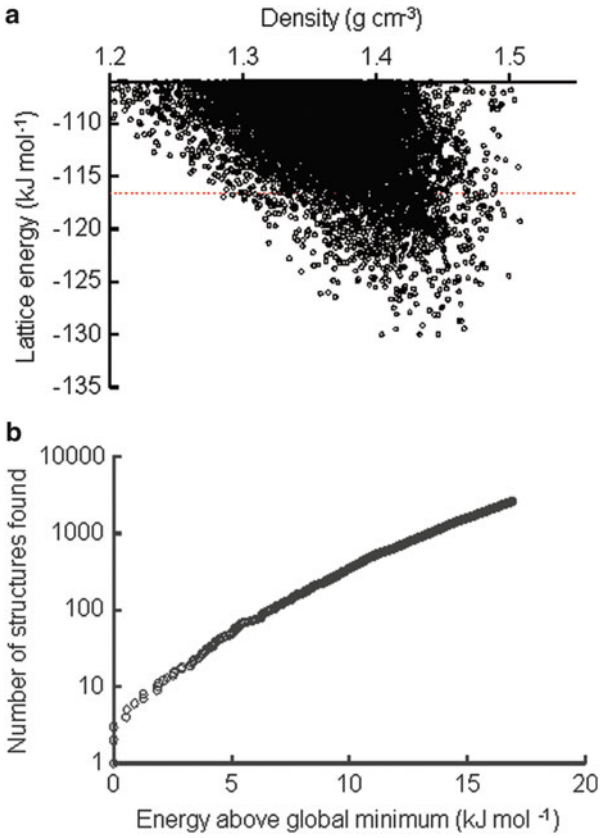


Fig. 4 Local minima of lattice energy surface for ROY molecule (5-methyl-2-[(2-nitrophenyl) amino]-3-thiophenecarbonitrile, [67]) identified by global search. (**a**) Energy vs density diagram; each point corresponds to a unique local minimum on the lattice energy surface. (**b**) Cumulative number of unique local minima identified vs energy difference from the global minimum

2.4 Implications for CSP Algorithm Design

The analysis presented above suggests that taking shortcuts in the accurate calculation and minimisation of crystal energy in an attempt to reduce computational complexity may be detrimental to the quality of the prediction, as are attempts to sample only a small part of the decision space (e.g. by using only hundreds or thousands of initial points in the global search). Such "savings" may prove highly counter-productive in applications (e.g. in the pharmaceutical industry) where failing to identify a low-energy polymorph or identifying too many fictitious ones can have serious implications. Accordingly, one needs to aim for algorithms that attempt to maximise reliability of prediction within currently available computational power.

The challenge for the CSP algorithm designer is how to reconcile the need for very accurate evaluation of energy and its partial derivatives for the purposes of local minimisation of lattice energy, with the extremely large number of such minimisations that have to be carried out during the search for low-energy structures. A practical way of achieving this is via a two-stage procedure where the *global search* is performed using a relatively simpler and computationally less expensive energy model. This allows a much smaller number of promising structures to be identified which can then serve as starting points for *refinement* via local minimisation using a more detailed model.

The two-stage approach to ab initio CSP is illustrated schematically in Fig. 5. It takes as input the stoichiometry of the crystalline phase and the molecular connectivity diagrams for the relevant chemical species, and produces as output the crystal structure with the lowest (globally minimum) lattice energy as well as other crystal structures that correspond to local lattice energy minima with energy values close to the global minimum.

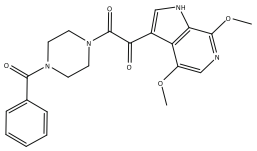
In practice it is usually necessary to have an additional "Stage 0" dedicated to the study of each individual species in order to:

- Identify important aspects of its molecular flexibility (e.g. the set of torsional angles that are likely to undergo significant deformation in the crystalline environment, and the likely range of any such deformation).
- Determine an appropriate level of theory of QM calculations (e.g. via comparison with any available experimental data on its gas-phase conformation or any already known polymorphs for crystals formed by it).

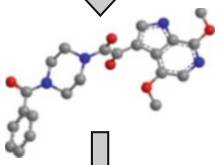
In some CSP methodologies the information necessary for computing intramolecular energy and/or intermolecular electrostatic contributions during Stages 1 and 2 is also generated via QM calculations during this Stage 0. Alternatively, these QM calculations may be performed "on-the-fly" when necessary during Stages 1 and 2 (see Sect. 3.3).

The multistage approach to CSP has been widely adopted [18, 25, 38, 68] and has been successfully used in the blind tests of crystal structure prediction [3–7, 9]. Its success hinges on the hypothesis that relatively simple models of the energy surface can provide energy minima whose geometry is in reasonably good agreement with that of energy minima on a more accurate surface – even if the actual energy values differ significantly, in both absolute and relative terms, between the simpler and the more rigorous models. The approach comes with its own potential pitfalls: for example, using too simplistic an energy model at the global search phase may result in some of the structures of interest either being missed altogether or being ranked so high in crystal energy that they are not selected for subsequent refinement. Therefore, the global search phase also exhibits an accuracy vs computational cost trade-off, and the way the balance between these two is struck differs significantly between algorithms.

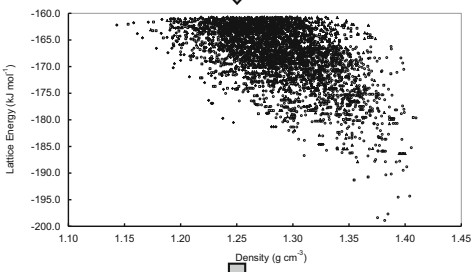
Input Crystal stoichiometry and molecular connectivity diagrams for chemical species



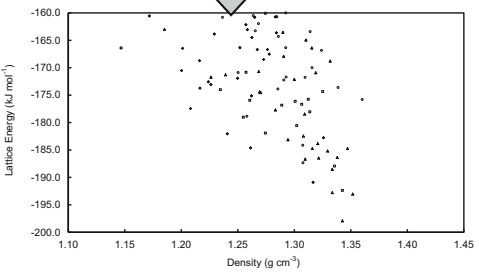
Stage 0 Isolated molecule studies



Stage 1 Global search for local minima in lattice energy surface



Stage 2 Refinement of low-energy structures identified at Stage 1



Output Likely polymorphs

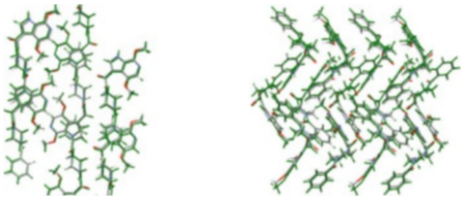


Fig. 5 Multistage approach for CSP, illustrated for molecule BMS-488043 (cf. Fig. 1b)

3 The CrystalPredictor and CrystalOptimizer CSP Algorithms

The CrystalPredictor and CrystalOptimizer algorithms (cf. Sect. 1.3) are aimed, respectively, at the global search and refinement stages of the general methodology described in Fig. 5. They are designed to be applicable to crystal structures which belong, in principle,[1] to any space group and which involve any number of chemical species of the same or different types within the asymmetric unit.

Based on the analysis presented in Sect. 2, and in order to ensure the maximum degree of consistency between the two algorithms, their overall design philosophy can be summarized as follows:

- In CrystalOptimizer, use the highest degree of accuracy in lattice energy computation that can be practically deployed at the refinement stage.
- In CrystalPredictor, apply the above subject to the *minimal* set of simplifications that are necessary to accommodate the additional computational complexity of the global search.

Inevitably, the practical implications of these general principles have been changing over the years, reflecting advances in our ability to describe efficiently and accurately various terms in the energy function. In this section we discuss the current state of the algorithms and their implementation in computer code.

3.1 Molecular Descriptions

The description of the molecular conformation is a key element of any CSP methodology. In CrystalPredictor and CrystalOptimizer each chemical entity in the crystal is assumed to be flexible with respect to all conformational degrees of freedom (CDFs), including torsion angles, bond angles and bond lengths.

In general, we divide the CDFs into two different sets[2]:

- The independent CDFs, θ, are those which are affected directly by intermolecular interactions in the crystalline environment.
- The dependent CDFs, $\overline{\theta}$, always assume values that minimise the intramolecular energy of an isolated molecule for given values of θ; therefore $\overline{\theta} = \overline{\theta}(\theta)$.

By spanning the whole range from an empty set θ (i.e. a rigid molecule calculation) to an empty set $\overline{\theta}$ (i.e. a fully atomistic computation), the above

[1] See Sect. 3.5.2 for details of the current implementation.

[2] In fact, the algorithms also recognise a third class of CDFs which can be fixed at user-provided values (e.g. in order to exploit a priori available experimental information in performing more targeted searches). However, in the interests of clarity of presentation, we omit this complication from the mathematical descriptions provided in this chapter.

partitioning provides a mechanism for adjusting the number of degrees of freedom that have to be manipulated during the energy minimisation (see Sect. 3.2). More specifically, it allows us to employ different degrees of molecular flexibility between the global search and refinement stages.

In earlier applications of our CSP methodology, θ would typically include only the more flexible torsion angles while $\overline{\theta}$ would comprise the remaining torsion angles, as well as bond angles and lengths. However, with increasing computational capability and more efficient ways of computing intramolecular energy contributions (see Sect. 3.3), one can afford to shift the balance from $\overline{\theta}$ to θ, taking direct account of a wider range of torsion angles and even some bond angles, especially during the refinement stage. A number of examples employing extended sets of independent CDFs θ, including fully atomistic computations, were reported in [28].

3.2 The Lattice Energy Minimisation Problem

The lattice energy minimisation problem is formulated in terms of the following independent decision variables:

- The unit cell lattice lengths and angles, collectively denoted by X
- The positions of the centres of mass and the orientation of the chemical entities within the unit cell, collectively denoted by β
- The independent CDFs, θ, of the chemical entities

As already mentioned, the dependent CDFs $\overline{\theta}$ can be computed as functions of the independent ones, i.e. $\overline{\theta}(\theta)$ via minimisation of the intramolecular energy, i.e.

$$\overline{\theta}(\theta) = \arg\min_{\overline{\theta}} U^{\text{intra}}(\overline{\theta}, \theta) \tag{3}$$

carried out as an isolated-molecule QM calculation. The latter also produces the information necessary for deriving an appropriate finite-dimensional description $Q(\theta)$ of the molecule's electrostatic field in terms of charges or distributed multipoles [52, 69]. The CDFs θ and $\overline{\theta}$ can also be used in conjunction with the molecular positioning variables β to determine the Cartesian coordinates Y of all atoms within a central unit cell, i.e. $Y = Y\left(\theta, \overline{\theta}, \beta\right)$. Finally, Y together with the unit cell parameters X determines the atomic positions in all periodic images of the central unit cell, which are required for the calculation of intermolecular energy contributions.

Overall, the lattice energy minimisation problem in both CrystalPredictor and CrystalOptimizer is formulated mathematically as[3]

[3] The actual implementations also include the $+pV$ term in the objective function which, therefore, corresponds to lattice enthalpy. However, in the interests of simplicity of presentation, this is omitted here and in subsequent discussion.

$$\min_{X,\beta,\theta} U(X,\beta,\theta) \equiv \Delta U^{\text{intra}}\left(\theta,\overline{\theta}\right) + U^{e}(Q,Y,X) + U^{\text{rd}}(Y,X) \tag{4}$$

where ΔU^{intra} represents the intramolecular energy contribution (after subtraction of the gas-phase internal energy of the chemical species in the crystal) and U^{e} and U^{rd} represent the intermolecular electrostatic and repulsion/dispersion contributions. Note that, in the interests of clarity, the above expression does not show explicitly the direct and indirect functional dependence of the quantities $\overline{\theta}, Q, Y$ on the independent decision variables X, β, θ.

3.3 Accounting for Molecular Flexibility During Lattice Energy Minimisation

The evaluation of the intramolecular contribution $\Delta U^{\text{intra}}\left(\theta,\overline{\theta}\right)$ in the above objective function can be done via a standard QM minimisation of configurational energy at given (fixed) values of θ. In general, such isolated-molecule calculations can provide the accuracy required for modelling the deformation of the molecular structure and energy within the crystal [70], although neglecting intramolecular dispersion can lead to inaccuracies for highly flexible molecules [71].

In principle this QM calculation could be embedded directly within the overall energy minimisation algorithm, as implemented in the DMAFlex algorithm [72]. This has the added advantage of also producing consistent values of the dependent CDFs $\overline{\theta}$, thereby allowing correct evaluation of atomic positions Y within the central unit cell, and consequently of the interatomic distances that are needed for the correct calculation of intermolecular contributions U^{e} and U^{rd} at each iteration. It also allows the derivation of consistent electrostatic descriptions Q which are also needed for the accurate evaluation of the intermolecular electrostatic contributions, U^{e}.

The obvious difficulty that arises from embedding an expensive QM calculation within an iterative optimisation procedure is one of computational cost, and this severely limits the number of independent CDFs that can be handled in practice. An alternative would be to replace the QM calculations by molecular mechanics intramolecular potentials (cf. the use of the DREIDING and COMPASS potentials in the RCM approach reported in [9]). Such techniques can approximate the effects of θ on ΔU^{intra} to a varying degree of accuracy; however, they do not take account of the secondary effects on intermolecular contributions arising from the effects of θ on $\overline{\theta}$ and Q. Overall, there is some doubt regarding the suitability of such models for CSP [19, 68].

A different way of addressing the above difficulties is via the use of pre-constructed interpolants for ΔU^{intra} (and, in principle, $\overline{\theta}$ and Q) based on a multi-dimensional grid of values of θ. An example of such an approach was the restricted multidimensional Hermite interpolants used in earlier versions of

CrystalPredictor [27]. However, the size of the required grid effectively imposed a limit on the number of independent CDFs that could be handled to typically 3–6 torsion angles depending on the complexity of the molecule(s) under consideration. For molecules exhibiting higher degrees of flexibility, one had to resort to artificial approximations, such as grouping the flexible torsion angles into multiple, supposedly non-interacting groups, and then constructing the interpolants using QM calculations based on surrogate simpler molecules, each involving a different group of torsions. For example, in the case of molecule XX of the fifth blind test, the six flexible torsion angles considered during the global search were decomposed into two "independent" groups (of three angles each) located at either end of the molecule; the QM calculations were then performed using two simpler surrogate molecules derived from the original molecule (cf. Sect. 2.1 in [9]). While in this case the global search was ultimately successful in identifying a structure corresponding to the experimentally observed crystal, in general such an approach is cumbersome, involves elements of subjective judgment and may not be applicable in cases where the torsion angles interact more closely with each other; all these factors make it less than ideal, especially in the context of the general principles and requirements set out in Sect. 1.2.

In view of the above, the approach used in the most recent versions of our algorithms is based on Local Approximate Models (LAMs) [28]. LAMs are essentially multidimensional quadratic Taylor expansions of the functions $\Delta U^{\text{intra}}(\theta) \equiv \min_{\overline{\theta}} U^{\text{intra}}\left(\overline{\theta},\theta\right) - U^{\text{gas}}$ and $\overline{\theta}(\theta) \equiv \arg\min_{\overline{\theta}} U^{\text{intra}}\left(\overline{\theta},\theta\right)$, and multidimensional linear Taylor expansions of the functions $Q^*(\theta) \equiv Q\left(\overline{\theta},\theta\right)$. As their name implies, they are local approximations constructed around certain points $\theta^{[1]}$, $\theta^{[2]}$, $\theta^{[3]}$ ... in the space of the independent CDFs θ. Because of the continuity and differentiability of the functions $\Delta U^{\text{intra}}(\theta), \overline{\theta}(\theta), Q^*(\theta)$, each LAM can be guaranteed to be accurate within a required tolerance within a finite non-zero volume surrounding the point at which it was derived. Consequently, in principle the entire θ-domain of interest can be covered with a finite number of LAMs. In practice the range of applicability of LAMs for each molecule of interest is determined based on test calculations at Stage 0 of the methodology of Fig. 5.

The use of LAMs can provide accurate values of $\Delta U^{\text{intra}}(\theta)$, $\overline{\theta}(\theta)$ and $Q^*(\theta)$ for the computation of the lattice energy function at minimal cost. It may also potentially improve the performance of the optimisation algorithm as LAMs are not subject to the numerical noise that may arise because of the iterative nature of the QM calculations. However, certain adjustments need to be made to the optimisation algorithm to account for the discontinuities that may arise as the iterates move from one LAM to a neighbouring one.

LAMs were originally introduced in the context of CrystalOptimizer [28]. In this case the θ-domain of interest cannot normally be determined a priori, and therefore the sequence of Taylor expansion points $\theta^{[1]}$, $\theta^{[2]}$, $\theta^{[3]}$ ... is determined "on-the-fly" during the optimisation iterations. This is illustrated schematically in Fig. 6 for a hypothetical molecule involving two independent CDFs, θ_1 and θ_2. Once a LAM is derived, it is kept in memory even if the optimisation iteration moves out of its

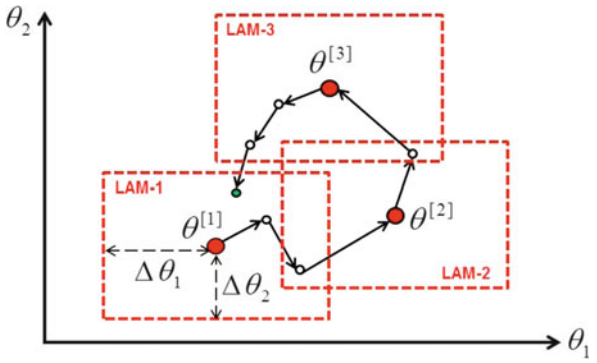


Fig. 6 Use of LAMs during lattice energy minimisation by CrystalOptimizer for a molecule involving two independent CDFs θ_1 and θ_2. The *points* and *solid arrows* indicate the progress of the optimisation iterations in the two-dimensional [θ_1, θ_2] domain. *Large red circles* indicate points at which new LAMs have to be derived, while the *smaller circles* indicate other iterates at which an existing LAM can be used. The *dashed rectangles* indicate the limits of applicability of each LAM; these are usually expressed in terms of ranges $\pm$ $\Delta\theta$ which are established at Stage 0 of the procedure in Fig. 5

range of applicability; this allows the LAM to be re-used should the optimisation iterates return to within range at a later state of the optimisation iterations (cf. the green point in Fig. 6). Moreover, at the end of the calculation, the relevant QM results that have been used to derive LAMs are stored in persistent storage ("LAM databases"), thereby allowing them to be re-used in future CSP calculations involving this particular molecule.

The introduction of LAMs in CrystalOptimizer over the last 3 years has significantly increased the range of molecular flexibility that can be handled from only a few (typically no more than six) torsional angles to large numbers of torsion and bond angles and indeed all the way to fully atomistic calculations [28]. For example, the successful prediction of molecule XX in the fifth blind test involved treating 14 torsion angles and 5 bond angles as independent CDFs (cf. the "FCC" approach reported in [9]).

The benefits realized from the use of LAMs in CrystalOptimizer and also the experience gained with the application of earlier versions of CrystalPredictor to the global searches undertaken in the context of the fifth blind test [4, 9] and other challenging systems [36] have motivated the introduction of LAMs in CrystalPredictor [27]. In this case, the θ-domain of interest is known a priori since the global search algorithm (see Sect. 3.5.3) will, by design, cover the entire allowable space of θ, as well as those of the other optimisation decision variables X and β. Therefore, in this case there is no advantage in computing the LAMs on-the-fly during the search; instead, it is more efficient to compute them before the start of the global search based on a regular grid, as illustrated in Fig. 7.[4] This recent

[4] As already mentioned, these LAMs can be stored in persistent LAM databases to be re-used in later calculations, such as those required for the subsequent refinement stage.

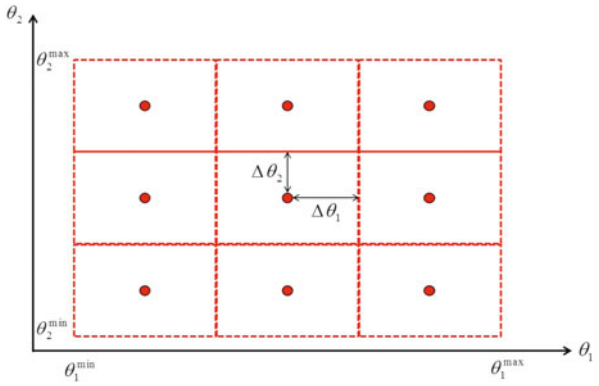


Fig. 7 LAMs for use by global search in CrystalPredictor for a molecule involving two independent CDFs θ_1 and θ_2. LAMS are derived at points (indicated by the *large red circles*) placed on a regular grid defined over the θ-domain of interest $[\theta_1^{min},\theta_1^{max}] \times [\theta_2^{min},\theta_2^{max}]$. The *dashed rectangles* indicate the limits of applicability of each LAM; these are usually expressed in terms of ranges $\pm\ \Delta\theta$ which are established at Stage 0 of the procedure in Fig. 5

development has made it possible to consider much higher degrees of molecular flexibility during the global search without the need for ad hoc approximations such as the molecular decomposition described earlier.

3.4 Intermolecular Contributions to the Lattice Energy

Both CrystalPredictor and CrystalOptimizer calculate the intermolecular electrostatic contributions to the lattice energy using finite representations of the electrostatic potential determined via isolated-molecule QM computations (cf. Sect. 3.3). The main difference between the two codes is in the form of this finite representation. In the interest of computational efficiency during the global search, CrystalPredictor employs simple charges located at the atomic positions. On the other hand, in order to ensure higher accuracy during the crystal structure refinement stage, CrystalOptimizer makes use of distributed multipoles, placing an expansion comprising charge, dipole, quadrupole, octupole and hexadecapole terms at each atomic position. The expansion is derived directly from the isolated molecule wavefunction [52] using the GDMA [69] program. Distributed multipole moments have been shown to be successful in predicting the highly directional (anisotropic) lone-pair interactions, π–π stacking in aromatic rings and hydrogen bond geometries in molecular organic crystals [73–75].

CrystalPredictor and CrystalOptimizer employ empirical isotropic potentials for the computation of repulsion/dispersion contributions to the lattice energy. The energy contribution arising from a pair of atoms (i,i') located at a distance r from each other is given by the Buckingham potential [76]:

$$U_{ii'}(r) = A_{ii'} e^{-B_{ii'} r} - \frac{C_{ii'}}{r^6} \tag{5}$$

where $A_{ii'}, B_{ii'}, C_{ii'}$ are given parameters. For atoms of the same type (i.e. $i = i'$), the values of the latter are taken from [62]; for unlike pairs ($i' \neq i$), they are computed via the combining rules:

$$A_{ii'} = \sqrt[2]{A_{ii} A_{i'i'}}; \quad B_{ii'} = \frac{B_{ii} + B_{i'i'}}{2}; \quad C_{ii'} = \sqrt[2]{C_{ii} C_{i'i'}} \tag{6}$$

The summations of intermolecular atom-atom interactions between the central unit cell and its neighbouring cells are handled via a combination of direct and Ewald [64] summations.

3.5 The Global Search Algorithm in CrystalPredictor

CrystalPredictor performs a global search by generating very large numbers of structures, each one of which may be used as an initial guess for a local minimisation of the lattice energy function (cf. Sect. 3.2). The key aspects of this algorithm are described below.

3.5.1 Exploitation of Space Group Symmetry

Physically, any crystal structure will have to belong to one of the 230 crystallographic space groups. For a given space group, only a subset of the optimisation decision variables X, β, θ may be independent, while the rest can be determined via space group symmetry relations. In practical terms this means that the global search for this particular space group only needs to explore the space of the independent subset, thereby improving the coverage of the decision space that can be achieved with any given number of candidates.

In its current implementation, CrystalPredictor generates candidate structures in 59 space groups chosen among those most frequently encountered in the CSD. The total number of structures to be generated is specified by the user, and so is the distribution of these structures among the 59 space groups. Typical choices include the numbers of structures generated being either the same for all space groups, or in direct proportion to the space groups' frequency of occurrence in the CSD.

3.5.2 Search Domains for Conformational Variables

The domain of independent CDFs θ (cf. Sect. 3.2) that needs to be searched is an important aspect of the global search algorithm given the complexity and cost

associated with handling the effects of these variables on both intramolecular and intermolecular energy contributions (cf. Sect. 3.3). For example, the size of the domain $[\theta_1^{\min}, \theta_1^{\max}] \times [\theta_2^{\min}, \theta_2^{\max}]$ illustrated in Fig. 7 directly determines the number of LAMs that are needed to cover it, which is an important consideration given the fact that the construction of each LAM requires a computationally expensive isolated-molecule QM calculation.

In view of the above, establishing appropriate ranges of the independent CDFs for each chemical entity that appears in the crystal is an important part of the preliminary conformational analysis carried out at Stage 0 of the algorithm of Fig. 5. Typically, this involves varying each independent CDF θ around its value in the in vacuo conformation of the molecule while keeping all other θ constant at their in vacuo values. The variations that are assumed to be relevant for CSP purposes are those which increase intramolecular energy by up to a given threshold (typically +20 kJ/mol) from its minimum value at the in vacuo conformation.

Overall, the above procedure establishes the range of interest for each independent CDF θ_i in terms of lower and upper bounds $[\theta_i^{\min},\theta_i^{\max}]$. The θ-domain of interest is assumed to be the Cartesian product $[\theta_1^{\min},\theta_1{}^{\max}] \times [\theta_2^{\min},\theta_2^{\max}] \times [\theta_3^{\min},\theta_3^{\max}] \times \ldots$. Theoretically, the one-dimensional scans used to determine this could result in inadvertently excluding certain *combinations* of multiple θ_i that would result in intramolecular energy increases below the specified threshold. However, this has not been found to be a problem in practice; this may be a result of setting the threshold at a conservatively high value.

It is worth noting that, in some cases, the values of interest may belong to multiple ranges that are disjoint from each other, e.g. $[a, b]$ and $[c, d]$ with $c > b$. In such cases, the CrystalPredictor global search is applied separately to each range. If more than one independent CDF has multiple ranges, then the search needs to be applied to each combination of the ranges of these CDFs.

3.5.3 Generation of Candidate Structures

An important decision in any global search algorithm is the precise way in which candidate points are generated over the space of independent variables being searched. Typical choices include creating points on a uniform grid in multidimensional space, or as random samples from a uniform probability distribution (the Monte–Carlo approach). CrystalPredictor [26, 27] makes use of deterministic low-discrepancy sequences [77]. These normally lead to better coverage of the search space for any given number of points being generated.

By way of illustration, Fig. 8 shows 225 points being placed on a two-dimensional search space according to the 3 schemes mentioned above. By construction, the low-discrepancy sequence approach (cf. Fig. 8c) places each new point so as to maximise a measure of distance from all previous points; this leads to better coverage of the domain than that achievable using random samples (cf. Fig. 8b). Moreover, the projection of each point onto each of the axes corresponds to a *distinct* value of each decision variable, i.e. no two points in the

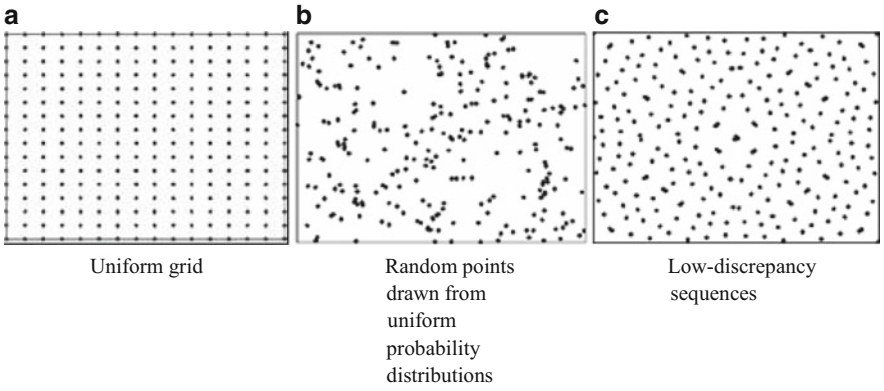


Fig. 8 Different schemes for candidate point generation during global search. (**a**) Uniform grid. (**b**) Random points drawn from uniform probability distributions. (**c**) Low-discrepancy sequences

low-discrepancy sequence in Fig. 8c have the same abscissa or ordinate; in practical terms this means that the search samples 225 distinct values of *each* variable in the search space, as compared with only 15 distinct values in the uniform grid case of Fig. 8a. A further advantage of low-discrepancy sequences over uniform grids is that the final number of candidate points does not have to be decided a priori. Should the initial search be deemed to be insufficient for whatever reason, more points can be added and optimally placed with respect to all previously generated points.

3.5.4 Local Minimisation of Lattice Energy

The crystal structures generated by the approach described in Sect. 3.5.3 are used as starting points for minimisation of the lattice energy function. In practice, before doing this, CrystalPredictor applies a pre-screening based on density, lattice energy and steric hindrance criteria, aimed at eliminating from further consideration any structures that are clearly unrealistic.

The minimisation of lattice energy, subject to the symmetry constraints determined by the space group currently under consideration (cf. Sect. 3.5.1) and simple bounds on the decision variables, is performed via a sequential quadratic programming (SQP) algorithm [66]. For efficiency and robustness, the algorithm makes use of exact first-order derivatives of the objective function and the constraints, determined via analytical differentiation and application of the chain rule on the dependent quantities $\overline{\theta}(\theta), Q(\theta), Y\left(\theta, \overline{\theta}, \beta\right)$.

The CrystalPredictor code is designed to use distributed computing environments involving arbitrarily large numbers of processors for the simultaneous minimisation of multiple structures.

3.5.5 Post-Processing of Generated Structures

The successful execution of CrystalPredictor typically results in a large number of structures, each of which is a local minimum of the lattice energy within a given space group. Given the even larger number of initial candidate structures that are generated and minimised, not all of these final structures will be unique. Accordingly, CrystalOptimizer applies a clustering step intended to remove any duplicate structures among the final set based on their lattice energy, density and interatomic distances.

Finally, because of the space group symmetry constraints, it is possible that a structure is a local minimum only with respect to the space group under which it was generated but only a saddle point as far as the lattice energy surface is concerned. This is assessed by generating the corresponding Hessian matrix of the lattice energy via centered finite differences and evaluating its eigenvalues. For a true local minimum, all of these have to be positive; if the Hessian is found to have one or more zero or negative eigenvalues, then a small perturbation is applied to this structure and it is then used as a starting point for a lattice energy minimisation without any space symmetry constraints. This leads to a lower-energy structure that is a true local minimum on the lattice energy surface.

3.6 Crystal Structure Refinement Via CrystalOptimizer

The crystal structures of lowest energy determined at the end of the CrystalPredictor global search stage (cf. Sect. 3.5.5) are selected for refinement by CrystalOptimizer using a more detailed model of lattice energy. One common criterion for determining whether a given structure is to be refined is based on the difference between the structure's lattice energy and the globally minimum lattice energy determined during the search, with typical cut-off points being placed at around +10–20 kJ/mol. Alternatively, a fixed number of structures (e.g. the lowest 1,000) may be chosen for refinement. Inevitably, there is a degree of subjective judgment in both of the above criteria, the overall objective being to apply refinement to the minimum possible number of structures but without leaving out any polymorphs that are likely to occur in nature. In general, the number of structures that need to be refined becomes lower as more physical detail is added to the lattice energy computation during the global search.

At the fundamental level, CrystalOptimizer employs a very similar lattice energy description as CrystalPredictor with two important differences:

- Molecular flexibility: both algorithms employ the concept of partitioning CDFs into independent θ and dependent $\overline{\theta}$ (cf. Sect. 3.1) and the LAMs described in Sect. 3.3. However, in order to achieve higher accuracy, CrystalOptimizer calculations typically involve more independent and fewer dependent CDFs.

Because of the use of LAMs, the incremental computational cost is usually acceptable, especially given the relatively few structures that have to be refined.

- Intermolecular electrostatic interactions: as has already been stated, CrystalPredictor uses atomic charges while CrystalOptimizer employs distributed multipole expansions.

At the implementational level, the minimisation of lattice energy in CrystalOptimizer is unconstrained: there is little advantage in explicitly enforcing space group symmetry constraints in order to reduce the number of independent decision variables during the optimisation. The space groups of the final structures can be determined by a posteriori analysis using tools such as PLATON [78].

Finally, CrystalOptimizer poses the lattice energy minimisation as a bilevel optimisation problem of the form

$$\min_{\theta}\left(\Delta U^{\text{intra}}\left[\theta,\overline{\theta}(\theta)\right]+U_{*}^{\text{inter}}\left[\theta,\overline{\theta}(\theta),Q(\theta)\right]\right) \tag{7}$$

where the function U_{*}^{inter} is the intermolecular energy corresponding to the *minimum* lattice energy crystal incorporating rigid molecule(s) described by CDFs $\theta,\overline{\theta}$ and distributed multipole expansions Q, i.e.

$$U_{*}^{\text{inter}}\left[\theta,\overline{\theta},Q\right]\equiv\min_{X,\beta}\left(U^{e}\left[Q,Y\left(\theta,\overline{\theta},\beta\right),X\right]+U^{\text{rd}}\left[Y\left(\theta,\overline{\theta},\beta\right),X\right]\right). \tag{8}$$

Thus, the bilevel optimisation problem comprises:

- An outer optimisation problem in the independent CDFs θ
- An inner optimisation problem in the unit cell parameters X and molecular positioning variables β.

CrystalOptimizer employs a quasi-Newton algorithm for the solution of the outer problem, and the DMACRYS code [79, 80] for the solution of the inner problem. The partial derivatives of the function U_{*}^{inter} are obtained via centered finite differences.

The application of the refinement algorithm to the structures selected at the end of the global search stage may result in the same structure being generated more than once. This often arises because the additional molecular flexibility taken into account by CrystalOptimizer allows two or more structures identified by CrystalPredictor as being distinct to relax into the same structure. Accordingly, a clustering algorithm based on the root mean square deviation in the 15-molecule coordination sphere [81] is applied to eliminate any crystallographically identical structures. This finally leaves the list of distinct structures that are reported to the user in ascending order of lattice energies as potential polymorphs.

4 Concluding Remarks

The methodology presented in this chapter is applicable to the prediction of a wide range of crystal structures of organic molecules, including those involving highly flexible molecules and containing multiple molecules (of the same or different types) or ions in the asymmetric unit.

Based on a two-stage global search/refinement approach, the methodology incorporates some major recent advances such as the use of low-discrepancy sequences for the systematic coverage of the space of decision variables during the global search, the efficient and accurate handling of molecular flexibility during both the global search and the refinement stages via LAMs and the accurate description of electrostatic interactions via distributed multipole expansions at the refinement stage. The CrystalPredictor and CrystalOptimizer codes are based on a careful implementation of these ideas, together with efficient numerical optimisation algorithms and exploitation of modern distributed computing resources.

4.1 Predictive Performance of CSP Methodology

There is currently a growing body of experience (cf. the references mentioned in Sect. 1.3) on the performance of these codes on a range of systems; some of this experience has been gained under blind test conditions. It may be worth noting in this context that, as the codes have been evolving over the last decade, results reported in different publications may have been obtained with different versions. However, an improvement in applicability and predictive accuracy is clearly discernible over this period, and we have now reached a point where, for example, we can usefully study molecules of relevance to the pharmaceutical or agrochemical industries.

Although the predictive performance of the methodology varies from one case to another, we believe the following statements to be a fair general representation of the current state of the technology to the extent that this has been explored both by us and by others:

[S1] Experimentally observed crystal structures are generally identified successfully.

[S2] In general, the accuracy of structure reproduction is reasonably good for crystals involving a single chemical species, and less good for co-crystals, salts and hydrates.

[S3] Experimentally observed structures are generally predicted to have low rank (i.e. high relative stability).

[S4] For systems where multiple crystal structures have been identified experimentally (cf. the ROY molecule [36]), the predicted stability ranking is not always correct.

[S5] Low-energy structures that have not (yet) been identified experimentally are often also reported, and some of them may be more stable than the experimentally observed ones.

4.2 Errors and Approximations in CSP Methodology

At the fundamental level, our CSP approach incorporates a number of approximations, including:

- The use of a lattice enthalpy[5] criterion, i.e. the omission of entropic contributions from the Gibbs free energy.
- The separation of lattice energy into intramolecular and intermolecular electronic and repulsive/dispersive contributions.
- The calculation of the intermolecular contributions as sums of pairwise interactions.
- The use of finite descriptions of electronic charge density based on isolated molecule calculations, and not taking account of polarisability effects.
- The use of empirical isotropic repulsion/dispersion potentials.

At a less fundamental, but potentially also important, level the application of the methodology to a particular system may be subject to practical limitations relating to:

- The level of theory of QM calculations that can be employed for a given chemical species within practical computational limits.
- The partitioning between independent and dependent CDFs.
- The use of empirical repulsion/dispersion potential parameters that were estimated from experimental data using molecular descriptions and lattice energy models which were different to those used by our methodology (e.g. in accounting for molecular flexibility, in the description of electrostatic interactions, and in the QM level of theory); we shall return to consider this issue in more detail in Sect. 4.4.

4.3 The Free Energy Residual Term

Mathematically, we can summarize the discussion of Sect. 4.2 via the following expression for the Gibbs free energy, G, of the crystal structure:

[5] Including the $+pV$ term.

$$G(x) = \hat{G}(x) + \boldsymbol{\mathcal{E}}(x) \tag{9}$$

where x is the set of variables defining the crystal structure, $\hat{G}$ is the free energy approximation that is computed[6] by a CSP methodology and $\boldsymbol{\mathcal{E}}$ is a residual term that combines the errors from all the approximations, both physical and mathematical/numerical, listed in Sect. 4.2.

To date we have not reached firm conclusions regarding the relative importance of these approximations in the context of our methodology and their relation to observations [S1]–[S5]. However, some of these factors (e.g. the effects of polarisability or of anisotropic repulsion/dispersion interactions) have been studied in the CSP literature, and it would be useful to repeat this type of analysis with the more detailed energy model presented here. From the general mathematical and algorithmic perspectives:

- [S1] indicates that the molecular representations (e.g. in terms of flexibility), the nature and extent of the global search and the criteria used for selecting the crystal structures to be refined are generally satisfactory.
- [S2] suggests that, at least for crystals comprising single chemical species and notwithstanding the various approximations listed in Sect. 4.2, the local minima of the computed lattice energy function are close to true minima of the Gibbs free energy. Thus, the local sensitivities (gradients) of the residual term $\boldsymbol{\mathcal{E}}$ with respect to the variables x are likely to be significantly smaller than the gradients of the computed free energy $\hat{G}$, i.e.:

$$\left\| \frac{\partial \boldsymbol{\mathcal{E}}}{\partial x} \right\| \ll \left\| \frac{\partial \hat{G}}{\partial x} \right\| \Rightarrow \frac{\partial G}{\partial x} \approx \frac{\partial \hat{G}}{\partial x}. \tag{10}$$

On the other hand, the term $\frac{\partial \boldsymbol{\mathcal{E}}}{\partial x}$ may be more significant for crystals involving multiple types of chemical species.

- [S4] indicates that the errors $\boldsymbol{\mathcal{E}}$ depend on the variables x to an extent sufficient to alter the relative stability order of two crystal structures x_1 and x_2, both of which correspond to local minima, i.e. $\hat{G}(x_1) < \hat{G}(x_2)$ while $G(x_1) > G(x_2)$.

One practical implication of the above analysis is that, at least in some cases, it may be useful to:

1. Use our CSP methodology as a way of identifying, with reasonable accuracy, a small number n of (likely) stable structures x_k, $k = 1, \ldots, n$, and their corresponding energy values $\hat{G}_k$.

[6] In the case of our CSP methodology, this is the *computed* value of the lattice enthalpy (as opposed to the *true* lattice enthalpy) of the crystal structure.

2. Keep these structures fixed at the values x_k and apply to them more computationally demanding calculations in order to compute more accurate values $\hat{G}'_k$.
3. Re-rank the structures x_k according to the new values $\hat{G}'_k$.

Overall, such a procedure may lead to a more accurate ranking of structures x_k, $k = 1, \ldots, n$ at a relatively moderate cost and without actually introducing additional complex calculations within the optimisation carried out at the refinement stage. Examples of a posteriori calculations that could be applied at step 2 include QM calculations at very high levels of theory, and the use of harmonic approximation techniques for estimating the entropic contributions to the free energy [82].

4.4 Combining Experimental Information and Ab Initio CSP

The free energy residual term $\boldsymbol{\mathcal{E}}$ also provides a useful way of thinking about the potential role of experimental information and empirical models derived from it in CSP. We note that *any* method for constructing an ab initio approximation of free energy, irrespective of its accuracy, is likely to have a non-zero residual, $\boldsymbol{\mathcal{E}}$, and this will inevitably lead to non-zero deviations between predictions and available experimental data. Therefore, a more accurate estimate of the free energy may be achievable by assuming an empirical parameterized functional form, $\boldsymbol{\mathcal{E}}(x, \alpha)$, i.e.

$$G(x) = \hat{G}(x) + \boldsymbol{\mathcal{E}}(x, \alpha) \tag{11}$$

and then using the experimental data to estimate the parameters α so as to minimise some measure of the deviation between data and predictions.

In fact, the use of empirical "repulsion/dispersion" potentials (cf. Sect. 3.4) may be interpreted as one example of the introduction of a residual term. In particular, equations (5) and (6) essentially define the functional form of a parameterized residual function $\boldsymbol{\mathcal{E}}(x, \alpha)$, where the set of parameters α comprises the interaction parameters A_{ii}, B_{ii} and C_{ii} for pairs of atoms of type i. Interestingly, the analysis presented above indicates that:

- Albeit ostensibly intended to account for repulsion/dispersion interactions, this residual term actually acts as an all-purpose "garbage bin", attempting to capture all errors and approximations listed in Sect. 4.2, some of which may be at least as important as repulsion/dispersion.
- The values of the parameters α obtained by any experimental data fitting procedure will depend on the form of the computed energy term $\hat{G}(x)$ used for this procedure; using them in conjunction with a *different* $\hat{G}(x)$ is, to say the least, questionable.

- If the above considerations are not taken into account properly, the use of more sophisticated calculations[7] in an attempt to mitigate the effects of some of the approximations listed in Sect. 4.2 may be ineffective or even counterproductive. For example, employing higher levels of theory in QM calculations may sometimes lead to a worse quality of predictions.

Finally, it could be argued that the immediate objective of introducing *any* empirical function $\mathcal{E}(x, \alpha)$ should be to improve CSP accuracy for a *specific* system of interest. Therefore it would make sense to estimate the parameters α using experimental data that are more directly relevant to the system of interest, in conjunction with the same model $\hat{G}(x)$ as the one that will be used for carrying out the CSP. Examples of appropriate experimental data would include already resolved polymorphs for the same system, or indeed structures in the CSD arising from similar molecules. We note that such an approach would be substantially different to the common practice of using information in the CSD to provide qualitative guidance as to likely high-level features (e.g. packing motifs) in crystal structures; instead, parameter estimation would extract *quantitative* lower-level information on energetic contributions that would complement the ab initio computed energy $\hat{G}(x)$ in the context of formal CSP algorithms. We believe that this area, and the fundamental and practical challenges associated with it, constitute a fruitful subject for further research.

Acknowledgements We wish to acknowledge the major contributions made by P.G. Karamertzanis, M. Vasileiadis and M. Habgood to the fundamentals and implementation of CrystalPredictor and CrystalOptimizer. We are grateful to Professor S.L. Price for many useful discussions and collaboration, and for supplying the DMACRYS code for use within CrystalOptimizer. Financial support for the work reported here was provided by the United Kingdom's Engineering & Physical Sciences Research Council (EPSRC) under grants EP/E016340, EP/J003840/1 and EP/J014958/1.

References

1. Storey RA, Ymén I (2011) Solid state characterization of pharmaceuticals. Wiley, Chichester
2. Bauer J, Spanton S, Henry R, Quick J, Dziki W, Porter W, Morris J (2001) Ritonavir: an extraordinary example of conformational polymorphism. Pharm Res 18:859–866
3. Lommerse JPM, Motherwell WDS, Ammon HL, Dunitz JD, Gavezzotti A, Hofmann DWM, Leusen FJJ, Mooij WTM, Price SL, Schweizer B, Schmidt MU, van Eijck BP, Verwer P, Williams DE (2000) A test of crystal structure prediction of small organic molecules. Acta Crystallogr B 56:697–714
4. Bardwell DA, Adjiman CS, Arnautova YA, Bartashevich E, Boerrigter SXM, Braun DE, Cruz-Cabeza AJ, Day GM, Della Valle RG, Desiraju GR, van Eijck BP, Facelli JC, Ferraro MB, Grillo D, Habgood M, Hofmann DWM, Hofmann F, Jose KVJ, Karamertzanis PG,

[7] Either as part of the energy minimisation calculations or in the form of an a posteriori adjustment of the type discussed in Sect. 4.3.

Kazantsev AV, Kendrick J, Kuleshova LN, Leusen FJJ, Maleev AV, Misquitta AJ, Mohamed S, Needs RJ, Neumann MA, Nikylov D, Orendt AM, Pal R, Pantelides CC, Pickard CJ, Price LS, Price SL, Scheraga HA, van de Streek J, Thakur TS, Tiwari S, Venuti E, Zhitkov IK (2011) Towards crystal structure prediction of complex organic compounds—a report on the fifth blind test. Acta Crystallogr B 67:535–551

5. Day GM, Cooper TG, Cruz-Cabeza AJ, Hejczyk KE, Ammon HL, Boerrigter SXM, Tan JS, Della Valle RG, Venuti E, Jose J, Gadre SR, Desiraju GR, Thakur TS, van Eijck BP, Facelli JC, Bazterra VE, Ferraro MB, Hofmann DWM, Neumann MA, Leusen FJJ, Kendrick J, Price SL, Misquitta AJ, Karamertzanis PG, Welch GWA, Scheraga HA, Arnautova YA, Schmidt MU, van de Streek J, Wolf AK, Schweizer B (2009) Significant progress in predicting the crystal structures of small organic molecules – a report on the fourth blind test. Acta Crystallogr B 65:107–125
6. Day GM, Motherwell WDS, Ammon HL, Boerrigter SXM, Della Valle RG, Venuti E, Dzyabchenko A, Dunitz JD, Schweizer B, van Eijck BP, Erk P, Facelli JC, Bazterra VE, Ferraro MB, Hofmann DWM, Leusen FJJ, Liang C, Pantelides CC, Karamertzanis PG, Price SL, Lewis TC, Nowell H, Torrisi A, Scheraga HA, Arnautova YA, Schmidt MU, Verwer P (2005) A third blind test of crystal structure prediction. Acta Crystallogr B 61:511–527
7. Motherwell WDS, Ammon HL, Dunitz JD, Dzyabchenko A, Erk P, Gavezzotti A, Hofmann DWM, Leusen FJJ, Lommerse JPM, Mooij WTM, Price SL, Scheraga H, Schweizer B, Schmidt MU, van Eijck BP, Verwer P, Williams DE (2002) Crystal structure prediction of small organic molecules: a second blind test. Acta Crystallogr B 58:647–661
8. Ismail SZ, Anderton CL, Copley RCB, Price LS, Price SL (2013) Evaluating a crystal energy landscape in the context of industrial polymorph screening. Crystal Growth Design 13: 2396–2406
9. Kazantsev AV, Karamertzanis PG, Adjiman CS, Pantelides CC, Price SL, Galek PTA, Day GM, Cruz-Cabeza AJ (2011) Successful prediction of a model pharmaceutical in the fifth blind test of crystal structure prediction. Int J Pharm 418:168–178
10. Kendrick J, Stephenson GA, Neumann MA, Leusen FJJ (2013) Crystal structure prediction of a flexible molecule of pharmaceutical interest with unusual polymorphic behavior. Crystal Growth Design 13:581–589
11. Fakes MG, Vakkalagadda BJ, Qian F, Desikan S, Gandhi RB, Lai C, Hsieh A, Franchini MK, Toale H, Brown J (2009) Enhancement of oral bioavailability of an HIV-attachment inhibitor by nanosizing and amorphous formulation approaches. Int J Pharm 370:167–174
12. Campeta AM, Chekal BP, Abramov YA, Meenan PA, Henson MJ, Shi B, Singer RA, Horspool KR (2010) Development of a targeted polymorph screening approach for a complex polymorphic and highly solvating API. J Pharm Sci 99:3874–3886
13. Sakai K, Sakurai K, Nohira H, Tanaka R, Hirayama N (2004) Practical resolution of 1-phenyl-2-(4-methylphenyl)ethylamine using a single resolving agent controlled by the dielectric constant of the solvent. Tetrahedron: Asymmetry 15:3405–3500
14. Friščić T, Lancaster RW, Fábián L, Karamertzanis PG (2010) Tunable recognition of the steroid ς-face by adjacent Π-electron density. Proc Natl Acad Sci 107:13216–13221
15. Uzoh OG, Cruz-Cabeza AJ, Price SL (2012) Is the fenamate group a polymorphophore? Contrasting the crystal energy landscapes of fenamic and tolfenamic acids. Crystal Growth Design 12:4230–4239
16. Arlin J-B, Price LS, Price SL, Florence AJ (2011) A strategy for producing predicted polymorphs: catemeric carbamazepine form V. Chem Commun 47:7074–7076
17. Price S (2013) Why don't we find more polymorphs? Acta Crystallogr B 69:313–328
18. Day GM (2010) Computational crystal structure prediction: towards *in silico* solid form screening. In: Tiekink ERT, Vittal J, Zaworotko M (eds) Organic crystal engineering: frontiers in crystal engineering. Wiley, Chichester, pp 43–66
19. Day GM (2011) Current approaches to predicting molecular organic crystal structures. Crystallogr Rev 17:3–52

20. Day GM (2012) Crystal structure prediction. In: Steed JW, Gale PA (eds) Supramolecular materials chemistry. Wiley, Chichester, pp 2905–2926
21. Kendrick J, Leusen FJJ, Neumann MA, van de Streek J (2011) Progress in crystal structure prediction. Chem Eur J 17:10736–10744
22. Oganov AR (2010) Modern methods of crystal structure prediction. Wiley-VCH, Berlin
23. Price SL (2008) Computational prediction of organic crystal structures and polymorphism. Int Rev Phys Chem 27:541–568
24. Price SL (2008) Computed crystal energy landscapes for understanding and predicting organic crystal structures and polymorphism. Acc Chem Res 42:117–126
25. Price SL (2008) From crystal structure prediction to polymorph prediction: interpreting the crystal energy landscape. Phys Chem Chem Phys 10:1996–2009
26. Karamertzanis PG, Pantelides CC (2005) Ab initio crystal structure prediction—I. Rigid molecules. J Comput Chem 26:304–324
27. Karamertzanis PG, Pantelides CC (2007) Ab initio crystal structure prediction. II. Flexible molecules. Mol Phys 105:273–291
28. Kazantsev AV, Karamertzanis PG, Adjiman CS, Pantelides CC (2011) Efficient handling of molecular flexibility in lattice energy minimization of organic crystals. J Chem Theory Comput 7:1998–2016
29. Baias M, Widdifield CM, Dumez J-N, Thompson HPG, Cooper TG, Salager E, Bassil S, Stein RS, Lesage A, Day GM, Emsley L (2013) Powder crystallography of pharmaceutical materials by combined crystal structure prediction and solid-state ^{1}H NMR spectroscopy. Phys Chem Chem Phys 15:8069–8080
30. Bhardwaj RM, Price LS, Price SL, Reutzel-Edens SM, Miller GJ, Oswald IDH, Johnston BF, Florence AJ (2013) Exploring the experimental and computed crystal energy landscape of olanzapine. Crystal Growth Design 13:1602–1617
31. Eddleston MD, Hejczyk KE, Bithell EG, Day GM, Jones W (2013) Determination of the crystal structure of a new polymorph of theophylline. Chem Eur J 19:7883–7888
32. Eddleston MD, Hejczyk KE, Bithell EG, Day GM, Jones W (2013) Polymorph identification and crystal structure determination by a combined crystal structure prediction and transmission electron microscopy approach. Chem Eur J 19:7874–7882
33. Habgood M (2012) Solution and nanoscale structure selection: implications for the crystal energy landscape of tetrolic acid. Phys Chem Chem Phys 14:9195–9203
34. Habgood M, Lancaster RW, Gateshki M, Kenwright AM (2013) The amorphous form of salicylsalicylic acid: experimental characterization and computational predictability. Crystal Growth Design 13:1771–1779
35. Spencer J, Patel H, Deadman JJ, Palmer RA, Male L, Coles SJ, Uzoh OG, Price SL (2012) The unexpected but predictable tetrazole packing in flexible 1-benzyl-1H-tetrazole. CrystEngComm 14:6441–6446
36. Vasileiadis M, Kazantsev AV, Karamertzanis PG, Adjiman CS, Pantelides CC (2012) The polymorphs of ROY: application of a systematic crystal structure prediction technique. Acta Crystallogr B 68:677–685
37. Issa N, Barnett SA, Mohamed S, Braun DE, Copley RCB, Tocher DA, Price SL (2012) Screening for cocrystals of succinic acid and 4-aminobenzoic acid. CrystEngComm 14:2454–2464
38. Karamertzanis PG, Kazantsev AV, Issa N, Welch GWA, Adjiman CS, Pantelides CC, Price SL (2009) Can the formation of pharmaceutical cocrystals be computationally predicted? 2. Crystal structure prediction. J Chem Theory Comput 5:1432–1448
39. Wu H, Habgood M, Parker JE, Reeves-McLaren N, Cockcroft JK, Vickers M, West AR, Jones AG (2013) Crystal structure determination by combined synchrotron powder X-ray diffraction and crystal structure prediction: 1: 1 L-ephedrine D-tartrate. CrystEngComm 15:1853–1859
40. Braun DE, Ardid-Candel M, D'Oria E, Karamertzanis PG, Arlin J-B, Florence AJ, Jones AG, Price SL (2011) Racemic naproxen: a multidisciplinary structural and thermodynamic comparison with the enantiopure form. Crystal Growth Design 11:5659–5669

41. Habgood M (2013) Analysis of enantiospecific and diastereomeric cocrystal systems by crystal structure prediction. Crystal Growth & Design 13:4549–4558
42. Braun DE, Bhardwaj RM, Florence AJ, Tocher DA, Price SL (2012) Complex polymorphic system of gallic acid—five monohydrates, three anhydrates, and over 20 solvates. Crystal Growth Design 13:19–23
43. Braun DE, Karamertzanis PG, Price SL (2011) Which, if any, hydrates will crystallise? Predicting hydrate formation of two dihydroxybenzoic acids. Chem Commun 47:5443–5445
44. Gelbrich T, Braun DE, Ellern A, Griesser UJ (2013) Four polymorphs of methyl paraben: structural relationships and relative energy differences. Crystal Growth Design 13:1206–1217
45. Admiraal G, Eikelenboom JC, Vos A (1982) Structures of the triclinic and monoclinic modifications of (2-oxo-1-pyrrolidinyl)acetamide. Acta Crystallogr B 38:2600–2605
46. Gavezzotti A, Filippini G (1995) Polymorphic forms of organic-crystals at room conditions – thermodynamic and structural implications. J Am Chem Soc 117:12299–12305
47. Clark SJ, Segall MD, Pickard CJ, Hasnip PJ, Probert MI, Refson K, Payne MC (2005) First principles methods using CASTEP. Zeitschrift fuer Kristallographie 220:567–570
48. Neumann MA, Perrin MA (2005) Energy ranking of molecular crystals using density functional theory calculations and an empirical van der Waals correction. J Phys Chem B 109: 15531–15541
49. Kim HS, Jeffrey GA (1969) The crystal structure of xylitol. Acta Crystallogr B 25:2607–2613
50. Karamertzanis PG, Pantelides CC (2004) Optimal site charge models for molecular electrostatic potentials. Mol Simulat 30:413–436
51. Stone AJ (1996) The theory of intermolecular forces. Clarendon, Oxford
52. Stone AJ, Alderton M (1985) Distributed multipole analysis – methods and applications. Mol Phys 56:1047–1064
53. Mooij WTM, van Duijneveldt FB, van Duijneveldt-van de Rijdt JGCM, van Eijck BP (1999) Transferable ab initio intermolecular potentials. 1. Derivation from methanol dimer and trimer calculations. J Phys Chem A 103:9872–9882
54. Cooper TG, Hejczyk KE, Jones W, Day GM (2008) Molecular polarization effects on the relative energies of the real and putative crystal structures of valine. J Chem Theory Comput 4:1795–1805
55. Misquitta AJ, Welch GWA, Stone AJ, Price SL (2008) A first principles prediction of the crystal structure of $C_6Br_2ClFH_2$. Chem Phys Lett 456:105–109
56. Stone AJ, Misquitta AJ (2007) Atom-atom potentials from ab initio calculations. Int Rev Phys Chem 26:193–222
57. Beyer T, Price SL (2000) Dimer or catemer? Low-energy crystal packings for small carboxylic acids. J Phys Chem B 104:2647–2655
58. Coombes DS, Price SL, Willock DJ, Leslie M (1996) Role of electrostatic interactions in determining the crystal structures of polar organic molecules. A distributed multipole study. J Phys Chem 100:7352–7360
59. Cox SR, Hsu LY, Williams DE (1981) Nonbonded potential function models for the crystalline oxohydrocarbons. Acta Crystallogr A 37:293–301
60. Williams DE (1965) Repulsion center of a bonded hydrogen atom. J Chem Phys 43:4424–4426
61. Williams DE (1999) Improved intermolecular force field for crystalline hydrocarbons containing four- or three-coordinated carbon. J Mol Struct 485–486:321–347
62. Williams DE (2001) Improved intermolecular force field for molecules containing H, C, N, and O atoms, with application to nucleoside and peptide crystals. J Comput Chem 22: 1154–1166
63. Williams DE, Cox SR (1984) Nonbonded potentials for azahydrocarbons: the importance of the coulombic interaction. Acta Crystallogr B 40:404–417
64. Ewald PP (1921) Die Berechnung Optischer Und Elektrostatischer Gitterpotentiale. Annalen der Physik (Berlin) 369:253–287
65. Nelder JA, Mead R (1965) A simplex method for function minimization. Comput J 7:308–313
66. Nocedal J, Wright SJ (2006) Numerical optimization, 2nd edn. Springer, New York

67. Yu L (2010) Polymorphism in molecular solids: an extraordinary system of red, orange, and yellow crystals. Acc Chem Res 43:1257–1266
68. Day GM, Motherwell WDS, Jones W (2007) A strategy for predicting the crystal structures of flexible molecules: the polymorphism of phenobarbital. Phys Chem Chem Phys 9:1693–1704
69. Stone AJ (2005) Distributed multipole analysis: stability for large basis sets. J Chem Theory Comput 1:1128–1132
70. Gavezzotti A (1997) Theoretical aspects and computer modeling of the molecular solid state. Wiley, Chichester
71. van Mourik T, Karamertzanis PG, Price SL (2006) Molecular conformations and relative stabilities can be as demanding of the electronic structure method as intermolecular calculations. J Phys Chem A 112:8–12
72. Karamertzanis PG, Price SL (2006) Energy minimization of crystal structures containing flexible molecules. J Chem Theory Comput 2:1184–1199
73. Brodersen S, Wilke S, Leusen FJJ, Engel G (2003) A study of different approaches to the electrostatic interaction in force field methods for organic crystals. Phys Chem Chem Phys 5:4923–4931
74. Day GM, Motherwell WDS, Jones W (2005) Beyond the isotropic atom model in crystal structure prediction of rigid molecules: atomic multipoles versus point charges. Crystal Growth Design 5:1023–1033
75. Mooij WTM, Leusen FJJ (2001) Multipoles versus charges in the 1999 crystal structure prediction test. Phys Chem Chem Phys 3:5063–5066
76. Buckingham RA (1938) The classical equation of state of gaseous helium, neon and argon. Proc R Soc Lond A Math Phys Sci 168:264–283
77. Sobol' IM (1967) On the distribution of points in a cube and the approximate evaluation of integrals. Comp Math Math Phys 7:86–112
78. Spek AL (2003) PLATON, a multipurpose crystallographic tool. Utrecht University, The Netherlands
79. Price SL, Leslie M, Welch GWA, Habgood M, Price LS, Karamertzanis PG, Day GM (2010) Modelling organic crystal structures using distributed multipole and polarizability-based model intermolecular potentials. Phys Chem Chem Phys 12:8478–8490
80. Willock DJ, Price SL, Leslie M, Catlow CRA (1995) The relaxation of molecular crystal structures using a distributed multipole electrostatic model. J Comput Chem 16:628–647
81. Chisholm JA, Motherwell WDS (2005) COMPACK: a program for identifying crystal structure similarity using distances. J Appl Crystallogr 38:228–231
82. Vasileiadis M (2013) Calculation of the free energy of crystalline solids. PhD thesis, Imperial College London, London

Top Curr Chem (2014) 345: 59–94
DOI: 10.1007/128_2013_502
© Springer-Verlag Berlin Heidelberg 2013
Published online: 1 December 2013

Accurate and Robust Molecular Crystal Modeling Using Fragment-Based Electronic Structure Methods

Gregory J.O. Beran, Shuhao Wen, Kaushik Nanda, Yuanhang Huang, and Yonaton Heit

Abstract Accurately modeling molecular crystal polymorphism requires careful treatment of diverse intra- and intermolecular interactions which can be difficult to achieve without the use of high-level ab initio electronic structure techniques. Fragment-based methods like the hybrid many-body interaction QM/MM technique enable the application of accurate electronic structure models to chemically interesting molecular crystals. The theoretical underpinnings of this approach and the practical requirements for the QM and MM contributions are discussed. Benchmark results and representative applications to aspirin and oxalyl dihydrazide crystals are presented.

Keywords Fragment methods · Molecular crystals · Polymorph prediction

Contents

G.J.O. Beran (✉), S. Wen, K. Nanda, Y. Huang, and Y. Heit
Department of Chemistry, University of California, Riverside, CA 92521, USA
e-mail: gregory.beran@ucr.edu

1 Introduction

Organic molecular crystals often exhibit a variety of different packing motifs, or polymorphs. These different crystal packing motifs can have diverse physical properties, making crystal structure critically important in a wide range of fields. Polymorphism plays a major role in the pharmaceutical industry, for example, where a substantial fraction of drugs including aspirin, acetaminophen, Lipitor, and Zantac have known polymorphs. Polymorphs of a given pharmaceutical can have drastically different solubilities and bioavailabilities, making the understanding of polymorphism critical for the drug industry.

Pharmaceutical polymorphism has led to several major drug recalls or withdrawals in recent years. For instance, the HIV drug ritonavir was temporarily removed from the market in 1998 when a new, insoluble polymorph appeared in production facilities, leading to shortages of this desperately needed medicine and costing its maker hundreds of millions of dollars in lost sales [1, 2]. Polymorphism is also believed to be behind multiple recalls of the anti-seizure drug carbamazepine, which exhibits several low-solubility polymorphs [3]. In 2008, Neupro brand skin patches for the Parkinson's disease drug rotigotine were withdrawn from the market when a less-effective crystal form appeared visibly on the patches as dendritic structures [4]. Pharmaceutical polymorphism has also been the subject of many legal battles arising from the fact that unique crystal forms are patentable. Major examples include the ulcer/heartburn medication Zantac and the antibiotic cefadroxil [5].

Crystal packing is also important for foods such as chocolate. Solid cocoa butter form V is desired to achieve chocolate with a shiny appearance, a 34°C melting point that causes it to melt pleasingly on the tongue, and other favorable characteristics. However, form VI cocoa butter, which leads to chocolate that is dull, soft, grainy tasting, and has a melting point a few degrees higher, is thermodynamically more stable [6]. At room temperature the transition from form V to form VI occurs on a timescale of months, and it occurs even faster at elevated temperatures. The chocolate industry expends considerable effort to produce and maintain chocolate in the proper form to ensure a high-quality product with a reasonable shelf life.

Many other areas of chemistry and materials science must cope with molecular crystal polymorphism as well. Crystal packing influences the stability, sensitivity, and detonation characteristics of energetic materials, for instance [7, 8]. It also can have drastic effects on organic semiconductor materials. Solid rubrene currently holds the record for the highest-known carrier mobility in an organic molecular crystal. However, a rubrene derivative with a different crystal packing motif exhibits no measurable carrier mobility [9].

Predicting molecular crystal structure from first principles is extremely challenging. The problem involves (1) a search over many (millions or more) possible crystal packings, (2) the accurate evaluation of the lattice energy of the possible structures (at 0 K), (3) calculation of the finite-temperature thermodynamic

contributions, and (4) an understanding of the competition between thermodynamically and kinetically preferred packing arrangements.

Significant progress on the "search" problem has been made in recent years. For example, the Price group uses a hierarchical series of ever-improving theoretical models to screen out unlikely structures and eventually identify the most stable forms [10, 11]. The initial screening might consider some ~10^7 randomly generated structures with different space group symmetries and numbers of molecules in the asymmetric unit cell using a very simple force field. This simple force field allows one to rule out a significant fraction of the energetically uncompetitive structures. Subsequent rounds improve the quality of the force field, winnowing down the potential structures toward a final prediction. Neumann and co-workers use a mixture of density functional theory (DFT) with dispersion corrections and a force-field fitted to match the DFT results to search over possible structures [12–15]. In their work the force field identifies a subset of likely structures which are then refined with DFT. Both strategies proved effective in the two most recent blind tests of crystal structure prediction [16, 17]. Recent progress in crystal structure global optimization algorithms may also help solve the search problem [18].

Once a relatively small number of candidate crystal structures have been identified, discrimination among them requires predicting the lattice energies or relative energies very accurately. This necessitates using a theoretical approach that can handle the subtle balances between intra- and intermolecular interactions that characterize conformational polymorphism [19]. One must treat the diverse non-covalent interactions – hydrogen bonding, electrostatics, induction (a.k.a. polarization), and van der Waals dispersion – with high and uniform accuracy to avoid biasing the predictions toward certain classes of structures (e.g., hydrogen-bonded vs π-stacking motifs). Overall, the energy differences between experimentally observed polymorphs are typically less than 10 kJ/mol, and they are often closer to ~1 kJ/mol.

After obtaining reliable predictions at 0 K, one can start to think about finite temperature effects. The computation of finite-temperature enthalpies, entropies, and free energies is much less mature. The entropic contributions to relative polymorph stabilities are commonly assumed to be smaller than the enthalpic ones [20], but of course there will be many exceptions. Thermal effects are typically estimated using simple (quasi-)harmonic approaches (e.g., [21]) though sampling-based free energy methods are starting to be explored more actively in this context (e.g., [22, 23]).

Finally, real-world crystallization is often driven by kinetics, not thermodynamics. Understanding the kinetics of crystallization is probably even more difficult than the thermodynamics, since it requires a dynamical understanding of the nucleation and crystal growth processes with sufficient accuracy to differentiate correctly among the different packing motifs. Progress in this direction is also being made, but it will likely be quite some time before one can reliably predict which crystal structures will form kinetically under a given set of crystallization conditions.

Nevertheless, the ability to predict the thermodynamically preferred structures reliably would be very useful for predicting temperature and pressure regions of phase stability, to determine whether a given pharmaceutical polymorph is stable or metastable, or to predict a variety of crystal properties. Toward this end, this chapter focuses on computing accurate lattice energies and relative energies at 0 K.

2 Theory

The theoretical treatment of molecular crystals has traditionally relied on either molecular mechanics (MM) force fields or electronic structure methods with periodic boundary conditions. Force field modeling of molecular crystals has improved dramatically over the past decade [19, 24–27], as demonstrated by the major successes in the recent blind tests of crystal structure prediction [11, 12, 15–17, 28]. Much of this success arises from the inclusion of increasing amounts of quantum mechanical information into the force fields, ranging from the parameterization to the determination of intramolecular conformation.

In that vein, one should be able to achieve even better accuracy by treating the systems fully quantum mechanically. This requires a careful balance between accuracy and computational expense. Most quantum mechanical (QM) calculations on molecular crystals are performed with periodic DFT. Widely used semi-local density functionals generally do not describe van der Waals dispersion interactions. However, there has been substantial progress toward including van der Waals dispersion interactions either self-consistently or as an a posteriori correction using a variety of empirical and non-empirical strategies [29–35]. Despite the tremendous progress in this area, it can sometimes be difficult to identify when DFT methods are performing well enough or to interpret the results when different density functionals make contradictory predictions.

Wavefunction methods offer the potential to improve the quality of the predictions systematically by improving the wavefunction. The simplest useful wavefunction technique for molecular crystals is second-order Møller–Plesset perturbation theory (MP2). A number of periodic MP2 implementations exist, including efficient ones based on local-correlation ideas [36–47]. These provide a nice alternative to DFT, but they remain relatively computationally expensive. Furthermore, MP2 correlation itself is often insufficient, as will be discussed below. Unfortunately, more accurate periodic coupled cluster implementations are too expensive to be applied to most chemically interesting molecular crystals [48–51].

2.1 Fragment-Based Methods in Electronic Structure Theory

Fragment-based methods provide a lower-cost alternative to traditional periodic boundary condition electronic structure methods. These techniques partition the

system in some fashion, perform quantum mechanical calculations on each individual fragment, and piece them together to obtain information about the system as a whole. While many individual fragment calculations are needed in a single crystal, each is relatively inexpensive compared to a full periodic crystal calculation. This allows one to utilize higher-level electronic structure methods at much lower computational cost.

Early fragment techniques include the fragment molecular orbital method [52, 53], divide-and-conquer techniques [54], and incremental schemes [55, 56], but there has been an explosion of fragment methods in recent years, as discussed in two reviews [57, 58] and in a thematic issue of *Physical Chemistry Chemical Physics* [59]. A couple of groups have also identified a common framework that unifies most or all of these fragment methods [60, 61].

One very common fragment strategy decomposes the total energy of a set of interacting molecules, whether in a cluster or a crystal, according to a many-body expansion:

$$E_{\text{total}} = E_{\text{1-body}} + E_{\text{2-body}} + E_{\text{3-body}} + \cdots. \tag{1}$$

The expansion is formally exact, but any computationally useful application of this expansion requires approximation of the higher-order terms in some fashion. For a typical molecular crystal, the three-body and higher terms account for 10–20% of the total interaction energy, making those terms necessary for accurate crystal modeling.

Approximations to the many-body expansion typically fall into two categories. The first category uses electrostatic embedding to incorporate polarization effects into the lower-order terms, thereby reducing the importance of the higher-order terms. The many-body expansion can then be truncated after two-body or three-body terms. Examples of such embedding approaches include the electrostatically-embedded many-body expansion approach [62–64], binary interaction [65–67], and the exactly embedded density functional many-body expansion [68].

The electrostatic embedding methods are very successful, but they suffer from two potential disadvantages. First, while it is often true that induction effects dominate the many-body contributions, many-body dispersion has also proved important in molecular crystal systems [30, 58, 69, 70], and those effects are not captured via electrostatic embedding. Second, embedding complicates the calculation of nuclear gradients and Hessians of the energy. When embedding a particular monomer or dimer in a potential arising from the other molecules, the embedding potential depends on the positions of all other atoms in the system. Therefore, the energy gradient of that monomer or dimer now depends on all 3N atomic coordinates, instead of just the coordinates of the monomer or dimer in question. These additional gradient contributions arising from the embedding-potential are often neglected, but they can sometimes be significant [65].

The methods in the second category of many-body expansion approximations make no truncation in the many-body expansion. Instead, the higher-order terms are approximated at some lower level of theory. The incremental method and

other related schemes treat the important low-order terms with accurate ab initio methods and approximate the higher-order terms using Hartree-Fock (HF), DFT, or any higher-level QM technique. Alternatively, the hybrid many-body interaction (HMBI) approach (and a nearly identical model simultaneously and independently proposed by Manby and co-workers [71]) approximates the higher-order terms using a polarizable force field. The advantage of this approach is that polarizable force fields are much less expensive to evaluate than electronic structure methods. Compared to embedding techniques, these approaches can avoid the need to make a priori assumptions about where to truncate the many-body expansion and provide the flexibility to build in whatever physical terms are needed in the system, including the aforementioned many-body dispersion effects.

2.2 The Hybrid Many-Body Interaction (HMBI) Method

In the HMBI model, the intramolecular (one-body) and short-range (SR) pairwise intermolecular (two-body) interactions are treated with electronic structure theory, while the long-range (LR) two-body and the many-body terms are approximated with the polarizable force field:

$$E_{\text{HMBI}} = E^{\text{QM}}_{\text{1-body}} + E^{\text{QM}}_{\text{SR2-body}} + E^{\text{MM}}_{\text{LR2-body}} + E^{\text{MM}}_{\text{many-body}}, \tag{2}$$

where the "many-body" term includes all three-body and higher interactions (Fig. 1). To evaluate the energy in practice, one exploits the fact that one can write a many-body expansion purely in terms of MM contributions:

$$E^{\text{MM}}_{\text{total}} = E^{\text{MM}}_{\text{1-body}} + E^{\text{MM}}_{\text{SR2-body}} + E^{\text{MM}}_{\text{LR2-body}} + E^{\text{MM}}_{\text{many-body}}, \tag{3}$$

which can be rearranged as

$$E^{\text{MM}}_{\text{LR2-body}} + E^{\text{MM}}_{\text{many-body}} = E^{\text{MM}}_{\text{total}} - E^{\text{MM}}_{\text{1-body}} - E^{\text{MM}}_{\text{SR2-body}}. \tag{4}$$

Substituting this expression into (2) leads to the following working HMBI energy expression:

$$E^{\text{HMBI}}_{\text{total}} = E^{\text{MM}}_{\text{total}} + \sum_i \left(E_i^{\text{QM}} - E_i^{\text{MM}}\right) + \sum_{ij} d_{ij}(R)\left(\Delta^2 E_{ij}^{\text{QM}} - \Delta^2 E_{ij}^{\text{MM}}\right), \tag{5}$$

where E_i corresponds to the energy of monomer i, and $\Delta^2 E_{ij} = E_{ij} - E_i - E_j$ is the interaction energy between monomers i and j evaluated as the difference between the total energy of the dimer E_{ij} and the individual monomer energies E_i and E_j. Note that the monomer energies in these expressions are evaluated at the same

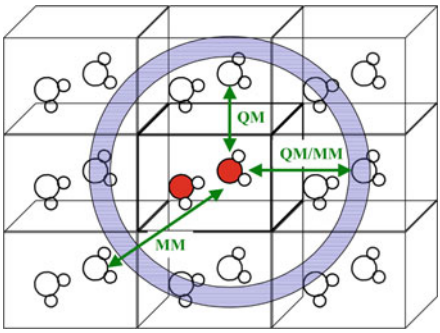


Fig. 1 Schematic of the HMBI method. Each individual molecule in the unit cell and its short-range pairwise interactions are treated with QM, while longer-range interactions and interactions involving more than two molecules are treated with MM. The *shaded region* indicates the smooth transition QM and MM via interpolation. Reprinted with permission from [115]. Copyright 2010 American Chemical Society

geometry as in the dimer (and in the full cluster or crystal). The $d_{ij}(R)$ term ensures no discontinuities arise in the potential energy surface as the model transitions from the short-range QM to the long-range MM two-body interaction regimes. This function decays from 1 to 0 as a function of the intermolecular distance R (defined here as the shortest distance between the two molecules) over a user-defined interval governed by two parameters, r_1 and r_0 [72]:

$$d_{ij}(R) = \frac{1}{1 + e^{2\left|r_1 - r_0\right|/(r_1 - R) - \left|r_1 - r_0\right|/(R - r_0)}}. \tag{6}$$

By default, we conservatively transition from QM to MM between $r_1 = 9.0$ Å and $r_0 = 10.0$ Å, but more aggressive cutoffs can often be used. In water/ice, for instance, using the ab initio force field described in the next section, one can transition from QM to MM between 4.5 and 5.5 Å with virtually no loss in accuracy (see Sect. 2.3).

The energy expression in (5) resembles the expressions used in a two-layer ONIOM-style QM/MM model. One computes the energy of the entire system at a low-level of theory, and then corrects it with smaller calculations at a higher level of theory. The physics described by the two types of models is quite different, however. In ONIOM QM/MM, one partitions the system into distinct QM and MM regions. In the fragment-based HMBI QM/MM approach, one instead partitions based on the nature of the interaction. There are no specific QM and MM regions. Each molecule has both QM and MM interactions in the HMBI model. The important interactions are treated with QM, while the less important ones are treated with MM. In this sense, the HMBI model is spatially homogeneous, as is appropriate for modeling a molecular crystal.

For systems with periodic boundary conditions like crystals, the HMBI energy expression in (5) must also include pairwise interactions between central unit cell molecules and their periodic images:

$$E_{\text{total}}^{\text{HMBI}} = E_{\text{total}}^{\text{MM}} + \sum_i \left(E_i^{\text{QM}} - E_i^{\text{MM}}\right) + \sum_{ij} d_{ij(0)} \left(\Delta^2 E_{ij(0)}^{\text{QM}} - \Delta^2 E_{ij(0)}^{\text{MM}}\right) + \frac{1}{2} \sum_i \sum_{j(\mathbf{n})}^{\text{images}} d_{ij(\mathbf{n})} \left(\Delta^2 E_{ij(\mathbf{n})}^{\text{QM}} - \Delta^2 E_{ij(\mathbf{n})}^{\text{MM}}\right). \tag{7}$$

In this expression, i runs over molecules in the central unit cell, while j can either be in the central unit cell ($j(0)$) or in the periodic image cell $\mathbf{n}$ ($j(\mathbf{n})$). Unit cell $\mathbf{n}$ is defined as the cell whose origin lies at vector $\mathbf{n} = n_{v_1}\mathbf{v}_1 + n_{v_2}\mathbf{v}_2 + n_{v_3}\mathbf{v}_3$ in terms of the three lattice vectors, $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$. In practice, thanks to the function $d_{ij}(R)$, the sum over $j(\mathbf{n})$ runs only over molecules within a distance r_0 of the current central unit cell molecule i.

The computational bottleneck in HMBI is the evaluation of the QM pairwise interaction energies $\Delta^2 E_{ij(0)}$ and $\Delta^2 E_{ij(\mathbf{n})}$. Because only short-range pairwise interactions are treated with QM, the HMBI model scales linearly with the number of molecules in the system (non-periodic) or unit-cell (periodic). More precisely, the MM terms do not scale linearly, but their cost is so much smaller than that of the QM terms for any practical system that linear scaling behavior is observed. For non-periodic systems, the onset of linear scaling occurs once the system becomes larger than the outer extent of the QM-to-MM transition region, r_0. In periodic systems, which are formally infinite, linear-scaling behavior is observed for unit cells of any size. This linear-scaling behavior is a key advantage of fragment methods compared to fully QM methods like periodic DFT or MP2. Large unit cells or supercells of the sort that might be used to perform lattice dynamics or to examine localized/defect behavior can be much cheaper with a fragment method than with a traditional periodic QM methods.

2.2.1 Nuclear Gradients and Hessians

The HMBI energy expression contains only additive energy contributions and does not use any sort of embedding, so derivatives of the energy can be computed straightforwardly. For instance, if q_l corresponds to the Cartesian x, y, or z coordinates (not its fractional coordinate) of the l-th atom in the central unit cell, then the gradient of the energy with respect to q_l is given by [73]

$$\frac{\partial E^{\mathrm{HMBI}}_{\mathrm{total}}}{\partial q_l} = \frac{\partial E^{\mathrm{MM}}_{\mathrm{total}}}{\partial q_l} + \sum_i \left(\frac{\partial E^{\mathrm{QM}}_i}{\partial q_l} - \frac{\partial E^{\mathrm{MM}}_i}{\partial q_l} \right) + \sum_{ij(\mathbf{0})} \frac{\partial d_{ij(0)}}{\partial q_l} \left(\Delta^2 E^{\mathrm{QM}}_{ij(0)} - \Delta^2 E^{\mathrm{MM}}_{ij(0)} \right)$$
$$+ \sum_{ij(\mathbf{0})} d_{ij(0)} \left(\frac{\partial \Delta^2 E^{\mathrm{QM}}_{ij(0)}}{\partial q_l} - \frac{\partial \Delta^2 E^{\mathrm{MM}}_{ij(0)}}{\partial q_l} \right)$$
$$+ \frac{1}{2} \sum_i \sum_{j(\mathbf{n})}^{\mathrm{images}} \frac{\partial d_{ij(\mathbf{n})}}{\partial q_l} \left(\Delta^2 E^{\mathrm{QM}}_{ij(\mathbf{n})} - \Delta^2 E^{\mathrm{MM}}_{ij(\mathbf{n})} \right)$$
$$+ \frac{1}{2} \sum_i \sum_{j(\mathbf{n})}^{\mathrm{images}} d_{ij(\mathbf{n})} \left(\frac{\partial \Delta^2 E^{\mathrm{QM}}_{ij(\mathbf{n})}}{\partial q_l} - \frac{\partial \Delta^2 E^{\mathrm{MM}}_{ij(\mathbf{n})}}{\partial q_l} \right). \tag{8}$$

The individual one-body and two-body energy gradient terms in (8) are obtained readily from the monomer and dimer gradients computed in standard electronic structure or MM software packages. The expression for the gradient of the QM-to-MM smoothing function d_{ij} has been provided previously [73].

To optimize the size and shape of the unit cell, one also needs the gradient with respect to the lattice vectors. When working in Cartesian coordinates instead of fractional coordinates, changing the lattice vectors does not affect the one-body or two-body terms within the central unit cell, but it does affect the two-body contributions due to interactions between molecules in the central unit cell and those in periodic image cells and the MM energy of the entire crystal. The resulting gradient with respect to the qth coordinate (x, y, or z) of the ϵth lattice vector ($\mathbf{v}_1$, $\mathbf{v}_2$, or $\mathbf{v}_3$) is given by

$$\frac{\partial E}{\partial v_{\epsilon q}} = \frac{\partial E^{\mathrm{MM}}_{\mathrm{total}}}{\partial v_{\epsilon q}} + \frac{1}{2} \sum_i \sum_{j(n_v)} n_v \sum_k \left\{ \frac{\partial d_{ij(n_v)}}{\partial q_k} \left(\Delta^2 E^{\mathrm{QM}}_{ij(n_v)} - \Delta^2 E^{\mathrm{MM}}_{ij(n_v)} \right) \right.$$
$$\left. + d_{ij(n_v)} \left(\left(\frac{\partial E^{\mathrm{QM}}_{ij(n_v)}}{\partial q_k} - \frac{\partial E^{\mathrm{QM}}_{j(n_v)}}{\partial q_k} \right) - \left(\frac{\partial E^{\mathrm{MM}}_{ij(n_v)}}{\partial q_k} - \frac{\partial E^{\mathrm{MM}}_{j(n_v)}}{\partial q_k} \right) \right) \right\}. \tag{9}$$

where k sums over the atoms in periodic image monomer j. Often, one expresses the unit cell in terms of three lattice constants (a, b, and c) and three angles (α, β, and γ). The expressions for the nine components of the lattice vector gradients in (9) can be transformed into expressions for these six lattice parameters [73].

Note that, except for the term $\partial E^{\mathrm{MM}}_{\mathrm{total}}/\partial v_{\epsilon q}$, all the terms needed to compute the lattice vector gradients in (9) are already available from the gradients with respect to the atomic positions (8). The $\partial E^{\mathrm{MM}}_{\mathrm{total}}/\partial v_{\epsilon q}$ term can be computed inexpensively, since it is at the MM level. Therefore, the calculation of the lattice vector gradient typically requires minimal additional work once the nuclear gradients with respect to atomic position have been obtained. The evaluation of the one-body and

two-body QM gradients forms the computational bottleneck in evaluating the full nuclear gradient.

The nuclear Hessian can be computed similarly by differentiating (8) with respect to a second nuclear coordinate, $q_{l'}$:

$$\begin{aligned}\frac{\partial^2 E_{\text{total}}^{\text{HMBI}}}{\partial q_l \partial q_{l'}} = &\frac{\partial E_{\text{PBC}}^{\text{MM}}}{\partial q_l \partial q_{l'}} + \sum_i \left(\frac{\partial^2 E_i^{\text{QM}}}{\partial q_l \partial q_{l'}} - \frac{\partial^2 E_i^{\text{MM}}}{\partial q_l \partial q_{l'}} \right) \\ &+ \sum_{ij(\mathbf{n})} \zeta_{ij} d_{ij(\mathbf{n})} \left(\frac{\partial^2 \Delta^2 E_{ij(\mathbf{n})}^{\text{QM}}}{\partial q_l \partial q_{l'}} - \frac{\partial^2 \Delta^2 E_{ij(\mathbf{n})}^{\text{MM}}}{\partial q_l \partial q_{l'}} \right) \\ &+ \sum_{ij(\mathbf{n})} \zeta_{ij} \frac{\partial d_{ij(\mathbf{n})}}{\partial q_l} \left(\frac{\partial \Delta^2 E_{ij(\mathbf{n})}^{\text{QM}}}{\partial q_{l'}} - \frac{\partial \Delta^2 E_{ij(\mathbf{n})}^{\text{MM}}}{\partial q_{l'}} \right) \\ &+ \sum_{ij(\mathbf{n})} \zeta_{ij} \frac{\partial d_{ij(\mathbf{n})}}{\partial q_{l'}} \left(\frac{\partial \Delta^2 E_{ij(\mathbf{n})}^{\text{QM}}}{\partial q_l} - \frac{\partial \Delta^2 E_{ij(\mathbf{n})}^{\text{MM}}}{\partial q_l} \right) \\ &+ \sum_{ij(\mathbf{n})} \zeta_{ij} \frac{\partial^2 d_{ij(\mathbf{n})}}{\partial q_l \partial q_{l'}} \left(\Delta^2 E_{ij(\mathbf{n})}^{\text{QM}} - \Delta^2 E_{ij(\mathbf{n})}^{\text{MM}} \right),\end{aligned} \tag{10}$$

where the sums over $j(\mathbf{n})$ run over molecules in both the central unit cell and in periodic image cells, $\zeta_{ij} = 1$ for dimers where both monomers lie within the central unit cell ($\mathbf{n} = 0$), and $\zeta_{ij} = 1/2$ for dimers where the second monomer lies outside the central unit cell. Evaluation of the Hessian requires gradients and Hessians for each individual monomer and dimer. Once the Hessian has been computed, one can compute harmonic vibrational frequencies, lattice dynamics, statistical thermodynamic partition functions, etc.

2.2.2 Crystal Symmetry

Molecular crystals often exhibit high symmetry, both translational (due to the periodic boundary conditions) and space group symmetry, and significant computational savings can be reaped by exploiting this symmetry in the monomer and dimer calculations. The straightforward approach would identify the space group of the crystal and use the symmetry operations of that group to determine which monomers and dimers are symmetry-equivalent. One need only compute the energies and forces for the symmetry-unique monomers and dimers and then scale their contributions based on the number of symmetry-equivalent monomers or dimers.

Alternatively, one can simply rotate the monomers and dimers to a common reference frame (e.g., aligned along the principle axes of inertia) and simply test which monomer/dimer geometries are identical within some numerical threshold.

This approach avoids the need (1) to specify the space group and (2) to program the symmetry operations for all 230 space groups.

A sizable fraction of organic molecular crystals exhibit $P2_1/c$ symmetry, for which roughly fourfold computational savings can be obtained. The savings can be even larger for other space groups. For acetamide crystals in the R3c space group, one obtains 18-fold speed-ups by exploiting symmetry. So while the details are system-dependent, for a crystal with one molecule in the asymmetric unit cell and a conservative QM to MM transition distance, one might typically need to perform ~50 symmetry-unique dimer calculations.

2.3 *Accurate Force Fields for Long-Range and Many-Body Interactions*

The success of the HMBI approach depends critically on the quality of the polarizable force field used for the long-range two-body and the many-body terms; c.f. (2). This means that it needs to capture two-body electrostatics, two-body van der Waals dispersion, self-consistent long-range and many-body induction, and many-body dispersion interactions (approximated here with only the leading three-body Axilrod–Teller term):

$$E^{\mathrm{MM}} = E_{2\text{-body es}} + E_{2\text{-body disp}} + E_{\mathrm{induction}} + E_{3\text{-body disp}}. \tag{11}$$

In our early work [74] we used the Amoeba force field [75, 76], which includes all of these contributions except for the many-body dispersion. It works fairly well in this context, but even better results are achieved by constructing an ab initio force field (AIFF) "on the fly" based on QM calculations for each individual monomer [77, 78].

The key idea behind the AIFF is to parameterize the force field in terms of atom-centered distributed multipole moments [79–81], distributed polarizabilities [82, 83], and distributed dispersion coefficients [84]. These are obtained from the molecular electron density, the static polarizabilities, and the frequency-dependent polarizabilities, respectively. The form of the force field is well-justified at long ranges from intermolecular perturbation theory, and empirical short-range damping functions help avoid serious problems for shorter-range interactions [85, 86].

This AIFF model mimics the much more expensive QM treatment very well, as shown in Figs. 2, 3, and 4. The use of multipolar expansions and the lack of exchange terms lead to some problems at short range, but the long-range interactions are modeled very accurately. Figure 2 shows that the predicted lattice energy of ice Ih is virtually invariant to the distance at which HMBI transitions from a QM to an MM description of the long-range pairwise interactions once the two molecules are ~4.5 Å apart. For other systems, the corresponding transition distance may

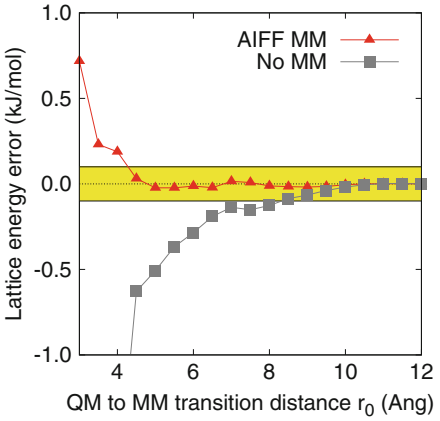


Fig. 2 Performance of the AIFF for long-range pairwise interactions: Change in the lattice energy of ice Ih as a function of the QM to MM transition distance (r_0 in (6), and $r_1 = r_0 - 1$ Å). Starting around ~4.5 Å, the AIFF reproduces the QM to within 0.1 kJ/mol. Neglecting the MM terms entirely leads to much larger errors [58]. Adapted with permission of the PCCP Owner Societies

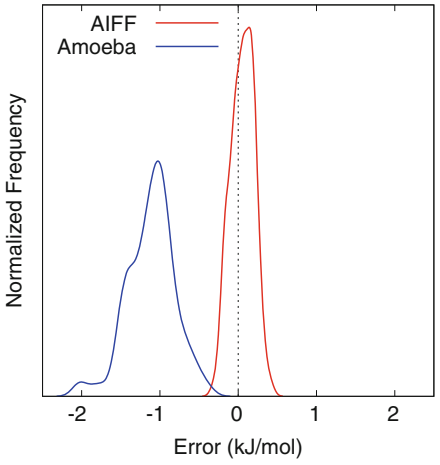


Fig. 3 Performance of the AIFF for many-body induction: error distributions for the AIFF and Amoeba force field many-body induction relative to MP2 for 101 $(\text{formamide})_8$ geometries. The AIFF errors are smaller than the Amoeba ones and are nearly centered on zero. Adapted with permission from [77]. Copyright 2010 American Chemical Society

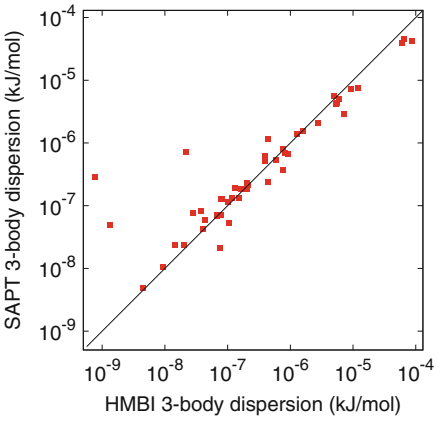


Fig. 4 Performance of the AIFF for three-body dispersion: Correlation between the HMBI three-body dispersion and the same contribution evaluated from symmetry adapted perturbation theory (SAPT) for 78 trimers taken from the formamide crystal [58]. Adapted with permission of the PCCP Owner Societies

need to be longer (e.g., ~7 Å in formamide), but the AIFF always behaves well at sufficiently long distances [58].

The AIFF also reproduces the QM many-body induction effects accurately [87]. Figure 3 shows the errors in the many-body induction in a set of 101 $(\text{formamide})_8$ geometries for the Amoeba force field and AIFF relative to RI-MP2. The Amoeba force field performs fairly well, but it systematically underestimates the many-body induction interactions by up to ~2 kJ/mol. The AIFF performs significantly better, with errors of roughly ±0.5 kJ/mol and a mean error very close to zero. Compared to Amoeba, the success of the AIFF stems from (1) its use of higher-rank multipoles (up to hexadecapole), (2) its use of higher-rank polarizabilities (up to quadrupole–quadrupole), and (3) the fact that these parameters are computed for each molecule in its current geometry rather than frozen at the values for some averaged/equilibrium geometry.

Finally, Fig. 4 demonstrates that the three-body dispersion model performs well compared to the three-body dispersion term in symmetry-adapted perturbation theory (SAPT). The simple, isotropic coupled Kohn–Sham dispersion coefficient representation provides a good approximation to the more complete SAPT calculation [58].

Overall, the AIFF does an excellent job of reproducing the QM interactions it replaces at much lower cost. HMBI predictions for the lattice energies of ammonia and carbon dioxide crystals differ from full periodic MP2 by only 1–2 kJ/mol, for instance [58].

2.3.1 AIFF Implementation

The long-range two-body electrostatics in the AIFF are implemented in standard fashion, with electrostatic interaction energy between two atoms A and B being given by

$$E_{\text{es}} \leftarrow \sum_{tu} Q_t^{\text{A}} T_{tu} Q_u^{\text{B}}, \tag{12}$$

where the T_{tu} matrix includes the distance- and orientation-dependent contributions for the interaction of two different spherical-tensor multipole moment components Q_t and Q_u. To evaluate the induction contributions, one first finds the induced multipole moments according to

$$\Delta Q_t^{\text{A}} = -\sum_{t'u} \alpha_{tt'}^{\text{A}} T_{t'u} \left(Q_u^{\text{B}} + \Delta Q_u^{\text{B}} \right), \tag{13}$$

where $\alpha_{tt'}^{\text{A}}$ is the static polarizability tensor on atom A and ΔQ is an induced multipole moment. Clearly the induced multipole moment on atom A depends on the induced multipole moment on atom B, so this process is done self-consistently until the induced multipoles reach convergence. Once the induced multipoles are

known, one can compute the induction energy contribution for the pair of atoms A and B according to

$$E_{\text{ind}} \leftarrow \sum_{tu} \Delta Q_t^{\text{A}} T_{tu} Q_u^{\text{B}} + Q_t^{\text{A}} T_{tu} \Delta Q_u^{\text{B}}. \tag{14}$$

The two-body dispersion between atoms A and B is evaluated via Casimir–Polder integration over the frequency-dependent polarizabilities:

$$E_{\text{2-body disp}} \leftarrow \sum_{tu} \sum_{t'u'} T_{tu} T_{t'u'} \int_0^\infty \alpha_{tt'}^{\text{A}}(i\omega) \alpha_{uu'}^{\text{B}}(i\omega) d\omega. \tag{15}$$

The integration is typically evaluated via numerical quadrature at ten frequencies ω. The expression in (15) corresponds to an anisotropic model for atom-atom dispersion. However, to a fairly good approximation, one can approximate this with a simple isotropic dispersion model (i.e., one which averages over the diagonal dipole–dipole and quadrupole–quadrupole elements of the frequency-dependent polarizability and neglects the off-diagonal elements). In that case, the dispersion model reduces to the standard C_6, C_8, etc., terms divided by the interatomic distance R to the corresponding power,

$$E_{\text{2-body disp}} \leftarrow \frac{C_6^{\text{AB}}}{R_{\text{AB}}^6} + \frac{C_8^{\text{AB}}}{R_{\text{AB}}^8} + \cdots. \tag{16}$$

The isotropic dispersion coefficients C_n are obtained from the Casimir–Polder integration over the appropriate elements of the isotropic frequency-dependent polarizabilities $\overline{\alpha}$:

$$C_n^{\text{AB}} \propto \int_0^\infty \overline{\alpha}^{\text{A}}(i\omega) \overline{\alpha}^{\text{B}}(i\omega) d\omega. \tag{17}$$

The many-body dispersion is approximated using the leading Axilrod–Teller three-body dispersion contribution. In the isotropic case, the three-body dispersion between atoms A, B, and C is computed according to

$$E_{\text{3-body disp}} \leftarrow C_9^{\text{ABC}} \frac{\left(1 + 3\cos\hat{A}\cos\hat{B}\cos\hat{C}\right)}{R_{\text{AB}}^3 R_{\text{BC}}^3 R_{\text{AC}}^3}, \tag{18}$$

where the R_{IJ} correspond to the distances between pairs of atoms and the cosines involve the interior angles of the triangle formed by the three atoms. The C_9 dispersion coefficient is obtained via Casimir–Polder integration over the frequency-dependent dipole–dipole polarizabilities on all three atoms:

$$C_9^{\text{ABC}} \propto \int_0^\infty \overline{\alpha}^{\text{A}}(i\omega) \overline{\alpha}^{\text{B}}(i\omega) \overline{\alpha}^{\text{C}}(i\omega) d\omega. \tag{19}$$

Note that most of the interaction terms described here also involve various empirical damping functions at short range to avoid divergences and the "polarization catastrophe," [78] but those are omitted here for simplicity.

The most difficult aspect of the force field implementation involves the evaluation of the $E_{\text{total}}^{\text{MM}}$ term in (7). The periodic crystal's induced multipoles are first computed self-consistently in a large, finite cluster consisting of the central unit cell and some number of periodic images (e.g., those within 25 Å). The algorithm works by inducing multipoles on the central unit cell molecules, replicating those induced moments on the periodic image molecules, and repeating the process until self-consistency is achieved. This procedure minimizes the edge effects arising from the finite cluster. Next, the permanent and induced multipole moments are combined and the overall interaction is evaluated via a multipolar Ewald sum [88]. Finally, the two-body and three-body dispersion contributions are evaluated via explicit lattice summation with large cutoffs. The evaluation of the AIFF interactions is cheap compared to the QM calculations, so one can afford to evaluate all of these contributions with tight cutoffs, thereby minimizing any errors due to truncating the lattice sums. See [78] for more details on the AIFF implementation and its performance for molecular crystals.

2.4 *Electronic Structure Treatment of the Intermolecular Interactions*

As the results described in Sect. 1 will demonstrate, high quality electronic structure methods need to be used for the QM one-body and two-body terms in HMBI. The electronic structure treatment must be capable of balancing the different types of intramolecular and intermolecular interactions to discriminate properly among different packing motifs. Ideally, one would employ high-level coupled cluster methods, like coupled cluster singles, doubles, and perturbative triples (CCSD(T)), with large basis sets, but this is usually impractical due to its N^7 scaling with system size N.

A more pragmatic approach will primarily use techniques that scale no more than N^5, like MP2. It is well known, however, that while MP2 describes van der Waals dispersion, it does so poorly [89–91]. Comparisons with intermolecular perturbation theory reveal the reasons for this poor performance: MP2 treats intermolecular dispersion interactions at the uncoupled HF (UCHF) level [92, 93]. In this view, the intermolecular dispersion interaction involves matrix elements of the intermolecular interaction between the ground state and excited state wavefunctions for molecules A and B divided by an energy denominator that depends on the excitation energies from ground to excited states:

$$E_{\text{disp}} = -\sum_{ab} \frac{< \psi_{\text{A}}^0 \psi_{\text{B}}^0 \left| \hat{V} \right| \psi_{\text{A}}^a \psi_{\text{B}}^b >}{E_{\text{A}}^a - E_{\text{A}}^0 + E_{\text{B}}^b - E_{\text{B}}^0}. \tag{20}$$

In UCHF (and MP2) these excited states for ψ^a and excitation energies $E^a - E^0$ are approximated using unrelaxed, ground-state HF orbitals and Fock matrix orbital eigenvalue energy differences.

Better results can be obtained by replacing the UCHF treatment of intermolecular dispersion with a coupled HF (CHF) [94] or coupled Kohn–Sham (CKS) [95, 96] one in which the excited states and excitation energies are computed with time-dependent HF or time-dependent DFT. Or one can obtain improved dispersion coefficients through other means [97]. The CKS variant has proved particularly successful, and is known as MP2C:

$$E_{\mathrm{MP2C}} = E_{\mathrm{MP2}} - E_{\mathrm{disp}}^{\mathrm{UCHF}} + E_{\mathrm{disp}}^{\mathrm{CKS}}. \tag{21}$$

MP2C works very well for a variety of intermolecular interaction types across a wide spectrum of intermolecular separations and orientations [89, 91, 96, 98–102]. The dispersion correction scales N^4, so MP2C retains the overall fifth-order scaling of MP2. However, the prefactor for the dispersion correction is relatively large, and it consumes a non-trivial amount of computer time. For example, computing the HMBI RI-MP2C/aug-cc-pVTZ single-point energy of the aspirin crystal requires ~2,450 h for the MP2 and ~390 h for the dispersion correction. So while the underlying MP2 calculation clearly dominates, accelerating the MP2C dispersion correction would be beneficial.

MP2C appears to work very well for molecular crystal problems [70, 103]. Of course, other alternatives for accurately describing intermolecular interactions with similar computational cost exist, including spin-component-scaled MP2 methods [104–106], dispersion-weighted MP2 [107], and van der Waals-corrected density functionals. Recent improvements in the random-phase approximation (RPA) are also promising and may prove useful in the near future [108–111].

2.4.1 Faster MP2C for Molecular Crystals

Evaluating the MP2C dispersion correction requires computing UCHF and CKS density–density-response functions χ for each monomer at a series of frequencies ω,

$$[\chi_0(\omega)]_{PQ} = -4\sum_{ia} \frac{(P|ia)\epsilon_{ia}(ia|Q)}{\epsilon_{ia}^2 + \omega^2}, \tag{22}$$

and then computing the dispersion energy via Casimir–Polder integration [96]:

$$E_{\mathrm{disp}} = -\frac{1}{2\pi}\int_0^\infty d\omega \widetilde{\chi}^{\mathrm{A}}(\omega)\mathbf{J}^{\mathrm{AB}}\left(\widetilde{\chi}^{\mathrm{B}}(\omega)\right)^T\left(\mathbf{J}^{\mathrm{AB}}\right)^T. \tag{23}$$

In these expressions, i and a are occupied and virtual molecular orbitals, P and Q are auxiliary basis functions, ϵ_{ia} is the difference between the HF orbital energies for orbitals i and a, $J^{\mathrm{AB}} = (P^{\mathrm{A}}|r_{12}^{-1}|Q^{\mathrm{B}})$, $\widetilde{\chi} = \mathbf{S}^{-1}\chi\mathbf{S}^{-1}$, and $S = (P|r_{12}^{-1}|Q)$. For CKS, one must obtain the coupled density response functions from the uncoupled one in (22) by solving the Dyson equation [96].

Equations (22) and (23) are typically computed using a so-called dimer-centered (DC) basis, in which the calculations on the isolated monomers are carried out in the presence of ghost basis functions on the other monomer it is interacting with. This has two disadvantages for fragment-based calculations in molecular crystals. First, the ghost basis functions substantially increase the basis set size and therefore the computational cost. Second, when computing the many different pairwise intermolecular interaction energies, one must recompute each monomer's response function repeatedly for each different dimer interaction in which it is involved.

However, while the calculation of accurate CKS or UCHF dispersion energies requires large basis sets [112], it turns out that the energy difference between them is much less sensitive to the basis set [103]. In fact, one obtains nearly identical results when using a comparable monomer-centered (MC) basis (i.e., with no ghost basis functions). This indicates that the basis-set dependence of UCHF and CKS dispersion are similar, and the higher-order basis-set effects are well described in the DC UCHF dispersion that is inherent in the supermolecular MP2 calculation. It is unnecessary to replace those contributions with CKS ones. Figure 5 demonstrates the excellent agreement between the MP2C results obtained in monomer-centered and dimer-centered basis sets, nearly independent of whether one uses a double-zeta basis or extrapolated complete-basis-set limit.

In practice, switching to an MC basis accelerates the calculation of the dispersion correction for a single dimer roughly fivefold. Much more dramatic savings can be obtained in a molecular crystal, however. Thanks to space group symmetry and the use of periodic boundary conditions, typical molecular crystal unit cells contain no more than a handful of symmetry-unique monomers. Therefore, one needs only calculate the monomer density–density response functions in (22) for those few unique monomers. Then one can compute the dispersion interaction for each dimer according to (23) with trivial effort. For instance, using this approach reduces the computational time for evaluating the MP2C/aug-cc-pVTZ dispersion correction in crystalline aspirin by two orders of magnitude, from 390 CPU hours to only 2.8 CPU hours. At the same time, the MC MP2C approach affects the relative energies of the two polymorphs by no more than a few tenths of a kJ/mol relative to the DC MP2C values (e.g., Fig. 6). For all practical purposes, this approach makes computing the MP2C correction "free" for molecular crystals. With its high accuracy and low cost, MP2C makes an excellent choice for describing the non-covalent interactions in molecular crystals.

Of course, the aforementioned speed-ups only affect the evaluation of the dispersion correction. As noted previously, the underlying RI-MP2/aug-cc-pVTZ calculations in aspirin require ~2,450 CPU hours. Further computational savings must come from the MP2 part such as through local MP2 or other similar methods. One must be careful, however, that the truncation schemes do not hinder the ability to describe the non-covalent interactions at intermediate distances (i.e., beyond nearest-neighbor interactions but not yet far enough apart to be treated by the force field in HMBI).

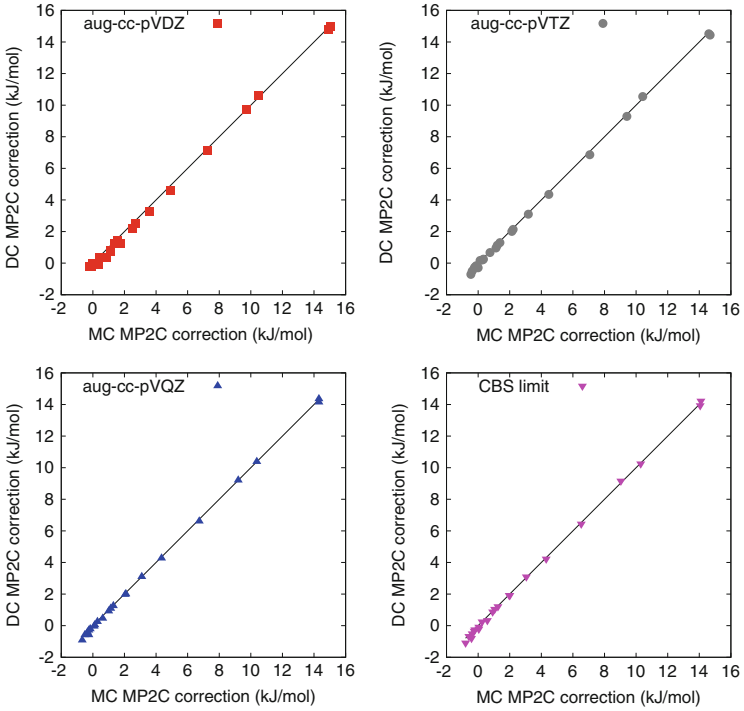


Fig. 5 For the dimers in the S22 test set [139], the MP2C dispersion correction calculated in an MC basis is nearly identical to that computed in a DC basis, even for a small aug-cc-pVDZ basis. Adapted with permission from [103]. Copyright 2013, American Institute of Physics

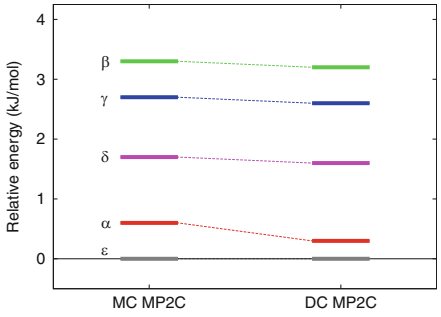


Fig. 6 Comparison between MC MP2C and DC MP2C for the relative energies of the five experimentally known polymorphs of oxalyl dihydrazide. The relative polymorph energies differ by 0.3 kJ/mol or less. Adapted with permission from [103]. Copyright 2013, American Institute of Physics

2.4.2 Basis Sets

It is well know that describing intermolecular interactions with correlated wave function methods requires large basis sets, so it is no surprise that molecular crystal energetics can often be very sensitive to the basis sets. This is particularly true for cases of conformational polymorphism, where the crystal packing motifs differ in both intramolecular conformation and intermolecular packing. Counterpoise corrections can help correct for basis set superposition error (BSSE). Different conformational polymorphs, however, may exhibit varying degrees of intramolecular BSSE, which is much harder to correct. In the oxalyl dihydrazide example discussed in Sect. 3.3, for example, the qualitative ordering of the experimental polymorphs differs dramatically as a function of basis set for exactly this reason. The intramolecular conformations of certain functional groups, such as the pyramidalization of nitrogen groups, can also be sensitive to basis sets.

In general, one should take care to ensure that molecular crystal calculations are converged with respect to basis set. Basis set extrapolation toward the complete basis set (CBS) limit, where feasible, can be very useful. One can also employ explicitly correlated techniques like MP2-F12 to achieve large-basis accuracy at reasonable computational cost [113, 114].

3 Performance and Applications of HMBI

To demonstrate the capabilities of the HMBI method, this section discusses molecular crystal benchmarks for predicting crystal geometries and lattice energies (Sect. 3.1) along with applications to polymorphic aspirin (Sect. 3.2) and oxalyl dihydrazide (Sect. 3.3) crystals.

The ability to predict crystal lattice energies accurately provides a demanding test for any theoretical treatment of molecular crystals. Lattice energies measure the energy required to dissociate the crystal to isolated gas-phase molecules. Whereas relative polymorph energies allow for some degree of error cancellation between the treatments of two different sets of non-covalent interactions, lattice energies expose any errors in calculating the strength of those interactions (at least for small, rigid molecules for which the intramolecular geometry is similar in the gas and crystal phases). The situation is analogous to the differences between computing reaction energies vs atomization energies, with the former being much easier due to cancellation of errors. In molecular crystals, the degree of error cancellation in the relative energies between polymorphs will depend on the differences in intramolecular configurations and intermolecular packing. In cases like aspirin, where the two polymorphs exhibit similar structures, substantial error cancellation occurs. On the other hand, cases like oxalyl dihydrazide exhibit more diverse interactions and packing modes, leading to less error cancellation and making it much more difficult to obtain accurate polymorphic energies.

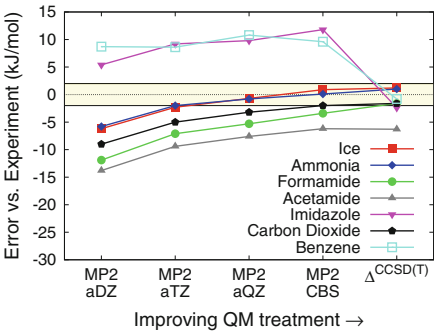


Fig. 7 Benchmark lattice energy predictions compared to experiment (in kJ/mol). Improving the QM treatment systematically improves the lattice energy predictions. The *shaded region* highlights ±2 kJ/mol to indicate the excellent agreement with the nominal experimental values. See [78] for complete details. Adapted with permission from [78]. Copyright 2011 American Chemical Society

3.1 Predicting Molecular Crystal Lattice Energies and Geometries

3.1.1 Lattice Energies

Benchmark lattice energy predictions have been performed on a series of seven small-molecule crystals: ammonia, acetamide, benzene, carbon dioxide, formamide, ice Ih, and imidazole. These crystals were chosen because they span a diverse range of intermolecular interactions, ranging from hydrogen-bonded to van der Waals dispersion bonded. The calculations start in a relatively small basis (aug-cc-pVDZ) RI-MP2, then increase the basis set toward the TQ-extrapolated basis set limit, and then finally examine higher-order correlation effects evaluated with CCSD(T) in a modest basis. Initially the Amoeba force field was used for the MM terms [115], but even better results are obtained when the AIFF (in the Sadlej basis set [116, 117]) described in Sect. 2.3 is used [78].

The primarily hydrogen-bonded crystals exhibit generally consistent improvement toward the experimental lattice energies as the electronic structure treatment of the one-body and short-range two-body terms is improved, as shown in Fig. 7. For those cases, the higher-order CCSD(T) correlation contributions are generally small. However, for the benzene and imidazole crystals, where the π-electron van der Waals dispersion interactions are significant, MP2 over-binds the crystals by ~10–12 kJ/mol (~10–20%). This sort of error is particularly problematic in the context of molecular crystal polymorphism, since MP2 will be biased toward such dispersion-bound packing motifs over hydrogen-bonded ones.

The CCSD(T) results correct the MP2 overestimate of the dispersion interactions in these two crystals. Overall, for six of the seven crystals, the estimated CBS-limit CCSD(T) results lie within 1–2 kJ/mol of the nominal experimental values, which

Table 1 Molecular crystal lattice energies in kJ/mol. MP2C/CBS-limit performs much better than MP2/CBS-limit relative to both the estimated complete-basis-set CCSD(T) results and experiment

Crystal	MP2[a]	MP2C[b]	CCSD(T)[a]	Experiment[c]
Ice (Ih)	59.9	60.3	60.2	59
Ammonia	39.3	40.5	40.2	39
Formamide	78.6	78.7	80.4	82
Acetamide	79.8	79.8	79.7	86
Carbon dioxide	29.1	26.3	29.5	31
Imidazole	102.8	93.1	88.6	91
Benzene		48.6	61.6	52

[a]Wen and Beran [78]
[b]Huang et al. [103]
[c]Beran and Nanda [115]

is likely within the experimental errors. The reason for the larger errors in acetamide is unclear. Of course, CCSD(T) calculations will often be cost-prohibitive in molecular crystals. MP2C can provide a lower-cost alternative that substantially improves upon MP2. Table 1 demonstrates that including the MP2C dispersion correction has relatively small effects on the hydrogen-bonded crystals, but it reduces the large MP2 errors for the imidazole and benzene crystals from ~10–12 to ~2–3 kJ/mol, which is only slightly worse than the CCSD(T) results; see [103] for further details.

3.1.2 Crystal Structures

The HMBI model also provides accurate crystal geometries. Fully relaxed geometry optimizations at the HMBI counterpoise-corrected RI-MP2/aug-cc-pVDZ and Amoeba MM level for several of the benchmark crystals described above reproduced the experimental structures fairly accurately, including the space group symmetry [73]. Root-mean-square errors in the unit cell lattice parameters are only 1.6% relative to low-temperature crystal structures (100 K or below). For comparison, the same errors with B3LYP-D* [118] are 3.4% in the 6-31G** basis and 2.0% in the TZP basis. The HMBI RI-MP2 calculations also perform well for the root-mean-square deviations in the atomic positions (Table 2). Figure 8 shows overlays of the experimental and HMBI RI-MP2 optimized structures, highlighting the good agreement between them.

Overall, the RI-MP2/aug-cc-pVDZ structures are clearly better than dispersion-corrected B3LYP-D*/6-31G* [118]. They are slightly worse than those obtained with B3LYP-D* in a triple-zeta basis (TZP), though the MP2/aug-cc-pVDZ results exhibit slightly more uniform errors. In principle, larger-basis MP2 geometries ought to be even better, but optimizing in those larger basis sets becomes even more expensive. One must also worry about the effects of the MP2 treatment of dispersion on the geometries.

Given the sorts of practical calculations that are currently feasible, it is not obvious that fragment methods provide a significant advantage relative to DFT for molecular crystal structure optimization, especially since the basis set requirements

Table 2 Root-mean-square deviations ($rmsd_{15}$ [140] in Å) between the theoretically optimized and experimental geometries for clusters of 15 molecules taken from each crystal. The temperatures at which the experimental structures was obtained is also listed. See [73] for details

Crystal	B3LYP-D*/ 6-31G**	HMBI RI-MP2/ aug-cc-pVDZ	B3LYP-D*/ TZP	Experimental temperature (K)
Ice (Ih)	0.13	0.10	0.04	15
Formamide	0.29	0.16	0.22	90
Acetamide	0.16	0.08	0.08	23
Imidazole	0.20	0.12	0.14	103
Benzene	0.09	0.06	0.02	4

for correlated wavefunction methods like MP2 are steeper than those for DFT. For this reason, we often optimize the geometry using DFT and perform high-level single-point energies using HMBI. Of course, further improvements in MP2-like algorithms and continuing advances in computer hardware may change this.

3.2 Aspirin Polymorphism

The crystal structure of aspirin has been known since the 1960s [119], but suspicion of a second crystal form lingered for many years until 2004, when the structure of aspirin form II was predicted by the Price group [120]. Experimental confirmation for form II came a year later [121], though this report was followed by several more years of controversy in the crystallographic community [122, 123]. Only in the past couple years has the existence and structure of form II aspirin been firmly established [124–126].

The controversy arose from the fact that the two aspirin polymorphs exhibit very similar structures, and it proved very difficult to obtain pure crystals of form II. Rather, one often obtains mixed crystals containing domains of both forms I and II. The overall crystal packing of the two polymorphs is very similar except for the nature of the interlayer hydrogen bonding (Fig. 9). Whereas form I exhibits dimers, with each acetyl group hydrogen bonding to one other aspirin molecule in the adjacent layer, form II exhibits a catemeric structure, where each acetyl group hydrogen bonds to two adjacent molecules. This catemeric structure produces long chains of hydrogen bonds in form II.

Interestingly, earlier dispersion-corrected DFT studies suggested that form II was ~2–2.5 kJ/mol more stable than form I [127, 128], which would be surprising for two crystals that appear to grow together so readily. Other examples of crystals which form intergrowths are often separated by less than 1 kJ/mol in energy [129–131].

We investigated this system [132] using HMBI after optimizing the crystal structures for each form using B3LYP-D*/TZP. We performed a variety of calculations, including MP2, SCS(MI)-MP2 [104], and MP2C in both aug-cc-pVDZ and

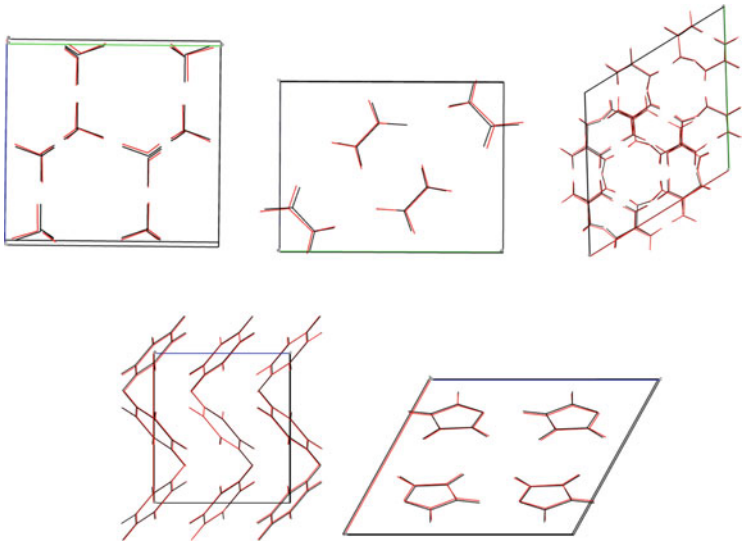


Fig. 8 Overlays of the experimental (*black*) and HMBI MP2 structures (*red*) for (clockwise from *top left*) ice Ih, formamide, acetamide, imidazole, and benzene. The experimental unit cell boundaries are drawn. Adapted with permission from [73]. Copyright 2012, American Institute of Physics

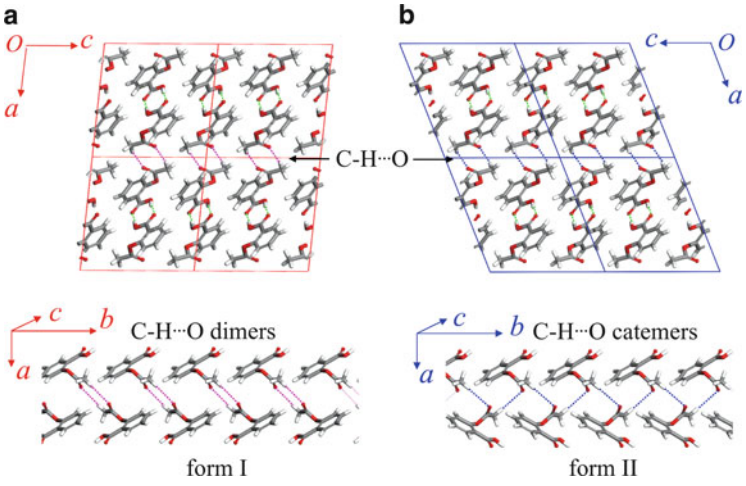


Fig. 9 Aspirin forms I and II exhibit very similar crystal packing. The key difference lies in whether the interlayer hydrogen bonding occurs as dimers (form I) or catemers (form II). Reprinted with permission from [132]. Copyright 2012 American Chemical Society

aug-cc-pVTZ basis sets. In all cases, the energy differences between the two polymorphs were only a few tenths of a kJ/mol, confirming the expected near-degeneracy of the two forms (Table 3). Interestingly, both MP2 and SCS(MI)-MP2

Table 3 Lattice energies and relative energy differences for the two polymorphs of aspirin (in kJ/mol). The two polymorphs are virtually degenerate across a range of model chemistries

	MP2[a]	SCS(MI) MP2[a]	MP2C[b]
aug-cc-pVDZ			
Form I	113.7	132.5	–
Form II	113.5	132.3	–
$\Delta E_{\mathrm{I} \rightarrow \mathrm{II}}$[c]	0.2	0.1	–
aug-cc-pVTZ			
Form I	132.1	135.6	116.1
Form II	132.0	135.5	116.3
$\Delta E_{\mathrm{I} \rightarrow \mathrm{II}}$[c]	0.1	0.0	−0.1

[a]Wen and Beran [132]
[b]Huang et al. [103]
[c]The apparent discrepancies between the lattice energies and the ΔE values arise from rounding

appear to overestimate significantly the ~115 kJ/mol experimental lattice energy [128, 133], predicting aug-cc-pVTZ values of 132 and 136 kJ/mol, respectively. In contrast, MP2C predicts a lattice energy of 116 kJ/mol, in much better agreement with the experimental value [103].

In addition to predicting the virtual degeneracy of the two aspirin polymorphs, physical insight into the nature of the polymorphism was obtained by decomposing the relative energies of the two polymorphs according to the different contributions in the HMBI many-body expansion. In particular, we found that while the intramolecular interactions favor form I, the intermolecular interactions favor form II (Fig. 10). The key differences arise from the nature of the interlayer hydrogen bonding. In form I, the acetyl group hydrogen bonds to only one adjacent aspirin molecule, which allows the acetyl group to adopt a slightly more favorable intramolecular conformation. In contrast, the catemeric hydrogen bonding in form II forces the acetyl group to adopt a conformation that is slightly less favorable, but in doing so it forms much better hydrogen bonds and achieves hydrogen bond cooperativity through the extended hydrogen-bond networks. It turns out that the energy differences in each case amount to about 1.5 kJ/mol and cancel one another almost perfectly, making the two polymorphs virtually and "accidentally" degenerate.

From a theoretical perspective, one other key feature emerged from this study: the strong similarity between the crystal packing in both polymorphs leads to excellent cancellation of errors in predicting the relative polymorph energetics. Unfortunately, such thorough error cancellation is probably the exception rather than the rule, as demonstrated by the case of oxalyl dihydrazide which is discussed in the next section.

3.3 Oxalyl Dihydrazide Polymorphism

Oxalyl dihydrazide provides another interesting example of molecular crystal polymorphism. It exhibits five experimentally known polymorphs, denoted $\alpha - \epsilon$, that differ in their degree of intramolecular and intermolecular hydrogen bonding

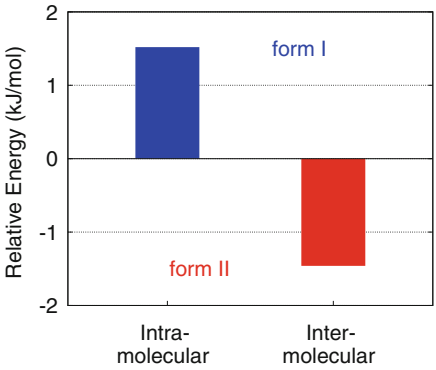


Fig. 10 Aspirin provides a classic example of conformational polymorphism: form I adopts a slightly more favorable intramolecular conformation, while form II exhibits stronger intermolecular interactions. The two effects have nearly identical energies, and the two polymorphs are virtually degenerate. Reprinted with permission from [132]. Copyright 2012 American Chemical Society

(Fig. 11). The relative stabilities of the five polymorphs is not known experimentally, but it is believed that the trend follows α, δ, $\epsilon < \gamma < \beta$ [134]. That is, the β form is the least stable, followed by the γ form. The other three are more stable, though their energetic ordering is unknown.

Earlier force field and DFT calculations had significant trouble reproducing these qualitative trends. The relative polymorph ordering varied widely with the choice of density functional; see, for example, [135]. Most of those functionals lacked van der Waals dispersion corrections, however. The empirical dispersion-corrected D-PW91 functional, which has been very successful in the blind tests of crystal structure prediction and elsewhere [12, 15, 87, 136–138], does however obtain a plausible ordering for the polymorphs ($\alpha < \epsilon < \delta < \gamma < \beta$), but the overall energy range of the polymorphs is ~15 kJ/mol, which is somewhat larger than the <10 kJ/mol typically found for experimentally observable polymorphs. On the other hand, a different dispersion corrected function, B3LYP-D* [118], gives a very different ordering that is inconsistent with experiment [70]. In other words, obtaining reasonable results for this system is not simply a matter of including dispersion. Rather, the energetics are very sensitive to the specific treatment of the electronic structure and the balance between intermolecular and intramolecular interactions. Note that vibrational zero-point energy is also important in oxalyl dihydrazide, and it is included in the results shown here. Finite-temperature effects may also be significant, but no attempt has been made to estimate them.

Obtaining plausible predictions for the relative polymorph energies proved challenging, even for correlated wavefunction methods (Fig. 12) [70]. In small basis sets, MP2 predicts the α form to be the *least* stable. However, increasing the basis set size toward the CBS limit dramatically reorders the polymorphs, preferentially stabilizing the α form. The slow basis set convergence in oxalyl

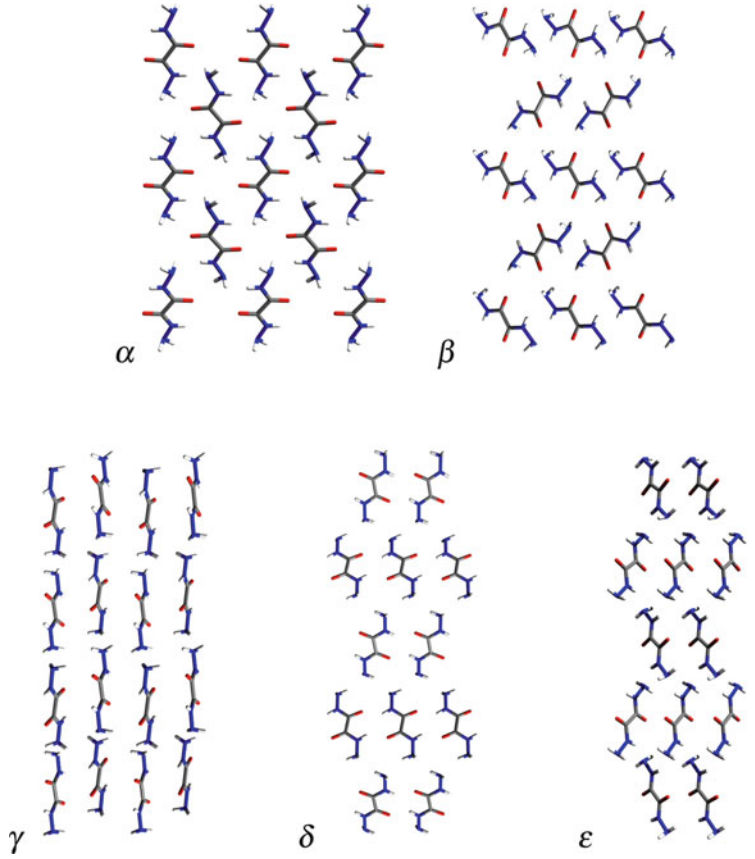


Fig. 11 The five experimentally known polymorphs of oxalyl dihydrazide

dihydrazide provides a sharp contrast to aspirin, where the relative energy difference between the two polymorphs was insensitive to the basis set.

The basis set sensitivity in oxalyl dihydrazide arises from BSSE. The α polymorph exhibits only intermolecular hydrogen bonds, while the other four forms contain mixtures of intermolecular and intramolecular hydrogen bonds. Applying a standard counterpoise correction to each QM dimer calculation helps correct the intermolecular BSSE, but it is harder to correct intramolecular BSSE. Therefore, the α polymorph, which has much less intramolecular BSSE, is destabilized relative to the other four forms in small basis sets. As the basis set size increases, the intramolecular BSSE decreases in the other four forms, and the balance between intermolecular and intramolecular hydrogen bonding is restored.

Beyond the basis set effects, oxalyl dihydrazide exhibits some π-stacking type van der Waals dispersion interactions, so it comes as no surprise that using MP2C instead of MP2 also has a significant effect. The dispersion correction reorders the β and γ polymorphs to the correct experimental ordering. It also stabilizes the α

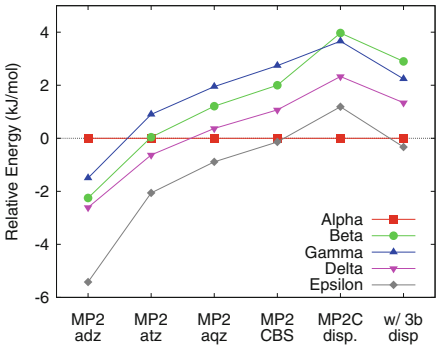


Fig. 12 Relative polymorph energies for oxalyl dihydrazide as a function of the model chemistry. Basis set effects, improved MP2C two-body dispersion, and AIFF three-body dispersion are all important. Adapted with permission from [70]. Copyright 2012 American Chemical Society

form substantially relative to the other polymorphs. The more empirical SCS(MI)-MP2 did not perform as well as MP2C in this system [70], perhaps due to the major differences in the optimal correlation energy scaling factors for intra- and intermolecular interactions.

Finally, Axilrod–Teller three-body dispersion effects prove important here. They destabilize the α form relative to the other four forms. The three-body dispersion contributions here are repulsive and die off with R^{-9}. Therefore, they are more repulsive in the dense α polymorph (1.76 g/cm^3) than in the four other, less dense polymorphs (1.59–1.66 g/cm^3). Without three-body dispersion, the calculations predict that the α polymorph is the most stable. However, including the three-body dispersion terms makes the ϵ polymorph slightly more stable than the α one, in contrast with the best dispersion-corrected DFT predictions. The predicted energetic preference for the ϵ polymorph over the α one is slight, and no experimental data currently exists to help resolve the issue. Nevertheless, it is interesting to note that the empirically dispersion-corrected DFT results in [70, 135] include only two-body dispersion and favor the α polymorph, just like the HMBI results without three-body dispersion. From this perspective it would be interesting to see what state-of-the-art many-body dispersion-corrected density functionals predict here.

4 Conclusions and Outlook

Fragment-based methods like HMBI provide a computationally viable means of achieving high-accuracy structures and lattice energies for chemically interesting molecular crystals. The systems examined here highlight the challenges inherent to modeling molecular crystals, where the subtle energetic competitions can require careful electronic structure treatments. It will not always be a priori obvious how elaborate an electronic structure treatment is needed for a given system. One can sometimes count on cancellation of errors, especially when the different crystal packing motifs are similar (like in aspirin), but much less error cancellation occurs in other cases (like oxalyl dihydrazide). Unfortunately, the reduced error-cancellation case is probably more typical for conformational polymorphs of flexible organic molecules, especially as the molecules become larger.

Periodic DFT methods can often provide acceptable accuracy for molecular crystal systems with reasonable computational costs. In many cases, however, one will wish to assess the reliability of those predictions or to obtain even more accurate predictions. From this perspective, one of the key strengths of fragment methods is their ability to improve systematically the quality of the results with respect to the underlying electronic structure treatments. The ability to demonstrate convergence of the predictions with respect to model chemistry is a powerful tool for making robust predictions. Moreover, for systems with large numbers of molecules in the unit cell, the inherently linear-scaling nature of the computationally dominant QM calculations in fragment methods will often make them cheaper than full periodic DFT calculations. Fragment methods also provide a natural decomposition scheme for understanding the nature of the energetic competitions in polymorphic crystals.

For evaluating the intermolecular interactions in molecular crystals, the dispersion-corrected MP2C method provides a useful balance between accuracy and efficiency, offering near coupled-cluster-quality results at MP2-like cost. In practice, MP2C (or its explicitly correlated variant, MP2C-F12), with an aug-cc-pVTZ basis, is practical for molecular crystals containing two to three dozen atoms per molecule and a handful of molecules in the asymmetric unit cell. In other words, these techniques are applicable to a number of interesting, small-molecule pharmaceuticals, organic semiconductors, and energetic materials (though many more such materials remain impractical for the moment!).

Thus, fragment methods like HMBI will likely play an important role in molecular crystal modeling for years to come. The next advances are likely to come from a couple of directions. First, further efficiency improvements will enable the application to larger systems. For instance, the MP2C timings described above indicate that the vast majority of the computational time is consumed on the MP2 calculations. A sizable fraction of that time comes from evaluating the long-range interactions, which are then largely discarded through the MP2C dispersion correction. New, more efficient strategies for achieving similar quality results can surely be developed. In addition, further improvements to the ab initio force field would enable one to treat fewer dimers quantum mechanically, thereby accelerating the calculations. This requires incorporating efficient treatments of the short-rate exchange interactions into the force field.

Second, one needs to consider finite-temperature entropies and free energies. This can be done through quasi-harmonic-type approximations or through dynamical free energy calculations. The latter is potentially more rigorous and capable of capturing anharmonic effects, but of course it is limited by the need for extensive computational sampling and the quality of the force fields that can be used to perform such sampling affordably. Quasiharmonic calculations are much less expensive computationally, but they involve their own severe approximations. Much more research is needed in this area to determine which techniques are useful under which circumstances. In other words, many theoretical opportunities remain in molecular crystal modeling to occupy researchers for quite some time!

Acknowledgements Funding for this work from the National Science Foundation (CHE-1112568) and supercomputer time from XSEDE (TG-CHE110064) are gratefully acknowledged.

References

1. Bauer J, Spanton S, Quick R, Quick J, Dziki W, Porter W, Morris J (2001) Ritonavir: an extraordinary example of conformational polymorphism. Pharm Res 18:859–866
2. Chemburkar SR, Bauer J, Deming K, Spiwek H, Patel K, Morris J, Henry R, Spanton S, Dziki W, Porter W, Quick J, Bauer P, Donaubauer J, Narayanan BA, Soldani M, Riley D, Mcfarland K (2000) Dealing with the impact of ritonavir polymorphs on the late stages of bulk drug process development. Org Process Res Dev 4(5):413–417
3. Raw AS, Furness MS, Gill DS, Adams RC, Holcombe FO, Yu LX (2004) Regulatory considerations of pharmaceutical solid polymorphism in abbreviated new drug applications (ANDAs). Adv Drug Deliv Rev 56(3):397–414. doi:10.1016/j.addr.2003.10.011
4. Goldbeck G, Pidcock E, Groom C (2012) Solid form informatics for pharmaceuticals and agrochemicals: knowledge based substance development and risk assessment. http://www.ccdc.cam.ac.uk/Lists/ResourceFileList/Solid Form informatics .pdf. Accessed 28 June 2013
5. Bernstein J (2002) Polymorphism in molecular crystals. Clarendon, Oxford
6. Roth K (2005) Von Vollmilch bis Bitter, edelste Polymorphie. Chem Unserer Zeit 39:416–428
7. Politzer P, Murray JS (eds) (2003) Energetic materials: part 1. Decomposition, crystal, and molecular properties. Elsevier, Amsterdam
8. Politzer P, Murray JS (eds) (2003) Energetic materials: part 2. Detonation, combustion. Elsevier, Amsterdam
9. Haas S, Stassen AF, Schuck G, Pernstich KP, Gundlach DJ, Batlogg B, Berens U, Kirner HJ (2007) High charge-carrier mobility and low trap density in a rubrene derivative. Phys Rev B 76:115203
10. Karamertzanis PG, Kazantsev AV, Issa N, Welch GWA, Adjiman CS, Pantelides CC, Price SL (2009) Can the formation of pharmaceutical cocrystals be computationally predicted? 2 Crystal structure prediction. J Chem Theory Comput 5:1432–1448
11. Kazantsev AV, Karamertzanis PG, Adjiman CS, Pantelides CC, Price SL, Galek PTA, Day GM, Cruz-Cabeza AJ (2011) Successful prediction of a model pharmaceutical in the fifth blind test of crystal structure prediction. Int J Pharm 418:168–178. doi:10.1016/j.ijpharm.2011.03.058
12. Kendrick J, Leusen FJJ, Neumann MA, van de Streek J (2011) Progress in crystal structure prediction. Chem Eur J 17(38):10736–10744. doi:10.1002/chem.201100689
13. Neumann MA (2008) Tailor-made force fields for crystal-structure prediction. J Phys Chem B 112(32):9810–9829. doi:10.1021/jp710575h
14. Neumann MA, Perrin MA (2005) Energy ranking of molecular crystals using density functional theory calculations and an empirical van der Waals correction. J Phys Chem B 109:15531–15541
15. Neumann MA, Leusen FJJ, Kendrick J (2008) A major advance in crystal structure prediction. Angew Chem Int Ed 47:2427–2430
16. Bardwell DA, Adjiman CS, Arnautova YA, Bartashevich E, Boerrigter SXM, Braun DE, Cruz-Cabeza AJ, Day GM, Della Valle RG, Desiraju GR, van Eijck BP, Facelli JC, Ferraro MB, Grillo D, Habgood M, Hofmann DWM, Hofmann F, Jose KVJ, Karamertzanis PG, Kazantsev AV, Kendrick J, Kuleshova LN, Leusen FJJ, Maleev AV, Misquitta AJ, Mohamed S, Needs RJ, Neumann MA, Nikylov D, Orendt AM, Pal R, Pantelides CC, Pickard CJ, Price LS, Price SL, Scheraga HA, van de Streek J, Thakur TS, Tiwari S, Venuti E, Zhitkov IK (2011) Towards

crystal structure prediction of complex organic compounds – a report on the fifth blind test. Acta Crystallogr B 67:535–551. doi:10.1107/S0108768111042868
17. Day GM, Cooper TG, Cruz-Cabeza AJ, Hejczyk KE, Ammon HL, Boerrigter SXM, Tan JS, Della Valle RG, Venuti E, Jose J, Gadre SR, Desiraju GR, Thakur TS, van Eijck BP, Facelli JC, Bazterra VE, Ferraro MB, Hofmann DWM, Neumann MA, Leusen FJJ, Kendrick J, Price SL, Misquitta AJ, Karamertzanis PG, Welch GWA, Scheraga HA, Arnautova YA, Schmidt MU, van de Streek J, Wolf AK, Schweizer B (2009) Significant progress in predicting the crystal structures of small organic molecules – a report on the fourth blind test. Acta Crystallogr B 65(Pt 2):107–125. doi:10.1107/S0108768109004066
18. Zhu Q, Oganov AR, Glass CW, Stokes HT (2012) Constrained evolutionary algorithm for structure prediction of molecular crystals: methodology and applications. Acta Crystallogr B 68:215–226. doi:10.1107/S0108768112017466
19. Price SL (2008) Computational prediction of organic crystal structures and polymorphism. Int Rev Phys Chem 27(3):541–568. doi:10.1080/01442350802102387
20. Gavezzotti A, Filippini G (1995) Polymorphic forms of organic crystals at room conditions: thermodynamic and structural implications. J Am Chem Soc 117:12299–12305
21. Otero-de-la Roza A, Johnson ER (2012) A benchmark for non-covalent interactions in solids. J Chem Phys 137(5):054103. doi:10.1063/1.4738961
22. Raiteri P, Martonák R, Parrinello M (2005) Exploring polymorphism: the case of benzene. Angew Chem Int Ed 44(24):3769–3773. doi:10.1002/anie.200462760
23. Schnieders MJ, Baltrusaitis J, Shi Y, Chattree G, Zheng L, Yang W, Ren P (2012) The structure, thermodynamics and solubility of organic crystals from simulation with a polarizable force field. J Chem Theory Comput 8(5):1721–1736. doi:10.1021/ct300035u
24. Kazantsev AV, Karamertzanis PG, Adjiman CS, Pantelides CC (2011) Efficient handling of molecular flexibility in lattice energy minimization of organic crystals. J Chem Theory Comput 7:1998–2016
25. Price SL (2004) The computational prediction of pharmaceutical crystal structures and polymorphism. Adv Drug Deliv Rev 56(3):301–319. doi:10.1016/j.addr.2003.10.006
26. Price SL, Price LS (2011) Computational polymorph prediction. In: Storey R, Ymen I (eds) Solid state characterization of pharmaceuticals, 1st edn. Blackwell, London, pp 427–450
27. Price SL, Leslie M, Welch GWA, Habgood M, Price LS, Karamertzanis PG, Day GM (2010) Modelling organic crystal structures using distributed multipole and polarizability-based model intermolecular potentials. Phys Chem Chem Phys 12:8478–8490. doi:10.1039/c004055j
28. Day GM, Motherwell WDS, Ammon HL, Boerrigter SXM, Della Valle RG, Venuti E, Dzyabchenko A, Dunitz JD, Schweizer B, van Eijck BP, Erk P, Facelli JC, Bazterra VE, Ferraro MB, Hofmann DWM, Leusen FJJ, Liang C, Pantelides CC, Karamertzanis PG, Price SL, Lewis TC, Nowell H, Torrisi A, Scheraga HA, Arnautova YA, Schmidt MU, Verwer P (2005) A third blind test of crystal structure prediction. Acta Crystallogr B 61(Pt 5):511–527. doi:10.1107/S0108768105016563
29. Dion M, Rydberg H, Schröder E, Langreth DC, Lundqvist BI (2004) Van der Waals density functional for general geometries. Phys Rev Lett 92:246401
30. DiStasio RA, von Lilienfeld OA, Tkatchenko A (2012) Collective many-body van der Waals interactions in molecular systems. Proc Natl Acad Sci U S A 109:14791–14795. doi:10.1073/pnas.1208121109
31. Grimme S (2011) Density functional theory with London dispersion corrections. WIREs Comput Mol Sci 1(2):211–228. doi:10.1002/wcms.30
32. Otero-de-la Roza A, Johnson ER (2013) Many-body dispersion interactions from the exchange-hole dipole moment model. J Chem Phys 138(5):054103. doi:10.1063/1.4789421
33. Thonhauser T, Cooper VR, Li S, Puzder A, Hyldgaard P, Langreth DC (2007) Van der Waals density functional: self-consistent potential and the nature of the van der Waals bond. Phys Rev B 76:125112
34. Tkatchenko A, Scheffler M (2009) Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data. Phys Rev Lett 102(7):073005. doi:10.1103/PhysRevLett.102.073005

35. Vydrov OA, Van Voorhis T (2010) Nonlocal van der Waals density functional: the simpler the better. J Chem Phys 133(24):244103. doi:10.1063/1.3521275
36. Ayala PY, Kudin KN, Scuseria GE (2001) Atomic orbital Laplacetransformed second-order Møller–Plesset perturbation theory for periodic systems. J Chem Phys 115:9698–9707
37. Hirata S, Iwata S (1998) Analytical energy gradients in second-order Moller–Plesset perturbation theory for extended systems. J Chem Phys 109(11):4147–4155
38. Hirata S, Shimazaki T (2009) Fast second-order many-body perturbation method for extended systems. Phys Rev B 80(8):1–7. doi:10.1103/PhysRevB.80.085118
39. Izmaylov AF, Scuseria GE (2009) Resolution of the identity atomic orbital Laplace transformed second-order Møller–Plesset theory for nonconducting periodic systems. Phys Chem Chem Phys 10:3421–3429
40. Marsman M, Grueneis A, Paier J, Kresse G (2009) Second-order Møller–Plesset perturbation theory applied to extended systems. I. Within the projector-augmented-wave formalism using a plane wave basis set. J Chem Phys 130:184103
41. Maschio L, Usvyat D, Manby FR, Casassa S, Pisani C, Schutz M (2007) Fast local-MP2 method with density-fitting for crystals. I. Theory and algorithms. Phys Rev B 76:075101
42. Ohnishi YY, Hirata S (2010) Logarithm second-order many-body perturbation method for extended systems. J Chem Phys 133(3):034106. doi:10.1063/1.3455717
43. Pisani C, Maschio L, Casassa S, Halo M, Schutz M, Usvyat D (2008) Periodic local MP2 method for the study of electronic correlation in crystals: theory and preliminary applications. J Comput Chem 29:2113–2124
44. Shiozaki T, Hirata S (2010) Communications: explicitly correlated secondorder Møller–Plesset perturbation method for extended systems. J Chem Phys 132(15):151101. doi:10.1063/1.3396079
45. Suhai S (1983) Quasiparticle energy-band structures in semiconducting polymers: correlation effects on the band gap in polyacetylene. Phys Rev B 27:3506–3518
46. Sun JQ, Bartlett RJ (1996) Second-order many-body perturbation-theory calculations in extended systems. J Chem Phys 104:8553–8565
47. Usvyat D, Maschio L, Manby FR, Casassa S, Pisani C, Schutz M (2007) Fast local-MP2 method with density-fitting for crystals. II. Test calculations and applications to the carbon dioxide crystal. Phys Rev B 76:075102
48. Förner W, Knab R, Čižek J, Ladik J (1997) Numerical application of the coupled cluster theory with localized orbitals to polymers. IV. Band structure corrections in model systems and polyacetylene. J Chem Phys 106:10248–10264
49. Hirata S, Podeszwa R, Tobita M, Bartlett RJ (2004) Coupled-cluster singles and doubles for extended systems. J Chem Phys 120:2581–2592
50. Reinhardt P (2000) Dressed coupled-electron-pair-approximation methods for periodic systems. Theor Chem Acc 104:426–438
51. Yu M, Kalvoda S, Dolg M (1997) An incremental approach for correlation contributions to the structural and cohesive properties of polymers. Coupled-cluster study of trans-polyacetylene. Chem Phys 224:121–131
52. Fedorov DG, Kitaura K (2007) Extending the power of quantum chemistry to large systems with the fragment molecular orbital method. J Phys Chem A 111:6904–6914
53. Kitaura K, Ikeo E, Asada T, Nakano T, Uebayasi M (1999) Fragment molecular orbital method: an approximate computational method for large molecules. Chem Phys Lett 313:701–706
54. Yang W (1991) Direct calculation of electron density in density functional theory. Phys Rev Lett 66(11):1438–1441
55. Paulus B (2006) The method of increments - a wavefunction-based ab initio correlation method for solids. Phys Rep 428(1):1–52. doi:10.1016/j.physrep.2006.01.003
56. Stoll H (1992) Correlation energy of diamond. Phys Rev B 46:6700–6704
57. Gordon MS, Fedorov DG, Pruitt SR, Slipchenko L (2012) Fragmentation methods: a route to accurate calculations on large systems. Chem Rev 112:632–672. doi:10.1021/cr200093j

58. Wen S, Nanda K, Huang Y, Beran GJO (2012) Practical quantum mechanics-based fragment methods for predicting molecular crystal properties. Phys Chem Chem Phys 14:7578–7590. doi:10.1039/c2cp23949c
59. Beran GJO, Hirata S (2012) Fragment and localized orbital methods in electronic structure theory. Phys Chem Chem Phys 14:7559–7561. doi:10.121/cg300358n
60. Mayhall NJ, Raghavachari K (2012) Many-overlapping-body (MOB) expansion: a generalized many body expansion for nondisjoint monomers in molecular fragmentation calculations of covalent molecules. J Chem Theory Comput 8(8):2669–2675. doi:10.1021/ct300366e
61. Richard RM, Herbert JM (2012) A generalized many-body expansion and a unified view of fragment-based methods in electronic structure theory. J Chem Phys 137(6):064113. doi:10.1063/1.4742816
62. Dahlke EE, Truhlar DG (2006) Assessment of the pairwise additive approximation and evaluation of many-body terms for water clusters. J Phys Chem B 3:10595–10601
63. Dahlke EE, Truhlar DG (2007) Electrostatically embedded many-body correlation energy, with applications to the calculation of accurate second-order Møller–Plesset perturbation theory energies for large water clusters. J Chem Theory Comput 3:1342–1348
64. Dahlke EE, Truhlar DG (2007) Electrostatically embedded many-body expansion for large systems, with applications to water clusters. J Chem Theory Comput 3:46–53
65. Hirata S (2008) Fast electron-correlation methods for molecular crystals: an application to the alpha, beta(1), and beta(2) modifications of solid formic acid. J Chem Phys 129(20):204104. doi:10.1063/1.3021077
66. Sode O, Keceli M, Hirata S, Yagi K (2009) Coupled-cluster and many-body perturbation study of energies, structures, and phonon dispersions of solid hydrogen fluoride. Int J Quantum Chem 109:1928–1939
67. Sode O, Keceli M, Hirata S, Yagi K (2009) Coupled-cluster and many-body perturbation study of energies, structures, and phonon dispersions of solid hydrogen fluoride phonon dispersions. Int J Quantum Chem 109:1928–1939. doi:10.1002/qua
68. Manby FR, Stella M, Goodpaster JD, Miller TF (2012) A simple, exact density-functional-theory embedding scheme. J Chem Theory Comput 8(8):2564–2568. doi:10.1021/ct300544e
69. Reilly AM, Tkatchenko A (2013) Seamless and accurate modeling of organic molecular materials. J Phys Chem Lett 4:1028–1033
70. Wen S, Beran GJO (2012) Crystal polymorphism in oxalyl dihydrazide: is empirical DFT-D accurate enough? J Chem Theory Comput 8:2698–2705. doi:10.1021/ct300484h
71. Neill DPO, Allan NL, Manby FR (2010) Ab initio Monte Carlo simulations of liquid water. In: Manby F (ed) Accurate quantum chemistry in the condensed phase. CRC, Boca Raton, pp 163–193
72. Subotnik JE, Sodt A, Head-Gordon M (2008) The limits of local correlation theory: electronic delocalization and chemically smooth potential energy surfaces. J Chem Phys 128:034103
73. Nanda K, Beran GJO (2012) Improved prediction of organic molecular crystal geometries from MP2-level fragment QM/MM calculations. J Chem Phys 137:174106. doi:10.1063/1.4764063
74. Beran GJO (2009) Approximating quantum many-body intermolecular interactions in molecular clusters using classical polarizable force fields. J Chem Phys 130:164115. doi:10.1063/1.3121323
75. Ponder JW, Wu C, Ren P, Pande VS, Chodera JD, Schnieders MJ, Haque I, Mobley DL, Lambrecht DS, DiStasio RA, Head-Gordon M, Clark GNI, Johnson ME, Head-Gordon T (2010) Current status of the AMOEBA polarizable force field. J Phys Chem B 114 (8):2549–2564. doi:10.1021/jp910674d
76. Ren P, Ponder JW (2003) Polarizable atomic multipole water model for molecular mechanics simulation. J Phys Chem B 107:5933–5947
77. Sebetci A, Beran GJO (2010) Spatially homogeneous QM/MM for systems of interacting molecules with on-the-fly ab initio force-field parameterization. J Chem Theory Comput 6:155–167. doi:10.1021/ct900545v

78. Wen S, Beran GJO (2011) Accurate molecular crystal lattice energies from a fragment QM/MM approach with on-the-fly ab initio force-field parameterization. J Chem Theory Comput 7:3733–3742. doi:10.1021/ct200541h
79. Stone AJ (1981) Distributed multipole analysis, or how to describe a molecular charge distribution. Chem Phys Lett 83:233–239
80. Stone AJ (2005) Distributed multipole analysis: stability for large basis sets. J Chem Theory Comput 1:1128–1132
81. Stone AJ, Alderton M (1985) Distributed multipole analysis – methods and applications. Mol Phys 56:1047–1064
82. Misquitta AJ, Stone AJ (2008) Accurate induction energies for small organic molecules: 1. Theory. J Chem Theory Comput 4:7–18
83. Misquitta AJ, Stone AJ, Price SL (2008) Accurate induction energies for small organic molecules: 2. Development and testing of distributed polarizability models against SAPT (DFT) energies. J Chem Theory Comput 4:19–32
84. Misquitta AJ, Stone AJ (2008) Dispersion energies for small organic molecules: first row atoms. Mol Phys 106(12):1631–1643. doi:10.1080/00268970802258617
85. Stone AJ (2002) The theory of intermolecular forces. Clarendon, Oxford
86. Stone AJ, Misquitta AJ (2007) Atom-atom potentials. Int Rev Phys Chem 26:193–222
87. Neumann MA, Perrin MA (2009) Can crystal structure prediction guide experimentalists to a new polymorph of paracetamol? CrystEngComm 11(11):2475. doi:10.1039/b909819d
88. Leslie M (2008) DL MULTI – a molecular dynamics program to use distributed multipole electrostatic models to simulate the dynamics of organic crystals. Mol Phys 106 (12):1567–1578. doi:10.1080/00268970802175308
89. Grafova L, Pitonak M, Rezac J, Hobza P (2010) Comparative study of selected wave function and density functional methods for noncovalent interaction energy calculations using the extended S22 data set. J Chem Theory Comput 6(8):2365–2376. doi:10.1021/ct1002253
90. Rezac J, Riley KE, Hobza P (2011) Extensions of the S66 data set: more accurate interaction energies and angular-displaced nonequilibrium geometries. J Chem Theory Comput 7:3466–3470. doi:10.1021/ct200523a
91. Riley KE, Pitonak M, Jurecka P, Hobza P (2010) Stabilization and structure calculations for noncovalent interactions in extended molecular systems based on wave function and density functional theories. Chem Rev 110(9):5023–5063. doi:10.1021/cr1000173
92. Chalasinski G, Szczesniak MM (1988) On the connection between the supermolecular Møller–Plesset treatment of the interaction energy and the perturbation theory of intermolecular forces. Mol Phys 63:205–224
93. Cybulski SM, Chalasinski G, Moszynski R (1990) On decomposition of second-order Møller–Plesset supermolecular interaction energy and basis set effects. J Chem Phys 92:4357–4363
94. Cybulski SM, Lytle ML (2007) The origin of deficiency of the supermolecule second-order Møller–Plesset approach for evaluating interaction energies. J Chem Phys 127:141102
95. Hesselmann A (2008) Improved supermolecular second order Møller–Plesset intermolecular interaction energies using time-dependent density functional response theory. J Chem Phys 128(14):144112
96. Pitonak M, Hesselmann A (2010) Accurate intermolecular interaction energies from a combination of MP2 and TDDFT response theory. J Chem Theory Comput 6(1):168–178. doi:10.1021/ct9005882
97. Tkatchenko A, Distasio RA, Head-Gordon M, Scheffler M (2009) Dispersion-corrected Møller–Plesset second-order perturbation theory. J Chem Phys 131:094106
98. Granatier J, Pitonak M, Hobza P (2012) Accuracy of several wave function and density functional theory methods for description of noncovalent interaction of saturated and unsaturated hydrocarbon dimers. J Chem Theory Comput 8:2282–2292
99. Hesselmann A, Korona T (2011) On the accuracy of DFT-SAPT, MP2, SCSMP2, MP2C, and DFT + Disp methods for the interaction energies of endohedral complexes of the C(60) fullerene with a rare gas atom. Phys Chem Chem Phys 13(2):732–743. doi:10.1039/c0cp00968g

100. Hohenstein EG, Jaeger HM, Carrell EJ, Tschumper GS, Sherrill CD (2011) Accurate interaction energies for problematic dispersion-bound complexes: homogeneous dimers of NCCN, P2, and PCCP. J Chem Theory Comput 7(9):2842–2851. doi:10.1021/ct200374m
101. Jenness GR, Karalti O, Al-Saidi WA, Jordan KD (2011) Evaluation of theoretical approaches for describing the interaction of water with linear acenes. J Phys Chem A 115:5955–5964
102. Karalti O, Alfe D, Gillan MJ, Jordan KD (2012) Adsorption of a water molecule on the MgO (100) surface as described by cluster and slab models. Phys Chem Chem Phys 14(21):7846–7853. doi:10.1039/c2cp00015f
103. Huang Y, Shao Y, Beran GJO (2013) Accelerating MP2C dispersion corrections for dimers and molecular crystals. J Chem Phys 138:224112. doi:10.1063/1.4809981
104. Distasio RA, Head-Gordon M (2007) Optimized spin-component-scaled second-order Møller–Plesset perturbation theory for intermolecular interaction energies. Mol Phys 105:1073–1083
105. Gerenkamp M, Grimme S (2004) Spin-component scaled second-order Møller–Plesset perturbation theory for the calculation of molecular geometries and harmonic vibrational frequencies. Chem Phys Lett 392:229–235
106. Hill JG, Platts JA (2007) Spin-component scaling methods for weak and stacking interactions. J Chem Theory Comput 3:80
107. Marchetti O, Werner HJ (2009) Accurate calculations of intermolecular interaction energies using explicitly correlated coupled cluster wave functions and a dispersion-weighted MP2 method. J Phys Chem A 113:11580
108. Eshuis H, Bates JE, Furche F (2012) Electron correlation methods based on the random phase approximation. Theor Chem Acc 131(1):1084. doi:10.1007/s00214-011-1084-8
109. Li Y, Lu D, Nguyen HV, Galli G (2010) van der Waals interactions in molecular assemblies from first-principles calculations. J Phys Chem A 114:1944–1952
110. Lu D, Li Y, Rocca D, Galli G (2009) Ab initio calculation of van der Waals bonded molecular crystals. Phys Rev Lett 102:206411
111. Ren X, Tkatchenko A, Rinke P, Scheffler M (2011) Beyond the random-phase approximation for the electron correlation energy: the importance of single excitations. Phys Rev Lett 106 (15):153003. doi:10.1103/PhysRevLett.106.153003
112. Williams HL, Mas EM, Szalewicz K, Jeziorski B (1995) On the effectiveness of monomer-, dimer-, and bond-centered basis functions in calculations of intermolecular interaction energies. J Chem Phys 103(17):7374–7391. doi:10.1063/1.470309
113. Hättig C, Klopper W, Köhn A, Tew DP (2011) Explicitly correlated electrons in molecules. Chem Rev 112:4–74. doi:10.1021/cr200168z
114. Kong L, Bischoff FA, Valeev EF (2011) Explicitly correlated R12/F12 methods for electronic structure. Chem Rev 112:75–107. doi:10.1021/cr200204r
115. Beran GJO, Nanda K (2010) Predicting organic crystal lattice energies with chemical accuracy. J Phys Chem Lett 1:3480–3487. doi:10.1021/jz101383z
116. Sadlej AJ (1988) Medium-size polarized basis sets for high-level correlated calculations of molecular electronic properties. Collect Czech Chem Commun 53:1995–2016
117. Sadlej AJ (1991) Medium-size polarized basis sets for high-level correlated calculations of molecular electronic properties II. Second-row atoms Si–Cl. Theor Chim Acta 79:123–140
118. Civalleri B, Zicovich-Wilson CM, Valenzano L, Ugliengo P (2008) B3LYP augmented with an empirical dispersion term (B3LYP-D*) as applied to molecular crystals. CrystEngComm 10:405–410. doi:10.1039/b715018k
119. Wheatley PJ (1964) The crystal and molecular structure of aspirin. J Chem Soc 6036–6048 doi: 10.1039/JR9640006036
120. Ouvrard C, Price SL (2004) Toward crystal structure prediction for conformationally flexible molecules: the headaches illustrated by aspirin. Cryst Growth Des 4(6):1119–1127. doi:10.1021/cg049922u
121. Vishweshwar P, McMahon JA, Oliveira M, Peterson ML, Zaworotko MJ (2005) The predictably elusive form II of aspirin. J Am Chem Soc 127(48):16802–16803. doi:10.1021/ja056455b
122. Bond AD, Boese R, Desiraju GR (2007) On the polymorphism of aspirin. Angew Chem Int Ed 46(4):615–617. doi:10.1002/anie.200602378

123. Bond AD, Boese R, Desiraju GR (2007) On the polymorphism of aspirin: crystalline aspirin as intergrowths of two ”polymorphic” domains. Angew Chem Int Ed 46(4):618–622. doi:10.1002/anie.200603373
124. Bauer JD, Haussuhl E, Winkler B, Arbeck D, Milman V, Robertson S (2010) Elastic properties, thermal expansion, and polymorphism of acetylsalicylic acid. Cryst Growth Des 10(7):3132–3140. doi:10.1021/cg100241c
125. Bond AD, Solanko KA, Parsons S, Redder S, Boese R (2011) Single crystals of aspirin form II: crystallisation and stability. CrystEngComm 13(2):399. doi:10.1039/c0ce00588f
126. Chan EJ, Welberry TR, Heerdege AP, Goossens DJ (2010) Diffuse scattering study of aspirin forms (I) and (II). Acta Crystallogr B 66:696–707
127. Li T (2007) Understanding the polymorphism of aspirin with electronic calculations. J Pharm Sci 96(4):755–760. doi:10.1002/jps
128. Li T, Feng S (2006) Empirically augmented density functional theory for predicting lattice energies of aspirin, acetaminophen polymorphs, and ibuprofen homochiral and racemic crystals. Pharm Res 23(10):2326–2332
129. Copley RCB, Barnett SA, Karamertzanis PG, Harris KDM, Kariuki BM, Xu M, Nickels EA, Lancaster RW, Price SL (2008) Predictable disorder versus polymorphism in the rationalization of structural diversity: a multidisciplinary study of eniluracil. Cryst Growth Des 8 (9):3474–3481. doi:10.1021/cg800517h
130. Torrisi A, Leech CK, Shankland K, David WIF, Ibberson RM, Benet-Buchholz J, Boese R, Leslie M, Catlow CRA, Price SL (2008) Solid phases of cyclopentane: combined experimental and simulation study. J Phys Chem B 112(12):3746–3758. doi:10.1021/jp710017y
131. Winkel K, Hage W, Loerting T, Price SL, Mayer E (2007) Carbonic acid: from polyamorphism to polymorphism. J Am Chem Soc 129(45):13863–13871. doi:10.1021/ja073594f
132. Wen S, Beran GJO (2012) Accidental degeneracy in crystalline aspirin: new insights from high-level ab initio calculations. Cryst Growth Des 12:2169–2172. doi:10.121/cg300358n
133. Perlovich GL, Kurkov SV, Kinchin AN, Bauer-Brandl A (2004) Solvation and hydration characteristics of ibuprofen and acetylsalicylic acid. AAPS PharmSci 6(1):22–30. doi:10.1208/ps060103
134. Ahn S, Guo F, Kariuki BM, Harris KDM (2006) Abundant polymorphism in a system with multiple hydrogen-bonding opportunities: oxalyl dihydrazide. J Am Chem Soc 128 (26):8441–8452. doi:10.1021/ja0573155
135. Karamertzanis PG, Day GM, Welch GWA, Kendrick J, Leusen FJJ, Neumann MA, Price SL (2008) Modeling the interplay of inter- and intramolecular hydrogen bonding in conformational polymorphs. J Chem Phys 128(24):244708. doi:10.1063/1.2937446
136. Perrin MA, Neumann MA, Elmaleh H, Zaske L (2009) Crystal structure determination of the elusive paracetamol form III. Chem Commun 22:3181–3183. doi:10.1039/b822882e
137. van de Streek J, Neumann MA (2010) Validation of experimental molecular crystal structures with dispersion-corrected density functional theory calculations. Acta Crystallogr B 66 (Pt 5):544–558. doi:10.1107/S0108768110031873
138. van de Streek J, Neumann MA (2011) Crystal-structure prediction of pyridine with four independent molecules. CrystEngComm 13(23):7135. doi:10.1039/c1ce05881a
139. Jurečka P, Šponer J, Černỳ J, Hobza P (2006) Benchmark database of accurate (MP2 and CCSD(T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs. Phys Chem Chem Phys 8:1985–1993
140. Chisholm JA, Motherwell WDS (2005) COMPACK: a program for identifying crystal structure similarity using distances. J Appl Crystall 38(1):228–231. doi:10.1107/S0021889804027074

Top Curr Chem (2014) 345: 95–138
DOI: 10.1007/128_2013_526
© Springer-Verlag Berlin Heidelberg 2014
Published online: 28 February 2014

Prediction and Theoretical Characterization of *p*-Type Organic Semiconductor Crystals for Field-Effect Transistor Applications

Şule Atahan-Evrenk and Alán Aspuru-Guzik

Abstract The theoretical prediction and characterization of the solid-state structure of organic semiconductors has tremendous potential for the discovery of new high performance materials. To date, the theoretical analysis mostly relied on the availability of crystal structures obtained through X-ray diffraction. However, the theoretical prediction of the crystal structures of organic semiconductor molecules remains a challenge. This review highlights some of the recent advances in the determination of structure–property relationships of the known organic semiconductor single-crystals and summarizes a few available studies on the prediction of the crystal structures of *p*-type organic semiconductors for transistor applications.

Keywords Charge transfer integral · Charge transport · Crystal structure · Crystal structure prediction · Mobility · Organic field-effect transistors · Organic semiconductors

Contents

Ş. Atahan-Evrenk (✉)
Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, MA 02138, USA

TOBB-ETU Medical School, Sogutozu Cad. No: 43, Sogutozu, Ankara 06560, Turkey
e-mail: atahan@fas.harvard.edu

A. Aspuru-Guzik
Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, MA 02138, USA

Abbreviations

BTBT	[1]Benzothieno[3,2-*b*][1]benzothiophene
BTBT-C_8	2,7-Dioctyl[1]benzothieno[3,2-*b*][1]benzothiophene
DATT	Dianthra[2,3-*b*:2′,3′-*f*]thieno[3,2-*b*]thiophene
DMA	Distributed multipole analysis
DNTT	Dinaphtha[2,3-*b*:2′,3′-*f*]thieno[3,2-*b*]thiophene
DPP	Diketo-pyrrolo-pyrrole
DPP(TBFu)$_2$	3,6-Bis(5-(benzofuran-2-yl)thiophen-2-yl)-2,5-bis(2-ethylhexyl)pyrrolo[3,4-*c*]pyrrole-1,4-dione
DTT-Ph-C(8,12)	2,6-Bis(4-{octyl,dodecyl}phenyl)-dithieno[3,2-*b*′:2′,3′-*d*]thiophene
GA	Genetic algorithms
ISC	Inorganic semiconductor
OSC	Organic semiconductor
PDIF-CN_2	*N*,*N*′-1*H*,1*H*-Perfluorobutyldicyanoperylene-carboxydi-imide
Rubrene	5,6,11,12-Tetraphenyltetracene
TbTH	Tetraceno[2,3-*b*]thiophene
TcTH	Tetraceno[2,3-*c*]thiophene
Tips-pentacene	6,13-Bis(triisopropylsilylethynyl)pentacene
TMTSF	Tetramethyltetraselenafulvalene

1 Introduction

Organic compounds with π-electron systems show interesting functionality such as conductivity and semiconductivity [1]. Thanks to their extended π-conjugation, organic semiconductors (OSCs) have delocalized orbitals where charge carriers can move. If the charge carriers are produced extrinsically by injection from electrodes or by the photoelectric effect, electrons or holes can find pathways to go from molecule to molecule or over a conjugated backbone of a polymer (Fig. 1a, b) Thus OSCs have great potential to be used as components in electronics or in solar cells. Especially in inexpensive, flexible, and large area applications such as radio frequency ID tags, chemical/pressure sensors, display drivers, and solar cells [8–11], they have the potential to be an alternative to silicon-based semiconductors.

A notable advantage of OSCs is their design potential. The versatility of carbon provides a vast chemical space which can be explored with in silico strategies. For example, in the Harvard Clean Energy project database, there are 2.6 million theoretical candidate OSC compounds combinatorially designed from 30 molecular fragments as potential semiconductor materials for solar cell applications

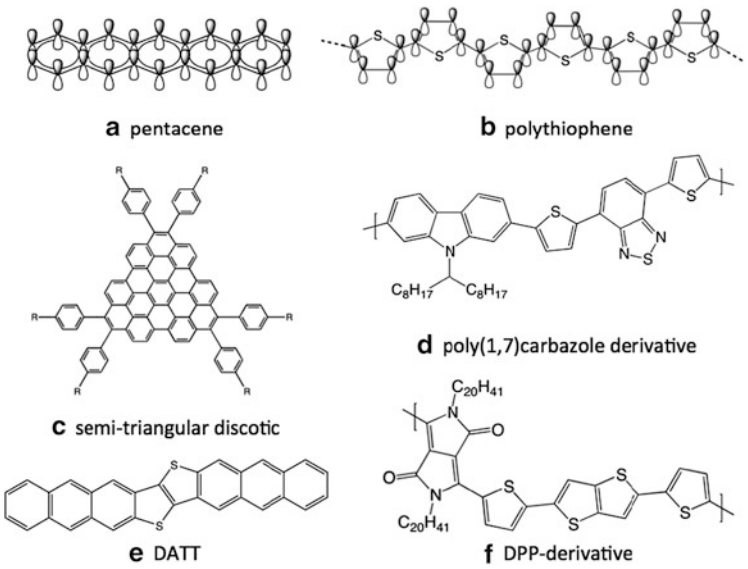


Fig. 1 Examples of molecular and polymer organic semiconductors. The references for the molecules: (**a**) [2], (**b**) [3], (**c**) [4], (**d**) [5], (**e**) [6], (**f**) [7]

[12–14]. In this vast compound space, theoretical modeling and prediction have the potential to guide the synthesis of new OSCs thus reducing the high cost of finding materials by trial and error.

In recent years, increasing numbers of new OSCs have been designed and improved through computational modeling. For example, the modification of liquid crystal molecules (Fig. 1c) to enhance columnar organization and optimize intermolecular couplings for enhanced charge transport [4] or the design of polycarbazole derivative donor polymers (Fig. 1d) for organic photovoltaic applications by first-principles screening of the prototypical oligomers [5] proved useful. The synthesis of a new high performance molecular material [6] (Fig. 1e) or diketopyrrolopyrrole (DPP)-based polymers (Fig. 1f) with the help of computational modeling are demonstrated [7]. Despite the common limitations of current theoretical studies based on approximate methods such as density functional theory (DFT), these studies, where theory guided the synthesis of high performance materials, are promising. Now we are just opening a door into a new era where quantum chemistry methods are integrated into material science and engineering problems for more intelligent search of better OSC materials [12, 13, 15–19].

The successful prediction and computational design of high performance OSCs will be possible if three main objectives are achieved. First, a detailed understanding of the structure–property relationship for known OSCs is crucial. The main goal is to know what works, why it works, and how the material properties could be further optimized. Second is the ability to predict the solid-state structure from the molecular structure. This requires the generation of all possible crystal structures

and computation of the lattice energies to rank and identify the lowest energy polymorphs. Third is the efficient computational simulation of charge transport to evaluate the crystal structures and identify the best semiconductors for particular applications. In this review we address the first and second objectives and briefly discuss the third.

With an ever-growing number of publications in OSC synthesis, design, and analysis [20, 21], there is now a substantial body of knowledge on the structure–property relationships for the OSCs in transistor applications. Undoubtedly, X-ray diffraction and analysis plays a crucial role in the elucidation of the structures and the derivation of the structure–property relationships experimentally [22]. In addition, theoretical characterization with molecular dynamics, quantum chemistry, especially density functional theory, and charge transport models, helped to derive important structure–property relationships in OSCs [16, 17, 23–25]. We will discuss some of the derived relationships in Sect. 2.

The crystal structure prediction for OSCs, on the other hand, remains mostly unexploited, not only because of the computational challenges but also, we believe, due to the lack of commercial applications that requires immediate attention to possible polymorphs and the structural factors affecting the device performance. One of the most important motivations for the crystal structure prediction for organic molecular solids is the polymorphism in solid forms of drug molecules [26]. For example, an unpredicted lower energy polymorph of the HIV drug Ritonavir [27], which manifested itself during the manufacturing and storage, not only was inactive as a drug but could also act as a seed to convert the active form into the inactive form on contact. In OSCs, interconversion between polymorphs may lead to loss of the desired electrical properties. Thus, we believe in the very near future, as more commercial applications of OSCs appear in the electronics market, the crystal structure prediction of OSCs will gain impetus.

There are various factors that complicate crystal structure prediction for OSCs. As shown in Fig. 1, most OSC molecules are large and flat with quasi-rigid conjugated backbones. Like other organic molecular crystals, they have many polymorphs with similar lattice energies. For example, there is less than 1 kcal/mol difference between C and H polymorphs of pentacene [28]. On one hand, it is advantageous to have a relatively simple molecular structure with higher symmetry that leads to mostly two-dimensional packing. This simplifies the search problem, that is the effective construction of likely packing configurations. On the other hand, phase transitions from one polymorph to another are more likely in these two-dimensional packing patterns. The energy barriers among different polymorphs are not as high as in organic crystals which pack in three-dimensional patterns, where the molecules are locked in place due to a lock and key kind of close-contacts. Therefore, in OSCs, as temperature fluctuates, transitions from one polymorph to another are more likely, sometimes yielding mobility differences as high as an order of magnitude as temperature increases [29]. In additon, solution processible materials usually have long alkyl chains as solubilizing groups [30], with many conformers, increasing the size of the search space dramatically.

With the exception of a few hydrogen bonding OSCs [31, 32], most OSC solids are held together by van der Waals forces, mostly London dispersion forces. These forces are quantum mechanical in nature and hard to capture in classical models such as molecular mechanics, which always treat the van der Waals forces as two body interactions. The transferability of the force field parameters is also an issue [33]. Due to the size of the crystalline systems, the conventional high-accuracy correlated wave function methods are too expensive. Therefore, density functional theory (DFT) methods are widely used. However, conventional DFT approximations cannot describe the attractive van der Waals forces (long-range dispersion) accurately [34]. For example, to rank the lattice energies of polymorphs of a relatively small molecular system such as para-diiodobenzene, DFT falls short, and computationally demanding methods such as diffusion quantum Monte Carlo is necessary [35]. Fortunately, the dispersion corrected DFT methods have now developed to such an extent that many body dispersion interactions can be included in the interaction potentials and can be used to calculate the lattice energies with chemical accuracy [36]. For example, many-body dispersion-corrected DFT has been shown to achieve chemical accuracy for the prediction of the sublimation enthalpies for a set of compounds ranging from pure hydrogen bonding molecules to only van der Waals bonded solids as well as compounds that uses both types of bonding in the solid state [36]. This is an exciting development yet it needs to be put to work for OSC crystal prediction. The practical use of DFT with dispersion to predict small molecular crystals has also been vetted in the last crystal structure blind test organized by Cambridge Structural Database [37].

Blind tests organized by Cambridge Structural Database provides an invalueable opportunity for the test of the predictive power of the state-of-the-art methodologies for crystal structure prediction. In the last blind tests, significant progress towards the prediction of crystal structures of molecular crystals was achieved [37, 38]. The most successful approaches included extensive initial structure searches to span all possible space groups, primary refinement of the initial structures by the tailor-made force fields [39], treatment of the internal degrees of freedom for the flexible molecules correctly [40], and refinement of the lattice energies with dispersion corrected DFT methods [41] or distributed multipole analysis approaches [42]. In the last blind test, from the perspective of the organic semiconductor prediction, especially important was the successful prediction of a compound with large internal flexibility. Two participating groups predicted the crystal structure of the largest (33 heavy atoms) and most flexible (8 rotatable bonds) molecule and reported the right polymorph as their top ranking choice. The positive implication of this success for the prediction of the crystal structures of OSCs is obvious as most of them are large and include solubilizing alkyl chains with many rotatable bonds.

Another challenge in the computational OSC design is the difficulty of testing the theories and structure predictions by comparison with experimental data. The main reason is the multitude of experimental factors affecting the solid-state structures and ensuing device performance. The nanoscale order, for example, is highly dependent on processing factors such as temperature, pressure, type of solvent, impurities, or substrate surface [15]. Small changes in these factors have

the potential to affect the experimental outcome drastically. The inclusion of all of these factors into the computer simulations is an important unsolved challenge for the theoretical community. We should also note that theoretical studies usually focus on single crystals, leaving out the polycrystalline materials which unfortunately constitute most of the reported compounds in the literature.

The charge transport processes in OSCs is another hard problem to model. To achieve de novo modeling, one needs general models where different regimes of transport are addressed accurately and efficiently so that libraries of molecules and their solid-state structures can be analyzed to identify the most promising ones. However, in principle, the movement of charge carriers in OSCs requires a many-body quantum mechanical treatment [43]. Since an exact solution is not possible, a number of approximate approaches are adapted such as the density matrix propagation, master equation approaches, dynamical mean-field theory, and so forth. The studies showed that the coupling of the motion of the charge carrier to the molecular and crystal degrees of freedom is crucial to describe the problem accurately and a self-consistent solution of the whole Hamiltonian is necessary [25]. The modeling of charge transport in OSCs is still a developing field and the structural parameters obtained from ab initio quantum chemistry are playing an important role in the development of more realistic models [23–25, 44, 45]. In addition, a comprehensive multi-scale approach also needs to address issues emerging from organic field-effect transistor (OFET) device configurations such as the contact resistance, and short and long-range effects of the gate dielectric, among others.

Thus, in this rapidly advancing field of OSC design, the theoretical characterization and prediction has tremendous potential for growth. Successful routine prediction and screening approaches may make themselves more useful in the future. Before moving on to the theoretical characterization and prediction techniques, in the rest of the introduction we briefly summarize the working principles of an OFET and the important structural parameters of charge transport for an ideal OSC for transistor applications.

1.1 The Organic Field-Effect Transistor

The transistor is the most important single element in an electronic circuit. For organic electronics to find viable commercial applications, a fast, reliable, and inexpensive OFET is crucial [46]. We have come a long way since the first organic transistors in the 1980s [3, 47]. Especially in the last 10 years there has been a great progress in the discovery of new OSCs and the optimization of the process conditions to achieve commercial viability [3, 21]. Today, some organic semiconductor materials surpass amorphous silicon in performance [48] and there are commercial applications of organic transistors in printed electronics [49], displays, and microelectronics [9].

An OFET is a voltage-controlled switch where an external voltage applied between the electrodes creates a current through an OSC layer. Then the flow of

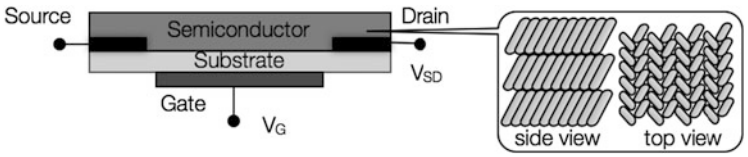


Fig. 2 Schematic of an OFET structure (*left*) and example of herringbone-type molecular packing in the OSC (*right*)

current is modulated by a gate voltage applied through the dielectric substrate [50, 51]. In a typical bottom-contact *p*-type OFET, an OSC crystal or a thin film is placed over a dielectric substrate (usually SiO_2) in contact with the source and the drain electrodes (Fig. 2). The negative bias gate voltage applied through the dielectric induces charge carriers (holes), creating a channel between the source and the drain electrodes. Current flows through the channel when a drain bias is applied.

In *p*-type materials the holes are the charge carriers. Therefore charge injection into the highest occupied molecular orbital (HOMO) of the OSC requires that the work function of the metal electrodes should match the HOMO energy level of the OSC. The source and drain electrodes are usually chosen from low work function noble metals such as gold (Au work function: 5.1 eV). The dielectric layer separating the gate from the semiconductor could be a Si/SiO_2 layer preferentially treated with a hyrophobic self-assembled monolayer of alkylsilanes to promote molecular order and enhance transport [52].

The most important parameters defining the transistor performance are the mobility and drain-source current ratio when the gate voltage switches on and off, $I_{on/off}$ [50]. The mobility characterizes the mean drift velocity of the charge carrier when an electric field is applied and it is derived from the current–voltage curves of the transistor. The higher the mobility and $I_{on/off}$, the better the transistor performance. For example, in a liquid display, $I_{on/off}$ of 10^6 and a minimum mobility of 0.1 $cm^2\ V^{-1}\ s^{-1}$ are needed for good performance [50]. Another important factor which affects the performance is the operating voltage of the transistor. Traditionally, the operating voltages in OFETs could be as high as 100 V. Low operating voltages is a requirement for practical applications and has been demonstrated by the use of high capacitance dielectric layers [53, 54].

OFET devices are very complex since many factors affect their performance, such as the grain size and boundaries, substrate, substrate temperature, or contact resistance. In this review we only focus on the molecular and crystal structural factors affecting the charge transport ability of the OSC layer. Specifically, we consider the single-crystal layers because they provide a more consistent platform to study the structural effects [55].

Since charge carrier transport depends strongly on the π-orbital interactions, the molecular level ordering has important consequences for OSCs. Figure 3 schematically illustrates the correlation of the mobility to molecular level ordering for a set of materials from perfect single-crystals to disordered polymers. It is established

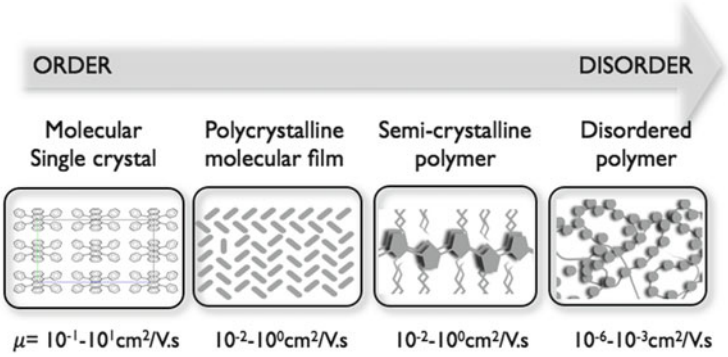


Fig. 3 OFET mobility correlates with structural order. As the order of the material decreases, in a progression from single molecular crystals to disordered polymers, carrier transport decreases by orders of magnitude

now that usually the better the molecular organization of an OSC, the higher the material performance. Surpassing amorphous silicon, the single crystalline OSCs have mobilities in the range of 0.1–40 $cm^2\ V^{-1}\ s^{-1}$ whereas in polycrystalline and disordered films the mobilities fall down to the range of 0.0001–0.1 $cm^2\ V^{-1}\ s^{-1}$ [46]. In polycrystalline films the grain boundaries become traps for charge carriers, thus reducing the performance considerably. The anisotropy in the mobilities due to the low symmetry of the molecules compared to inorganic semiconductors, for example, also underlines the importance of the packing patterns and microscopic order in OSCs.

Despite the success of crystalline small molecule OFETs, the commercialization of these devices has been limited due to their low solubility. Usually, to make single-crystal OFETs, micron-to-millimeter-size crystals are obtained by vacuum deposition and later handpicked to be aligned along the appropriate direction between the electrodes. This is not a yet practical option for commercialization. The large-scale semiconductor coating production similar to printing requires solution processable materials. To achieve solubility, molecular units are usually modified through the addition of aliphatic groups such as long alkyl chains. Although these additions/modifications sometimes increase the performance due the close-packing of the conjugated backbones enforced by the lipophilic interactions of the alkyl chains [31], the grain boundaries reduce the mobility to the range of 0.01–1 $cm^2\ V^{-1}\ s^{-1}$.

Recently, however, there has been important progress regarding the development of solution-processable and high-performance OSCs. For example, with controlled deposition of tips-pentacene films (Fig. 4a) with solution-shearing through nanoarray surfaces, a mobility of 11 $cm^2\ V^{-1}\ s^{-1}$ has been demonstrated by Diao et al. [59]. Also in microsheets and microribbons of dithieno[3,2-*b*:2′,3′-*d*] thiophene derivatives (Fig. 4b) a mobility of 10 $cm^2\ V^{-1}\ s^{-1}$ and above is observed [57]. Moreover, a record average high mobility of 16.4 $cm^2\ V^{-1}\ s^{-1}$ was observed in the single crystals of C_8-BTBT (Fig. 4c) deposited in an ink-jet printer [156].

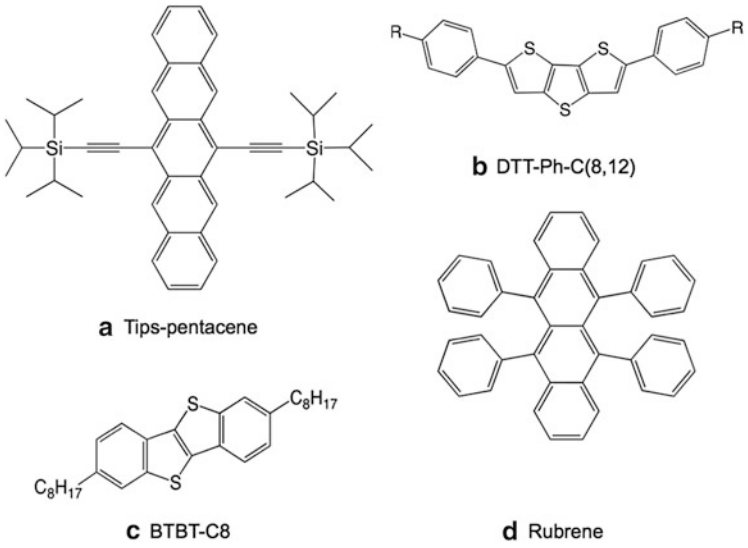


Fig. 4 Examples of high-performance OSCs. The references for the molecules: (**a**) [56], (**b**) [57], (**c**) [157], (**d**) [58]

Semiconducting polymers, on the other hand, naturally have the advantages of solution processability, thermal stability, and film uniformity [60]. The polymers usually form semi-crystalline films with lower mobilities as seen in Fig. 3. Compared to the first polymer-based OFET (polythiophene, Fig. 1b) fabricated in 1986 [3], the progress has been remarkable. For example, DPP-based polymers (Fig. 1f) with a record high mobility of 10 $cm^2\ V^{-1}\ s^{-1}$ were reported recently [61]. Such high mobility for a semicrystalline polymer is very unusual due to the presence of the disorder in these systems. Recent investigations revealed that a certain degree of polymerization is necessary for the high mobility since the polymeric chains connecting crystalline domains can act as bridges for charge transport [62]. Since the structure prediction for polymers is outside the scope of this review, we will not discuss polymer-based OFETs any further. However, we would like to underline the notion that once a good molecular or polymeric material is discovered, through the optimization of the processing conditions it is possible to engineer high-performance materials. This might entail the enhancement of the intermolecular orbital interactions or the reduction of the impurities or other structural traps.

Despite the remarkable progress made in the synthesis and process engineering in OFETs [15], up until now theoretical characterization and prediction has relied on the availability of crystal structures from X-ray diffraction analysis [22]. Once the structure of an OSC is known, the structural parameters affecting the performance are examined through theoretical characterization techniques. The structure–property relationships learned from the study of known OSCs provide a framework upon which new novel materials discovery routes rely. We dedicate the

rest of this section to the important parameters in OFET modeling and Sect. 2 to the structure–property relationships.

1.2 Charge Transport Models and Parameters

In contrast to the success of charge transport models for inorganic semiconductors (ISC), models for OSCs are still under development. The major reason for this lack of a comprehensive model is the difficulty of dealing with molecular behavior in the solid state. Unlike in an ISC, the OSC building blocks have high polarizability, low internal symmetry, and weak intermolecular interactions. In addition, the diversity of materials from molecules to polymers and to liquid crystals poses challenges for the development of generalized and consistent models.

To summarize the basic understanding of the charge transport mechanisms and underlying structural factors, it is useful to start with a Hamiltonian for a single particle moving on a periodic lattice. The following Hamiltonian [63] includes all the important energy contributions describing the motion of a charge carrier over regular lattice sites, i and j:

$$\begin{aligned}
\mathrm{H} &= \mathrm{H}^0_{\mathrm{el}} + \mathrm{H}^0_{\mathrm{ph}} + \mathrm{H}^{\mathrm{local}}_{\mathrm{el-ph}} + \mathrm{H}^{\mathrm{non\text{-}local}}_{\mathrm{el-ph}}, \\
\mathrm{H}^0_{\mathrm{el}} &= \sum_j \epsilon_j a_j^+ + \sum_{i \neq j} J_{ij} a_i^+ a_j, \\
\mathrm{H}^0_{\mathrm{ph}} &= \sum_{\mathbf{q}} \hbar\omega_{\mathbf{q}} \left(b_{\mathbf{q}}^+ b_{\mathbf{q}} + \frac{1}{2} \right), \\
\mathrm{H}^{\mathrm{local}}_{\mathrm{el-ph}} &= \sum_{\mathbf{q}} \sum_j \hbar\omega_{\mathbf{q}} g_{jj,\mathbf{q}} \left(b_{\mathbf{q}}^+ b_{-\mathbf{q}} \right) a_j^+ a_j, \\
\mathrm{H}^{\mathrm{non\text{-}local}}_{\mathrm{el-ph}} &= \sum_{\mathbf{q}} \sum_{i \neq j} \hbar\omega_{\mathbf{q}} g_{ij,\mathbf{q}} \left(b_{\mathbf{q}}^+ + b_{-\mathbf{q}} \right) a_i^+ a_j.
\end{aligned} \tag{1}$$

Here, $\mathrm{H}^0_{\mathrm{el}}$ is the electronic Hamiltonian, a^+ and a are the creation and annihilation operators, i and j label the molecular sites with energy ϵ, and J is the charge transfer integral which represents the strength of electronic coupling among neighboring sites. $\mathrm{H}^0_{\mathrm{ph}}$ represents the phonon contributions with frequency ω where $\mathbf{q}$ and is the wavevector. The operators $b^+_{\mathbf{q}}$ and $b_{\mathbf{q}}$ denote the creation and annihilation operators for the phonons with energy $\hbar\omega_{\mathbf{q}}$. As can be seen in (1), the electron–phonon coupling term has local (g_{ii}) and non-local (g_{ij}) contributions. Within this formulation, the local electron–phonon coupling term modulates the site energies whereas the non-local coupling modulates the site-to-site interaction terms J_{ij}. As a first-order approximation, the local and the non-local electron–phonon coupling terms could be obtained by Taylor expansion of the electronic Hamiltonian along the phonon modes [64].

Depending on the relative strength of the terms of this Hamiltonian, the models could be roughly categorized into band, polaron, and disorder models [63]. In the band transport models, J is the most important term. Due to strong coupling among the sites, the bandwidth is relatively large, and thus a mobility value above 10 $cm^2\ V^{-1}\ s^{-1}$ is possible. However, in OSCs, since the molecular interactions are weak, injection of a charge polarizes the molecule and its surroundings significantly, leading to formation of a polaron. The polaron is a quasi-particle representing the charge and the polarized lattice around it [43]. Thus in these models the electron–phonon coupling terms play a crucial role and determine the transport characteristics and the temperature dependence in the mobility. In the case of disorder-based models, the fluctuations of the site energies and the modulation of the coupling terms are so large that the charge transfer can be described as a series of uncorrelated hops in a broad density of states [63].

Temperature plays a crucial role in the charge transport in OSCs. It has been demonstrated for pentacene (Fig. 1a) [65] and rubrene (Fig. 4d) [58], for example, that, at low temperatures up to 280 K, the mobility shows an inverse dependence on temperature (T^{-n}) which is the evidence of band transport. The mobility decreases with increasing temperature as more and more scattering of the charge carriers happens due to the phonon modes of the medium. Somewhere around room temperature (280–300 K), the charges carriers localize to the molecular sites due to thermal fluctuations and the transition to activated hopping occurs so that the mobility increases as a function of increasing temperature [66].

Thus, naturally, the choice of the most appropriate model depends on the knowledge of the molecular and solid-state structure. Once the structure is known, ab initio molecular [25, 67] or solid-state electronic structure parameters [68] can be used to make more realistic models based on multi-scale QM/MM or semiclassical schemes [23, 24, 44, 69]. In the following sections we briefly discuss these important parameters for the charge transport models.

1.2.1 Electronic Coupling, *J*

The electronic coupling term, J, is one of the important structural parameters determining the charge transport properties of an OCS. In the literature, another more common name for the electronic coupling term is the charge transfer integral, or transfer integral.

For *p*-type OSCs, in a simple dimer approach, the transfer integral can be calculated as the electronic coupling of the HOMO orbitals of adjacent molecules as $J = \langle \psi_A^{\mathrm{HOMO}} | \mathrm{H_{el}} | \psi_B^{\mathrm{HOMO}} \rangle$, where the A and B indexes represent the adjacent molecules in a relative geometry extracted from the crystal structure. For symmetrically equivalent molecules, it is simply the splitting of the HOMO molecular orbitals in the dimer configuration. In the case of geometrically non-equivalent molecules, however, the effect of the site energy differences should be taken into account [70–72]. This dimer-based approximation neglects the effect of the crystal

environment on the transfer integrals. It is also possible to calculate the transfer integrals from fitting the dispersion of the bands into a tight-binding Hamiltonian [44, 73, 74]. This approach includes the crystal environment effects within the level of the theory used. Experimentally, the transfer integral values can be obtained from the dispersion of the HOMO bands determined by photoemission spectroscopy [68].

In the calculation of transfer integrals, DFT-based methods are usually employed as they provide a good enough accuracy with a reasonable cost. For a medium size molecule such as pentacene, hybrid functionals with double-zeta basis with polarization functions are usually used and shown to provide qualitatively accurate results [75]. For larger molecules, long-range corrected DFT methods provide better accuracy [76]. It has also been shown within the dimer-based approach that the change in the transfer integrals associated with the applied electric field is negligible [77].

The electronic coupling terms show the following characteristic features:

1. Exponentially decreasing as a function of increasing intermolecular distance. The smaller the intermolecular stacking distances, the stronger the molecular orbital couplings for configurations for which the transfer integral is not zero.
2. Oscillatory behavior as a function of molecular displacements in the direction perpendicular to the interaction of π-orbitals. The oscillation as a function of the slip along the long axis is called the *D*-modulation.
3. Dihedral angle dependence [78].
4. Anisotropy [79].

As can be seen in Fig. 5a–c, the transfer integral strength strongly depends on the dimer geometry of the interacting molecules [25, 64]. The dependence on the π-stacking distance is exponential (Fig. 5a) [25], and thus it is crucial to be able to control it. Slip-stacked geometries as a function of short and long axes show oscillations. Kojima and Mori analyzed the dihedral angle dependence of transfer integrals for a group of molecular OSCs with typical herringbone packing [78]. As expected, all these molecules show the characteristic oscillation behavior as a function of dihedral angle and *D*-modulation. They observed that the transfer integral strength is strongly dependent on the molecular orbital symmetry as well as the crystal structure packing. Due to the oscillatory nature of the transfer integrals as a function of geometrical orientation of the molecules, achieving charge transport through the optimization of the couplings is challenging.

The strength and the extent of *J* determine (an)isotropy of the mobility and the bandwidth. As shown in Fig. 5d for pentacene, moderate to strong anisotropy of mobility tensor is a common feature of the transport plane. The bandwidth is also strongly dependent on the transfer integral values. For example, within the tight-binding approximation, the bandwidth for a one-dimensional network of coupled sites is defined as 4*J* [80, 81]. For a two-dimensional OSC, direct relationships between the transfer integrals and the bandwidth are established [82, 83]. The percolation network's robustness, fragility, and dimensionality depend on the strength of the transfer integrals between the molecules in each direction.

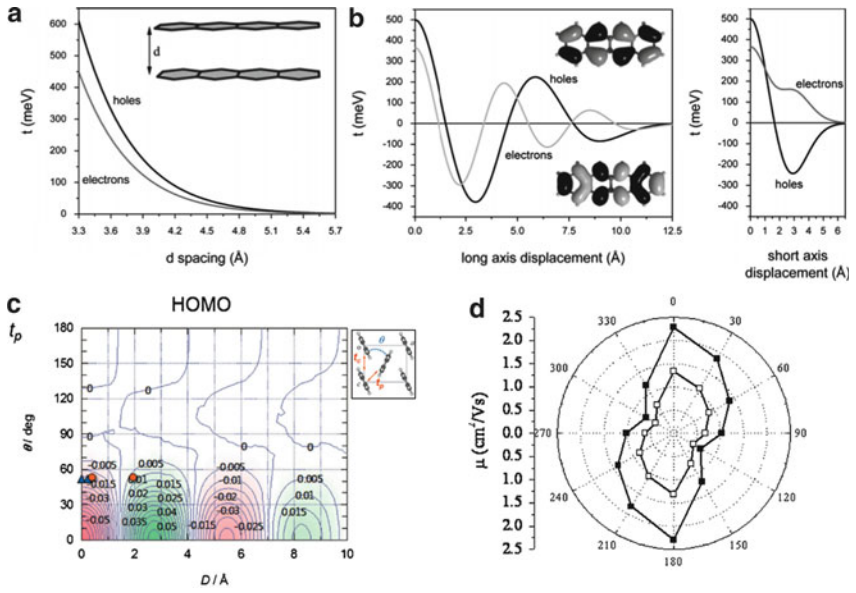


Fig. 5 (**a**, **b**) The dependence of the transfer integral on the intermolecular π-stacking distance in tetracene [25], *D*-modulation in tetracene [25]. (Reprinted with permission from [25]. Copyright (2007) American Chemical Society). (**c**) Dihedral angle dependence in pentacene [78]. (Reprinted with permission from [78]. Copyright (2011) The Chemical Society of Japan). (**d**) Anisotropy of the mobility of pentacene in the transport plane (*filled squares*: maximum mobility, *empty squares*: −10 V gate voltage). (Reprinted with permission from [79]. Copyright (2006) AIP Publishing LLC)

Depending on the packing patterns and the ensuing π-orbital overlap, one-, two-, or sometimes even three-dimensional conductance is possible. For example, Vehoff et al. analyzed the topological connectivity of four different single crystals, namely rubrene, benzo[1,2-*b*:4,5-*b*′]bis[*b*]benzothiophene derivatives with and without C_4H_9 side chains and indolo[2,3-*b*]carbazole with CH_3 side chains [84] (see Fig. 6). The transfer integrals for the adjacent molecules revealed the dimensionality of the percolation networks as illustrated in Fig. 6.

These observations pertaining to the nature of the transfer integral only include the geometries of the equilibrium structures. At finite temperature, since these materials are weakly bound van der Waals solids, the thermal motion of the molecules significantly affect (modulate) the transfer integrals. Instead of a single equilibrium value, a probability distribution of transfer integral values appears. The shape of the distribution strongly depends on the temperature. This modulation could naturally be described as the coupling of phonon modes of the crystal to the electronic degrees of freedom. There is indeed a great deal of work in the literature dedicated to this subject and we briefly discuss some important aspects of it in the next subsection.

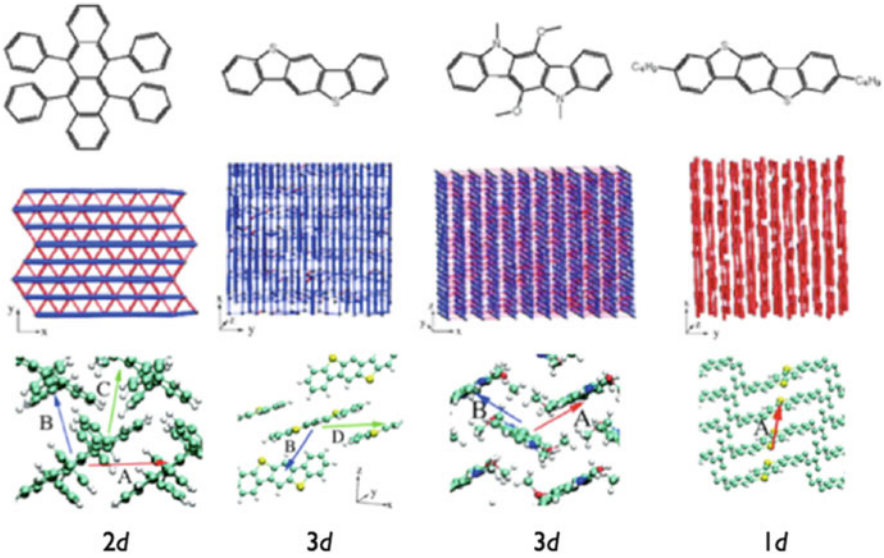


Fig. 6 The percolation networks and the dimensionality of the charge transport determined from the transfer integral calculations for the molecules shown. The centers are the center of mass of the each molecular site. The connectivity and colors refer to the strength and type of the transfer integrals, as indicated by the *colored arrows*. (Reprinted (adapted) with permission from [84]. Copyright (2010) American Chemical Society)

1.2.2 Electron–Phonon Coupling

Following (1), the electron–phonon coupling can be described as a sum of local and non-local electron–phonon coupling terms. This division comes naturally in OSCs as the intermolecular vibrational forces are much weaker compared to the intramolecular ones.

The electron–phonon coupling is usually calculated within the harmonic approximation. A charge carrier over a site changes the potential energy surface for the site such that rearrangement of the nuclei now have a different energy shifted from the original point of the potential well by an amount of $g_{jj}^2 \hbar\omega_{\mathbf{q}}$ [23]. This shift in energy is called the polaron binding energy in polaron transport models [25] and the reorganization energy in the terminology of the Marcus electron-transfer theory [85].

The intramolecular portion of the electron–phonon coupling is calculated approximately from the gas-phase geometries of the neutral and charged molecules by measuring the change in the total energy upon charging [25, 86]. This is commonly called the internal reorganization energy. This approach neglects the coupling of the intramolecular modes to the polarization of the neighbour molecules in the crystal environment. The aromatic conjugated molecules are highly polarizable; hence the lattice will distort to accommodate the charges [43]. Nevertheless, the external reorganization energy contributions are found to be small

[87, 88], providing some justification for their omission. In addition, the experimentally determined reorganization energy values for rubrene, pentacene, and perfluropentacene are comparable with this local coupling picture [89, 90].

The calculation of the reorganization energy for molecules with rotatable bonds is more complicated than the fused aromatic ones. In molecules with rotatable bonds, geometry of a charged state could be very different from the neutral one. Then the simple harmonic approximation fails to capture the dependency of the total reorganization energy on the anharmonic vibrational modes that are related to the torsional motion of the backbone [91]. Moreover, there is a chance that the gas phase calculations overestimate the reorganization energy of OSCs with rotatable bonds. Although in gas phase calculations the torsional degree of freedom is controlled only by the intramolecular interactions, in the solid state the torsional motion can be hindered due to intermolecular van der Waals interactions [25]. To compensate, sometimes the symmetry constraints of the molecule in the crystal environment are imposed for the gas phase reorganization energy calculations. However, there is no guarantee that this approach will provide an accurate picture in all cases.

The non-local electron–phonon coupling constant, g_{ij}, is more challenging to calculate as it involves the modification of the electronic couplings due to phonon modes of the crystal. Usually, finite temperature dynamics of the lattice is obtained from classical trajectories and the effect of phonon modes of the crystal on the electronic couplings is observed a posteriori by extracting unique dimers from an MD trajectory. Subsequently, the transfer integral calculations based on the dimer approach mentioned earlier are performed along the trajectory [92–95]. This approach provides insight into the thermal modulation of the transfer integrals by providing a probability distribution. How large is the standard derivation compared to the mean value and how strongly does it depends on the temperature? These questions can be answered by the use of this semiclassical approach. For example, for pentacene single-crystals at 300 K, the transfer integrals for three unique dimers have been studied with data from MD trajectories. The standard deviations were found to be of the same order of magnitude as the average transfer integrals [94].

The nonlocal couplings can also be studied by the calculation of the numerical derivatives of the transfer integrals with respect to the distortions of the crystal lattice [96, 97]. These couplings represent the zero Kelvin behaviour of the system. Subsequently, they can be extrapolated to higher temperatures with the use of classical or quantum statistics. With this approach, for pentacene slightly smaller standard deviations of the couplings were observed. However, they were still of the same order of magnitude as that of the mean values. Thus in the larger oligoacenes such as pentacene, it is safe to assume that the non-local electron–phonon coupling are significant and semiclassical models of charge transfer could be of use. We should note, however, that this type of analysis should be extended to other molecular systems, as different behaviour could be observed for systems with comparatively stronger or weaker intermolecular interactions.

2 Structure–Property Relationships

The structure–property relationships in OSC design can be characterized at the molecular and crystal scales. As mentioned previously, a molecular OSC is usually made of molecules which are mostly interacting by van der Waals forces [1]. Thus the formation of an organic solid preserves the molecular properties to a large extent with some perturbation due to the intermolecular interactions. Since the molecule in the solid state has many important characteristics resembling its gas phase properties, the choice of the right molecule is crucial. The molecular properties such as planarity, rigidity, conjugation length, size, symmetry, side groups, and their positions, together with chemical structure determine the electronic structure parameters as well as the packing in the solid state.

On the other hand, although the molecular properties are preserved to a certain degree in a molecular crystal, many new properties emerge. The experimental evidence shows that electronic, optical, or transport characteristics depend highly on the molecular packing [43], such as Davydov splitting [98], line broadening, and bathochromic shifts [99]. From the perspective of the OFET application, the most important feature is the semiconductivity. Since the interactions among the molecules are weak and low internal symmetry of the molecules dictates that only certain packing patterns can have π-orbital interactions, subtle changes in the packing patterns results in big differences in the semiconductance of a material.

2.1 Molecular Structure–Property Relationships

For a successful *p*-type OSC, the energy and shape of the HOMO is important. First of all, the ionization energy (IE) determines ambient stability and charge carrier polarity of the material, whether it be *p*-type or *n*-type.

Before we move on to the discussion of the structure–property relations at molecular scale, we should note here one technical issue, which is how to compare experimental and computational values of the electronic structure of the materials. Although the important quantity for material characteristics is the IE, usually the HOMO energy levels from the quantum chemical calculations are compared to the experimental IEs for practical reasons. As can be seen in Fig. 7, the HOMO energy levels closely follow the IEs from the photoelectron spectrum. In reality the difference between the gas and solid-state IEs is large and depends on the packing and, more precisely, on the nature of the π-orbital interactions. For example, in pentacene the gas phase IE is 6.6 eV whereas in thin films it is around 5 eV depending on the packing patterns; 4.8 eV if the molecules are standing up and 5.35 eV if they are lying over a substrate [101]. From the computational cost perspective, gas phase IE calculations are now routine, but IE calculations for the solid state are much more complicated. Therefore, the HOMO energy levels are

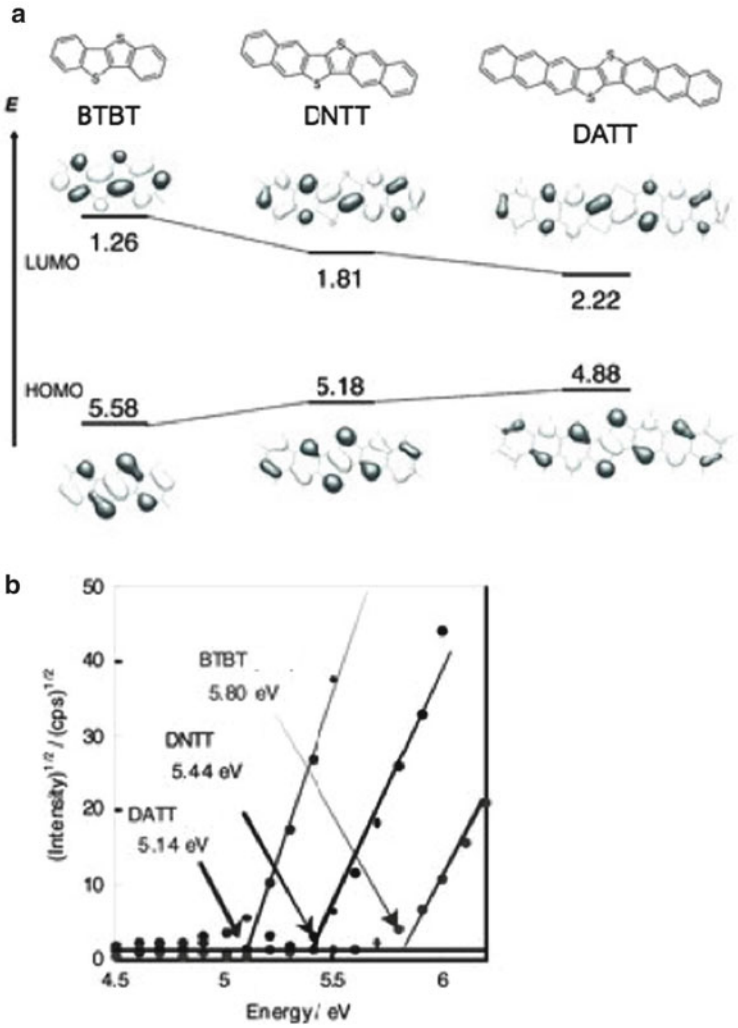


Fig. 7 The frontier orbitals of BTBT, DNTT, and DATT (**a**) and their photoelectron spectrum in air (**b**). (Reprinted with permission from [100]. Copyright (2011) Wiley-VCH, Weinheim)

usually used instead of the IE values for practical reasons. Our discussion will also follow the same approach.

For *p*-type OSCs, the HOMO level is usually around 5 eV, and thus very close to the oxidation threshold of 5.1 eV. The more extended the conjugation, the higher the HOMO level (see Fig. 7). Thus extending the conjugation without making the material unstable in ambient conditions is a challenge and an active area of research. For example, the introduction of a thienothiophene group in the middle of an acene (Fig. 5) has been shown to help stabilize a longer molecule [102], or the

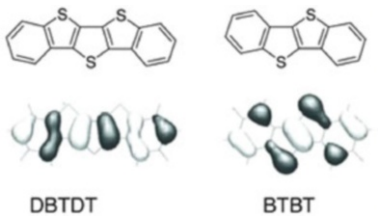


Fig. 8 The HOMO orbitals of DBTDT (*left*) and DNTT (*right*). (Reprinted with permission from [100]. Copyright (2011) Wiley-VCH, Weinheim)

substitutions with electron-withdrawing groups usually leads to deeper HOMO levels [103].

The shape of the HOMO is another important electronic property that affects the performance [100]. If the nodal planes of the HOMO lie on the atoms such as sulfur atoms, which are important for intermolecular interactions, the strength of the intermolecular interactions and thus the electronic coupling are reduced. For example, in molecules with thiophenes, sulfur–sulfur interactions are very important in the crystal and maintaining the short contacts among these atoms is important for stronger couplings. Figure 8 shows the shape of the HOMO level of molecules labeled as DBTDT and BTBT. DBTDT has nodal planes over the sulfur atoms whereas BTBT does not. As suggested by Takimiya and coworkers, DBTDT has larger S···S distances as well as much weaker transfer integral values: 11 and 17 meV [100] compared to transfer integral values of 67 and 26 meV in BTBT-C_8 [104].

Together with the chemical structure and availability of the sp^2 hybridized atoms, the size and rigidity of the molecule determines the extent of the conjugation. This, in turn, affects several parameters related to charge transport. One important consequence of the size of the molecule is the strength of the local electron–phonon coupling. Let us assume that a single charge carrier residing on a molecular site with a rigid conjugated backbone can simply be modeled as a particle in a box. Then we can consider that the reorganization energy required to accommodate the charge carrier scales with the length of the box. Usually, the more extended the conjugation, the smaller the reorganization energy and the better the charge transport properties. However, this is only true for a set of similar compounds. In the following examples, we explain what we mean by similar.

Since hexacene has recently been synthesized and its OFET mobility has been measured [105], the oligoacene family of compounds provide a good set to illustrate some of the structure–property relationships discussed above. The charge transport parameters of the members of the oligoacene family with $n = 2$–6 phenyl rings are presented in Table 1. The transfer integral values listed correspond to the unique dimers shown in Fig. 9b–c.

The trends observed for the reorganization energy, transfer integrals, and ensuing mobilities in oligoacenes confirm the discussion above. First, the more extended the conjugation, the higher the HOMO energy level and the smaller the reorganization energy. Second, although the transfer integral for the parallel arrangement, P, does not improve as the intermolecular distance also gets larger, the edge to face

Table 1 Calculated hole transport properties of oligoacenes. For details of the calculations see [105] and references therein. (Reprinted (adapted) with permission from [105]. Copyright (2012) Nature Publishing Group)

			R(Å), Transfer integrals (meV)				
Compound	HOMO (eV)	Reorganization energy (meV)	T_1	T_2	P	L	Mobility ($cm^2\ V^{-1}\ s^{-1}$)
Naphthalene	−5.80	183	5.01, 8	5.01, 8	5.93, 36	8.64, 0	0.0511
Anthracene	−5.24	138	5.22, 19	5.22, 19	6.01, 42	11.12, 0	0.158
Tetracene	−4.87	113	4.77, 70	5.13, 22	6.06, 37	13.44, 1	0.470
Pentacene	−4.61	95	4.76, 79	5.21, 45	6.27, 31	16.11, 1	0.832
Hexacene	−4.42	79	4.72, 88	5.22, 60	6.31, 37	18.61, 1	1.461

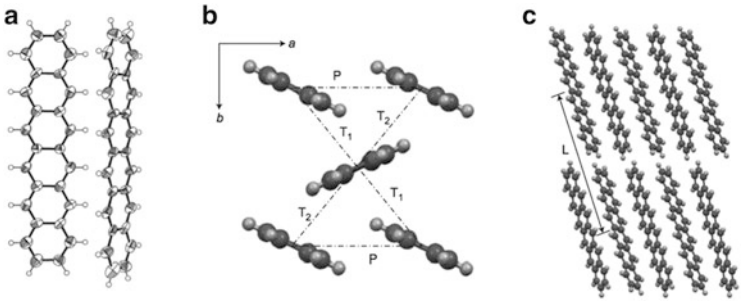


Fig. 9 (**a**) ORTEP drawing of two adjacent hexacene molecules. (**b**) Arrangement of hexacene molecules in the *ab*-plane. (**c**) Arrays of hexacene along *a*-axis. The labels *T*, *P*, *L* denote the pairs of molecules used in the transfer integral calculations. (Reprinted (adapted) with permission from [105]. Copyright (2012) Nature Publishing Group)

interaction terms, T_1 and T_2, improve going from $n = 2$ to $n = 6$. Together with the smaller reorganization energies for the larger oligoacenes, the estimated mobility values improve as well. Similar behavior was also observed for oligothiophenes: the longer the oligomers, the smaller the calculated reorganization energy [106]. Experimental evidence also shows that the longer oligothiophenes, four to six rings, have higher mobilities compared to shorter ones [107]. Usually four or more rings are required in an OSC molecule for good performance.

For structurally different families of compounds, such as heteroacenes, the reorganization energy trends need to be carefully analyzed and molecular similarity needs to be taken into account. For example, the thienoacenes usually have higher reorganization energies compared to oligoacenes (see Fig. 10). However, the polyphenyls have quite large reorganization energy values compared to acenes, even higher than polythiophenes. The trends can depend on structural factors such as the type of bonds, the availability of the torsional angles, or the spatial arrangement of the aromatic rings. Following the classification suggested by Takimiya and coworkers [100], in Figs. 10 and 11 we show that if the structural similarities are taken into account and the acene- and thiophene-based OSC molecules are grouped accordingly, the extended conjugation lowers the reorganization energy. However, the slopes of the curves representing the change in the reorganization energy as a function of the number of rings could be very different for each group of compounds.

One avenue through which to pursue a priori prediction of the high performance OSC molecules is the use of quantum chemical descriptors in quantitative structure–property (QSPR) studies. By combining molecular descriptors from cheminformatics with ab initio quantum chemistry, Misra et al. showed that it is possible to accelerate the search for smaller reorganization energy materials [108]. In a study of a family of 200 polycyclic aromatic hydrocarbons, by assuming that all of the molecules form similar liquid crystalline stacks, the reorganization energy was identified as the dominant factor influencing the charge transfer rates.

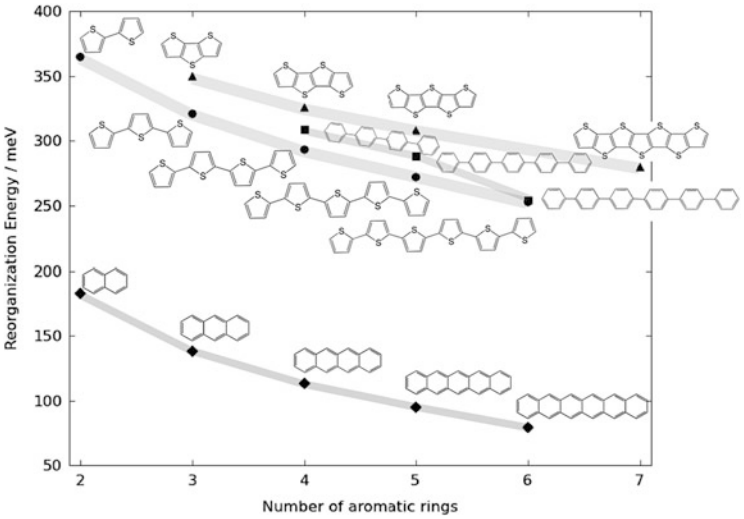


Fig. 10 The reorganization energies of oligoacene and thienoacene families of homocycles. References for the reorganization energy values: oligoacenes (*diamonds*) [105], polythiophenes (*circles*) [106], fused-thiophenes (*triangles*) [25], polyphenyls (*squares*). (Atahan-Evrenk and Aspuru-Guzik (2012), unpublished results)

Based on a QSPR study of various molecular descriptors, a weak correlation of the reorganization energy with the molecular signature (a canonical representation of the atom's environment up to a preferred height [109]) and another weak correlation with the electronic eigenvalue descriptors were identified. Then these two weakly correlated descriptors were combined into a QSPR model that estimated 80% of the reorganization energies within an error margin of 20 meV. This success rate was quite good as the only quantum chemistry calculation involved was the ground state geometry optimization of the neutral molecular structure.

Due to the benefits of low reorganization energy in the charge transport, a few strategies based on structural modifications are developed. For example, the presence of the nonbonding character in the HOMO level is identified as a reorganization energy lowering strategy. A comparison of tetraceno[2,3-*c*]thiophene (TcTH) and tetraceno[2,3-*b*]thiophene (TbTH) showed that the fusing of thiophene in a symmetric manner results in 30 meV smaller reorganization energy in TcTH (Fig. 12a) compared to TbTH (Fig. 12b) [110, 111]. Based on the same principle, the cyanide substitutions, and substitutions by weaker electron-withdrawing groups (for example, chlorination instead of fluorination) has been identified as a reorganization energy lowering strategy [112–115]. In addition, intra-ring substitutions such as the replacement of a carbon atom with nitrogen were also shown to lower the reorganization energy.

Last but not the least, choosing the right molecular structure has significant consequences for solid-state packing. The molecular group symmetry, polarizability, side group positions, the intermolecular interactions, etc., play a role in

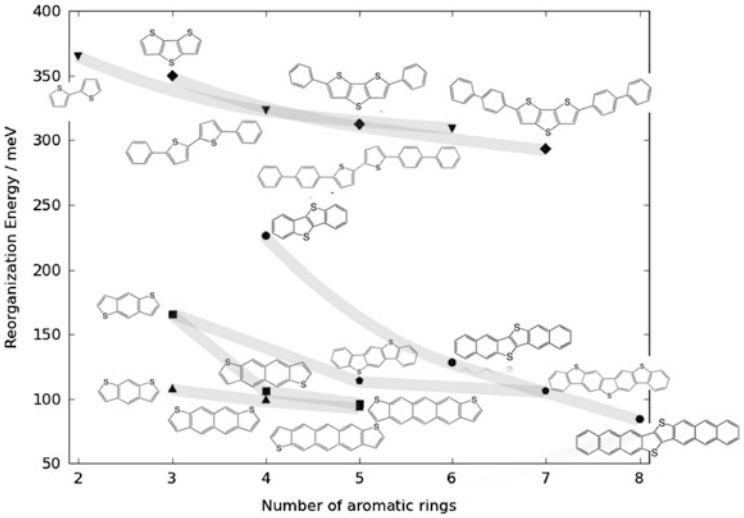


Fig. 11 The reorganization energies of thienoacene heterocycles: phenyl substituted dithienothiophenes (*diamonds*) (Atahan-Evrenk and Aspuru-Guzik (2012). Unpublished results), phenyl substituted dithiophenes (*down triangles*) (Atahan-Evrenk and Aspuru-Guzik (2012), unpublished results), diacene-fused thienothiophenes (*circles*) ([6], Atahan-Evrenk and Aspuru-Guzik (2012), unpublished results), benzene-thiophene alternating molecules (*pentagons*) ([25], Atahan-Evrenk and Aspuru-Guzik (2012), unpublished results), acene-anti-dithiophenes (*squares*) ([25], Atahan-Evrenk and Aspuru-Guzik (2012), unpublished results), acene-syn-dithiophenes (*triangles*) ([25], Atahan-Evrenk and Aspuru-Guzik (2012), unpublished results)

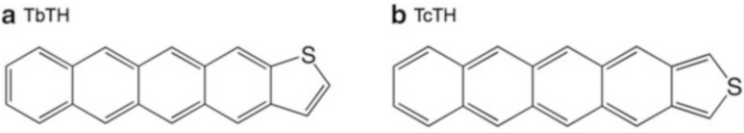


Fig. 12 Symmetric substitution leads to lower reorganization energy, 97 meV of TbTH compared to 66 meV in TcTH [110]

determining the stable lattice conformations, thus leading to enhanced or reduced mobility. For example, relative abundance of C–H (through the peripheral H-atoms) vs C–C interactions affect the tendency to pack in herringbone or π-stacked packing forms [116]. The molecules with H-bonding groups such as quinacridones can form supramolecular synthons [32, 117]. The symmetry of the molecular structures also affects the packing in the solid state. For example, molecules with centrosymmetry can form better short-contact networks [100]. In the following section we further discuss the effect of the modification of the molecular units to achieve high performance and the crystal structure–property relationships.

2.2 Crystal Structure–Property Relationships

Among the crystal structure–property relationship studies in OSCs, perhaps the most striking example is the case of rubrene (Fig. 4d). With a record single crystal OFET mobility of 20 $cm^2\ V^{-1}\ s^{-1}$ [58], rubrene was a remarkable improvement from tetracene. Although the phenyl groups added to tetracene backbone do not participate in the π-conjugation, they help the tetracene backbone to π-stack right very close to one of the maxima of the transfer integral surface [118]. The phenyl groups also provide a lock-in for the backbone so that displacement along the short axis, which usually yields lower transfer integrals in oligoacenes, is prevented. Understanding of the rubrene crystal structure provided motivation for further studies to control the solid-state packing to enhance performance.

To study further the structure–property relationships in rubrene and its derivatives, the effect of the substitutions of the external phenyl rings on the crystal structure was examined [119, 120]. Haas et al. showed that, depending on the positions of the substitutions, the π-stacking geometries and the interlayer distances between the backbones could be controlled. In particular, they have found that the substitutions on the 5,11 phenyls cause a large twist in the backbone leading to a polymorph with no charge carrier mobility, whereas the *tert*-butyl substitutions on the 5,12 phenyls showed a similar packing pattern to rubrene with similar OFET characteristics despite a 31% increase in the interlayer spacing. In a recent study, McGarry et al. [120] showed that the interlayer distance could also be controlled by the methyl and trifluoromethyl substitutions of the external phenyl rings. By means of this approach, rubrene derivatives with OFET mobility comparable to rubrene were synthesized. In particular, they had a success with dual substitutions by trifluoromethyls that resulted in ambipolar function as well as a high hole mobility.

Another successful example of charge transport tuning with structural modification is the family of substituted pentacenes [121]. Among the *peri*-functionalized pentacenes, tips-pentacene (Fig. 4a) with triisopropylsilylethynyl groups at the 6,13 positions emerged as one of the most successful engineering examples [122]. Along with solution processability and ambient stability, the substitutions changed the herringbone type of packing into π-stacking conformation. Subsequently by the solution shearing method a metastable state of tips-pentacene with better performance has been achieved with an order of magnitude increase in the mobility [123]. The theoretical investigation of the transfer integrals corroborated the experimental findings and showed that the solution-sheared films indeed have three times stronger transfer integrals [123]. Recently Bao and co-workers showed even higher mobility of 11 $cm^2\ V^{-1}\ s^{-1}$ for a solution processed film of tips-pentacene, demonstrating the fact that the control of the solution deposition is crucial for high performance materials [59]. Tips-pentacene and rubrene are among the few OSCs which show band-like transport in an OFET setup.

Another crystal packing control strategy is the engineering of sulfur–sulfur interactions. For example, the chalcogen substitutions at the peri-positions of pentacene promotes the π-stacking interactions [124]. In comparison to the

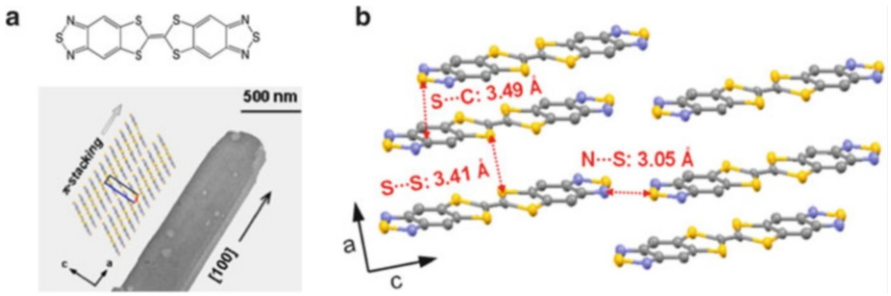


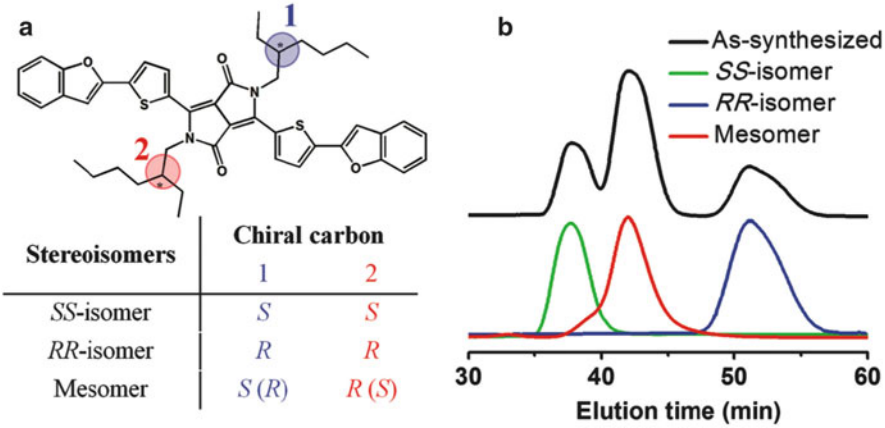
Fig. 13 Benzothiadiazole-tetrathiafulvalene with π-stacking interactions for single-crystal nanowire applications (**a**), short contacts (**b**). (Reprinted (adapted) with permission from [125]. Copyright (2013) American Chemical Society)

herringbone packing in pentacene, hexathiapentacene has strong π-stacking and unusually short intermolecular sulfur–sulfur distances. These substitutions usually lead to one-dimensional conduction and thus this strategy has the potential to develop semiconducting nanowires. Similarly for benzothiadiazole-tetrathiafulvalene single-crystal nanowires (see Fig. 13), the use of tetrathiafulvalene groups with strong S···S interactions and thiadiazole groups promotes the π-stacking interactions, and the intermolecular stacking distances could be controlled and short-contacts such as S···S (3.41 Å), S···C (3.49 Å), and S···N (3.05 Å) could be achieved [125]. Further discussion about additional strategies for inducing π-stacking could be found in [126].

Although one-dimensional transport could be advantageous for nanowire applications, reducing the dimensionality of the mobility has the potential to make the charge transport very susceptible to the presence of defects [127]. A higher dimensional transport network is desirable for OFET applications as it provides alternative pathways for transport. In recent work, for example, the methylated DNTT derivatives are synthesized and significant transfer integrals in all three dimensions is demonstrated [128].

The engineering of the crystal structures through bulky substitutions has also been studied in the case of oligothiophenes with trimethylsilane end groups [129]. It has been shown that through end-substitutions the in plane tilt and the offset of the oligothiophene backbones can be controlled and better or worse electronic couplings among the molecules can be enforced.

Another example of the substitutional engineering of the crystal structure is the end-substituted 5,5′bis(4-alkylphenyl)-2,2′-bithiophenes (P2TPs) [130]. Depending on the even and odd alkyl chain substitutions, the tilt angle and electronic couplings in the two dimensional deposition layers of the SCs can be controlled. The solid-state packing of P2TPs molecules with different length alkyl side chains ($n = 3,8$) were analyzed with grazing incidence X-ray diffraction. The data showed that having an even or an odd length chain controls the tilt angles of P2TP molecules on the self-assembled monolayers of ODTS. The even length alkyl chains yielded larger tilt angles and higher mobility compared to those with an odd number of



Stereoisomers	Chiral carbon	
	1	2
SS-isomer	*S*	*S*
RR-isomer	*R*	*R*
Mesomer	*S* (*R*)	*R* (*S*)

Fig. 14 The chemical structure of DPP(TBFu)$_2$ with two chiral carbons highlighted (1 and 2) and the resulting stereoisomers (**a**), elution profiles for the as-synthesized and isolated stereoisomers (**b**). (Copyright reference [131])

carbon atoms. The transfer integrals for the unique dimers extracted from the corresponding crystal structures also corroborated the odd-even trend. In addition, MD simulation of the surfaces showed that the tilt angle is indeed controlled by the parity of the side chains.

In addition to the position and the length of the side groups, stereoisomerism affects the FET mobility [131]. For example, due to *RS* arrangement of the chiral carbons in DPP(TBFu)$_2$ (labeled **1** and **2** in Fig. 14), the mesomer has been shown to have better π-orbital stacking interaction as well as higher mobility compared to the case of *SS* or *RR* isomers. The mesomeric form had a packing distance of 3.38 Å with a flat conjugated backbone compared to 3.47 Å in the case of SS or RR isomers. The mesomeric form also performed better than the case where no stereoisomer was selected (as-synthesized). Therefore, this study highlighted that chirality of the substituted groups also needs to taken into consideration for tuning of properties with structural modifications.

Another important concept in understanding the crystal structure–property relationships is the polymorphism in OSCs. Like in drug molecules, where polymorphism has direct consequences for the solubility and thus bioavailability of a drug, in OSCs polymorphism has important consequences for the charge transport properties. The polymorphs of an OSC can have totally different mobility tensors. For example, Fig. 15 illustrates the possible variations in the (an)isotropy in the mobility tensor for four polymorphs of pentacene [132]. Among four different polymorphs investigated, apart from one (polymorph IV), all of them showed remarkable anisotropy in the transport plane. This might have important implications for the device configurations.

It is also possible that several polymorphs of an OSC with different charge transport characteristics could be accessible in operating temperatures. It has been shown, for example, that, for a fluorinated derivative of an anthradithiophene, a

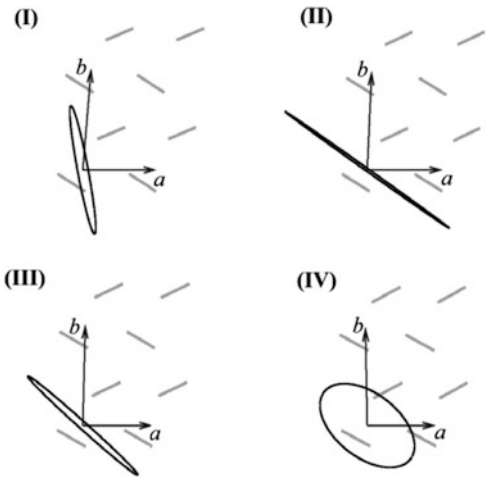


Fig. 15 The mobility tensors in the *ab*-plane for four polymorphs of pentacene. (Reprinted (adapted) with permission from [132]. Copyright (2005) American Chemical Society)

phase transition between two polymorphs occurs as the temperature rises from 260 K to 300 K [29]. For this system, the field-effect mobility of the films increases as a function of increasing temperature with particular slopes for each polymorph. This finding has immense significance for the consequences of polymorphism in OSCs in commercial applications.

Lastly, we would like to discuss the role of polarizability in the electronic structure and charge transport mechanisms [133]. The electronic structure is highly influenced by the molecular polarizability and the geometrical arrangement of the molecules. In OSCs, there is, for example, about 1 eV difference between the IEs in the gas phase and the solid state. This difference is due the polarization energy [134]. For example, the IE of pentacene in the gas phase is about 6.6 eV compared to about 5 eV in thin films. Moreover, it is 5.35 eV when the molecules lies flat over a substrate, and 4.8 eV when they are standing [101].

An important observation related to the correlation of the molecular polarizability to the band-like transport was discussed recently by Minder et al. [135]. In particular, they correlated the observation of the band-like charge transport in OFETs with the polarizability of the conjugated backbones as well as the collective dielectric response of the OSC crystal. In particular, for a group of compounds shown in Fig. 16, they identified two important factors potentially leading to band-like transport in OFETs: (1) the presence of substituents to influence the coupling among the adjacent OSC layers and (2) the orientation of the molecules and their polarizability tensors with respect to the charge transport direction. They argued that PDIF-CN_2 and BTBT-C_8 show band-like transport because the dielectric coupling of the adjacent layer as well as the gate is screened by the presence of the side groups. This notion is based on the understanding that the conductance channel is the first layer of the OSC film over the gate dielectric and charge transport mostly happens in this first layer. In addition, the alignment of the

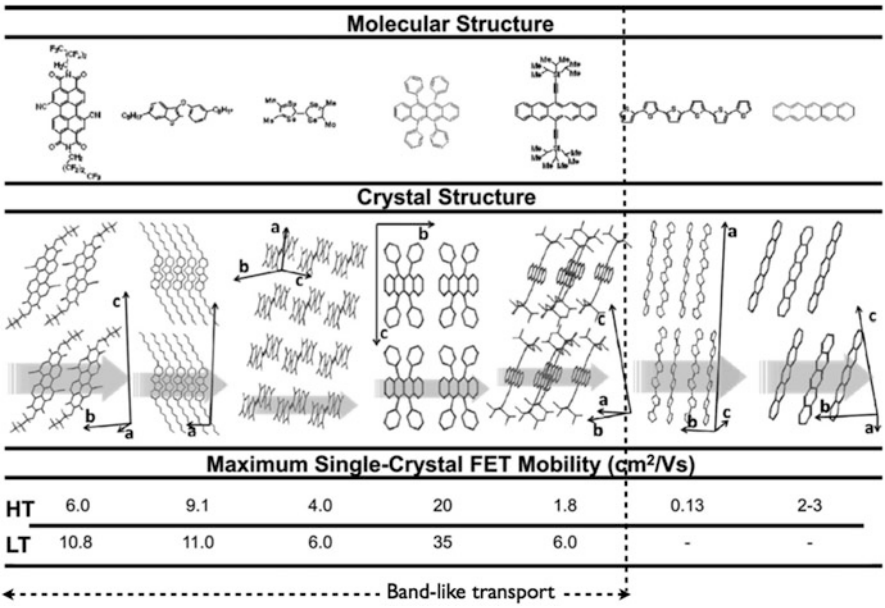


Fig. 16 Analysis of the molecular structure and packing for different organic molecules (from *left* to *right*: PDIF-CN_2, BTBT-C_8, TMTSF, rubrene, TIPS-pentacene, sexithiophene, and pentacene). To date, PDIF-CN_2, BTBT-C_8, TMTSF, rubrene, and TIPS-pentacene are the only organic semiconductors exhibiting band-like transport in an OFET configuration. The values of mobility are either at room temperature (HT) or low temperature (LT–for the molecules in which band-like transport has been observed). (Reprinted (adapted) with permission from [135]. Copyright (2012) Wiley-VCH, Weinheim)

molecules with respect to the gate is crucial: the dielectric coupling is smaller if they are parallel to the gate than when they are perpendicular to it. Therefore, band-like transport is also observed in TMTSF, rubrene, and tips-pentacene since the conjugated molecular backbones in these crystals lie parallel to the gate. They also argued that pentacene and oligothiophene have never shown band-like transport in an OFET setting because they are arranged perpendicular to the dielectric surface and have no terminal alkyl chains to reduce the coupling among the adjacent layers. Further work to quantify and measure these observations would be of great value to the design and engineering of high-performance OSC materials.

3 Crystal Structure Prediction for Organic Semiconductors

Arguably, the most important piece of information needed for the theoretical characterization of an OSC crystal is the availability of the crystal structure. As illustrated in the previous sections, the literature is full of analyses of the crystal

structures of experimentally known materials. The crystal structure prediction for OSCs a priori to synthesis, however, is scarce. Except for a few cases of limited solid-state structure prediction [4, 6, 136], it remains virgin territory. We associate the sporadic interest with lack of commercial applications. In the near future we expect that OSCs will be used in commercial applications ubiquitously and hence a more detailed understanding of the polymorphism or structural stability will be demanded by the market.

From the crystal structure prediction point of view, one advantage of the molecular OSCs with conjugated backbones is that they have mostly two-dimensional geometries and pack in a few specific patterns. For example, many years ago Desiraju and Gavezzotti [116] studied the crystal structures of polynuclear aromatic hydrocarbons. They showed that all of the aromatic hydrocarbons could be grouped into four distinct packing patterns: (1) herringbone pattern (polythiophene, pentacene), (2) pair-wise (*sandwich*) herringbone pattern (perylene (α phase), pyrene), (3) flattened-herringbone pattern (coronene), and (4) graphitic pattern (tribenzopyrene). Based on their analysis of the structures in terms of the presence of the type of intermolecular interaction, i.e., C$\cdots$H, C$\cdots$C, or H$\cdots$H, they successfully predicted the packing type of some of the unknown structures at the time such as the sandwich structure for the benzo(*e*)pyrene [137]. Nevertheless, the most practical applications of OSCs involve intra-ring substitutions, heterocycles, or side group additions to the backbone which complicate the crystal structure prediction.

Most crystal structure prediction studies for OSCs reported to date involve the structures of the known crystals of well-studied molecules such as oligoacenes or oligothiophenes. Usually a conventional molecular force field is used in conjunction with electrostatic potential (ESP) fitted point charges to describe the interactions in an experimentally known crystal structure. These studies provide a good basis for understanding the performance of the force fields and the type of improvements required in the force field parameters for particular systems. For example, Marcon and Raos [138] studied the crystalline oligothiophenes with improved MM3 force fields. In particular, they optimized the inter-ring torsion potentials and calculated the atomic charges as well as higher terms of the electrostatic interactions derived from distributed multipole analysis (DMA) with quantum chemistry methods. The systems studied ranged from herringbone packing structures (α-tetrathiophene, α-sexithiophene) to π-stacked configurations as well as an alkyl-chain substituted sexithiophene. Interestingly, they concluded that the MM3 force field with point atomic charges gave satisfactory results for all the *p*-type OSCs they have studied. Only the *n*-type OSC, perfluorosexithiophene, required the more accurate electrostatic modeling through the DMA. Therefore, they concluded that the more costly DMA approach is not justified for the *p*-type OSCs studied. MD simulations at the same level of theory are performed to investigate the effect of the temperature, and to ensure a better comparison of the predicted structures with the structures from room temperature X-ray diffraction. They found that MD simulations at room temperature systematically resulted in approximately 10% lower crystal densities. Among several variations of MM3 force field adapted for the

oligothiophenes, MM3 with the ab initio corrected torsion potentials, ESP charges, and adjustment of the dispersion terms with scaling of the dispersion parameters showed the best performance.

The known polymorphs of oligoacenes and oligothiophenes also provide a basis for the test of effective strategies for initial structure search and polymorph prediction. Della Valle and coworkers studied the polymorphs of tetracene [139], pentacene [140, 141], and sexithiophene [142], with the assumption of rigid bodies for the molecular structures. They employed a hybrid approach involving uniform sampling (low-discrepancy Sobol' sequence) of the energy landscape by including the crystallographic symmetry constraints. The completeness of the search space is maintained by adapting the capture-recapture method from wildlife ecology. The intermolecular potentials are described by atom–atom interactions where electrostatic interactions are defined in terms of the point charges derived from quantum mechanical calculations. In particular, for sexithiophene they employed an AMBER force field with restricted ESP fitted charges, and for pentacene and tetracene an atom–atom Buckingham model with Williams parameter set IV [143] (with ESP charges for the tetracene) is used. For tetracene and pentacene the initial search was constrained to triclinic structures with two independent molecules per unit cell. After the lattice energy minimizations, the crystal space groups are assigned to the optimized lattices by the use of PLATON program [144]. For sexithiophene the initial search was constrained to 16 structural classes with triclinic, monoclinic, and orthorhombic lattice types. In the case of tetracene and pentacene, most of the minima belonged to the $P1$ and $P2_1/c$ groups and have the layered herringbone-type packing. The deepest minima for tetracene corresponded to the high temperature-low pressure polymorph known by X-ray diffraction. From the global search for the low energy polymorphs of pentacene, they successfully found the C and H polymorphs as their rank 1 and rank 2 structures. In the case of sexithiophene among nine deepest minima structures, six different space groups are observed, lowest energy polymorphs also having herringbone-type packing (see Fig. 17). Again, the two known polymorphs of sexithiophene have been identified correctly with this methodology. Unfortunately, the same approach failed in the blind tests and Della Valle et al. concluded that the success in the case of tetracene, pentacene, and sexithiophene could be attributed to the symmetry, rigidity, and planarity of these molecules and the presence of the short-range isotropic interactions [142].

Another approach to speed up the optimization of the solid form of materials is based on genetic algorithms (GA), which are widely used in the crystal structure prediction of inorganic materials [145]. Facelli and co-workers [146] adapted the GA for the prediction of the crystal structures of benzene, naphthalene, and anthracene. In the GA approach, genes are a set of geometric parameters defining the spatial arrangements of the rigid bodies in three-dimensional space. Then the genomes are constructed as the collection of parameters defining the crystallographic axes, molecular positions, orientations, and number of molecules in the unit cell. Although no assumptions are made for the crystallographic axes, the unit cell parameters are limited to a certain region of the space to make the problem more

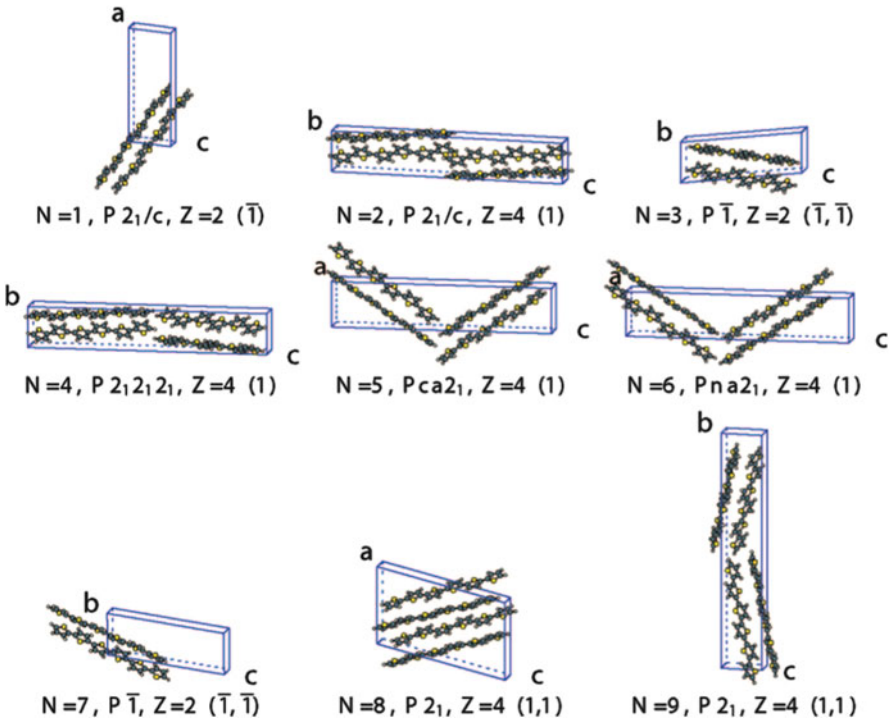


Fig. 17 Structure of the nine deepest minima, shown with an orientation in which the shortest cell axis (either *a* or *b*) is approximately perpendicular to the plane of the page. Minima are labeled by their energy rank *N* (also indicated in Table 2) and structural class (space group, *Z*, and site symmetry). (Reprinted (adapted) with permission from [142]. Copyright (2008) American Chemical Society)

tractable. One advantage of this approach is that there are no assumptions on the *Z* values, i.e., the number of molecules in the unit cell. At every step of the optimization, the lattice energies are calculated with an empirical force field and the lattices leading to higher energies are discarded. Then through the mutations and crossovers performed on the genomes, new structures were generated and minimized until a certain convergence criteria for the lattice energy was reached. Again, here the empirical force fields used were based on van der Waals and electrostatic terms with atomic point charges [143]. Despite the many assumptions involved, the resulting structures were surprisingly successful with lattice energy differences of 1% from the experimental crystals. In addition to the experimentally known structures, a new set of polymorphs was determined by the GA approach. Although this work assumed rigid-bodies, an adaptation of the same algorithm to flexible molecules was also demonstrated [147]. Despite the use of a more advanced force field, the Amber force field used in CHARMM optimizations, the same success for the structure prediction was not observed for a set of flexible molecules. It is perhaps indicated that the conjugated backbone, and hence relative

rigidity and flatness of the OSCs, significantly simplify the search problem, and therefore relatively simpler force field parameter sets can provide enough accuracy.

One of the strategies to limit the search space is to take advantage of the fact that 93% of all the space groups of organic molecular solids reported in the Cambridge Structural Database belongs to the 18 most frequent space groups among a total number of 230 [148]. Moreover, about 92% of all organic molecular solids have only one molecule in the asymmetric unit cell [149]. Therefore the initial search space with different packing forms are usually limited to a subset of the most prevalent space groups. We should note however that reported structures in the database does not include all the polymorphs for various reasons [150]. Although the omission of the less likely space groups could provide enough statistics, it does not guarantee that some of the polymorphs of a material might not have been left out. Indeed, in the last blind test, the most successful searches were those that spanned all possible space groups.

For screening many molecules with the potential for high performance, the speed at which the lattice configurations can be ranked is crucial. In the following we summarize an even simpler approach for the structure prediction based on the structural and molecular similarities among molecules. This practical approach is analogous to homology modeling in drug design. Taking advantage of the molecular structural similarities, a known crystal structures is adapted for a new molecule and subsequently optimized with electronic structure methods. A similar strategy, ligand replacement, is also adapted for the prediction of hybrid frameworks [151].

In 2007 Takimiya and coworkers reported two important compounds: BTBT and DNTT (Fig. 7a) with high hole mobility and extraordinary shelf-life. DNTT showed characteristics comparable to pentacene but better ambient stability. Our analysis [152] of the microscopic charge transport parameters of DNTT confirmed that extended aromatic structures have small reorganization energy and mildly anisotropic electronic couplings in the herringbone packing plane. Moreover, the non-local electron–phonon couplings investigated in terms of the thermal modulation of the transfer integrals were found to be weaker, for example, compared to pentacene. The transfer integrals were calculated (at the level of B3LYP/6-31G*) for the unique dimers extracted from an MD trajectory with an MM3 force field with ESP fitted charges. Figure 18 shows the moderate anisotropy of the mobility (Fig. 18b, c) and the weak thermal modulation of the transfer integrals (Fig. 18d).

This analysis motivated us to study a small library of thienoacene derivatives shown in Fig. 19 [6].

As a first strategy of ranking of the molecules in this library, the gas phase geometries were optimized and internal reorganization energies calculated. The two compounds, **2** and **7**, with the smaller reorganization energies (smaller than pentacene reorganization energy of 95 meV) of 85 and 77 meV respectively, were identified as potential high performance materials. Subsequently, the crystal structure minimizations for these two molecules were performed. This involved the optimization of the unit cells derived from the parent compound (DNTT) with Dreiding force field and ESP fitted charges as implemented in the Forcite module [154]. The symmetry imposed ($P2_1$) and the symmetry relaxed (P1) optimizations

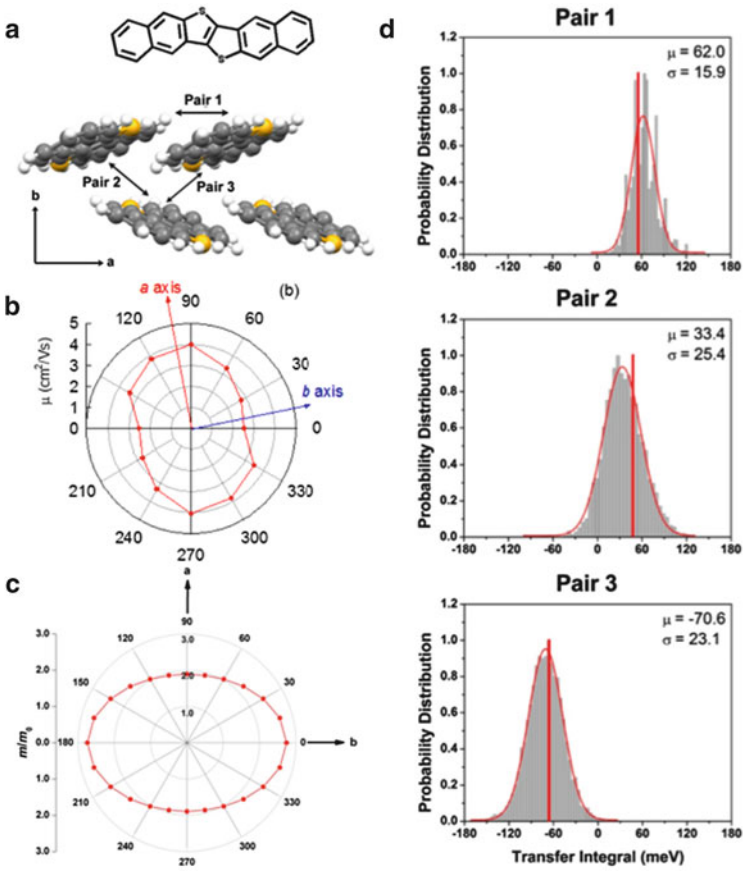


Fig. 18 (**a**) The packing in the *ab*-plane in DNTT crystals and molecular dimer pair investigated for transfer integral calculations. (**b**) The experimental moderate anisotropy in the *ab*-plane. (Reprinted with permission from [153]. Copyright (2009) AIP Publishing LLC). (**c**) Effective charge carrier masses obtained from the dispersion of the valence bands [95]. (**d**) The probability distributions for the transfer integrals extracted from an MD trajectory; the *vertical lines* correspond to the transfer integrals from the optimized equilibrium structures with the same level of theory. (Reprinted with permission from [95]. Copyright (2010) American Chemical Society)

lead to similar packing. Finally, the transfer integrals for the dimers from the optimized structures were calculated and, based on approximate Marcus electron transfer rates, the more promising material was identified as molecule **2**. This material, for which a performance better than pentacene had been anticipated, was eventually synthesized and OFET measurements of the single crystals revealed a record high mobility of 12 $cm^2\ V^{-1}\ s^{-1}$ [6].

The crystal structure was confirmed with powder X-ray with remarkably good agreement. We further adjusted the unit cell parameters by matching the experimental and predicted patterns (see Fig. 20). The lower density predicted by the

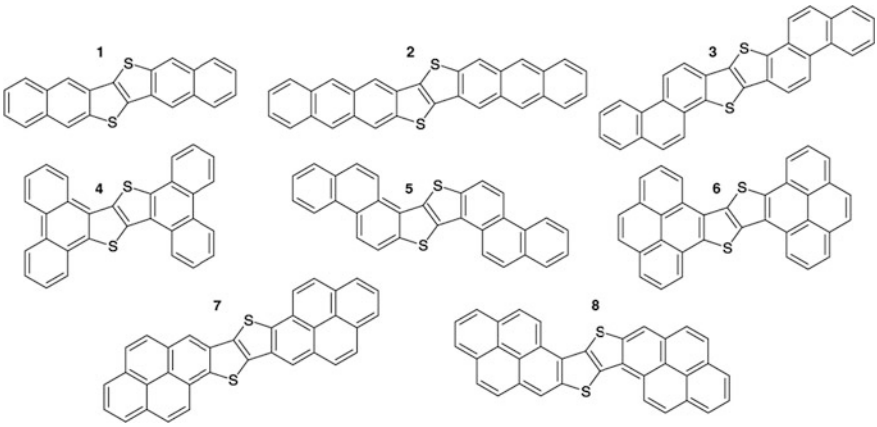


Fig. 19 DNTT (**1**) derivatives studied for charge transport capacity [6]

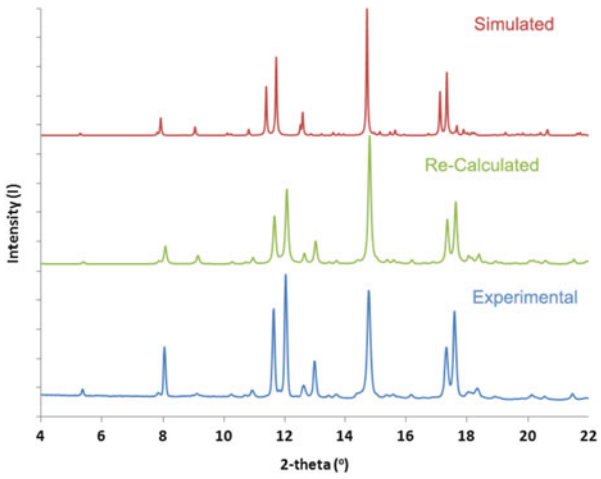


Fig. 20 The powder X-ray diffraction pattern of molecule 2 from experimental powder (*bottom*), the re-calculated structure by matching the simulated structure to the experimental powder pattern (*middle*), and the original predicted structure (*top*). (Powder simulation wavelength = 0.9758)

MM3 with ESP charges is a known property of this approximation unless the dispersion coefficients of the R^{-6} terms are optimized for the particular system [138]. Although the cell volume error in Table 2 is 6%, in reality it is larger as there is usually an expansion of 6% of the volume as the temperature rises from 0 to 300 K. Since the density of the experimental structure is larger, the ensuing transfer integral values improved. Despite its simplicity, this approach has led to the discovery of a new high-performance materials as well as helped the elucidation

Table 2 The unit cell parameters for the predicted crystal structure for molecule 2 and for the crystal structure matched to the powder X-ray spectrum

Cell properties	Predicted	Matched to powder X-ray
Space group	$P2_1$	$P2_1$
A (Å)	6.444	6.225
B (Å)	7.6217	7.577
C (Å)	21.203	20.824
α (deg)	90	90
β (deg)	89.24	87.23
γ (deg)	90	90
Cell volume ($Å^3$)	1,041.21	981.05

of the crystal structures from powder patterns. Again, for molecule 2 the rigid and planar molecular structure simplified the crystal prediction process tremendously. Similarly, Chang et al. predicted the polymorph structures for the substituted tetracenes with Dreiding forcefield with ESP fitted charges [136]. Subsequently, based on the semiclassical charge transfer rate from Marcus theory, they have identified two π-stacking high-performance compounds. These compounds as yet to be synthesized.

Although we have focused on the single-crystals of *p*-type OSCs, we would like to point out that the prediction of thin-film structures and the effect of various substrates on the film growth have drawn considerable interest over the years and a good summary discussing the recent developments can be found in [155].

4 Conclusions and Outlook

In this chapter we have discussed the molecular and crystal structure–property relationships as well as the state-of-the-art crystal structure prediction for OSCs. We focused on the *p*-type OSCs for OFET applications. Although we have a good understanding of structure–property relationships through the theoretical and experimental characterization of OSC crystal structures obtained from X-ray diffraction, the prediction of new OSCs from molecular structures is still in its infancy. There have been successful predictions of solid forms of rigid and planar OSC molecules such as pentacene; nevertheless, the more general CSP methodologies to predict crystals of OSCs with functional groups or long alkyl chains remain to be carried out.

The recent blind tests organized by the Cambridge Structural Database showed significant progress towards the prediction of the crystal structures of organic molecular solids [37, 38]. To date, studies on the crystal structure prediction of OSCs have been limited to classical force fields with mostly ESP fitted charges. We believe that there is room for a great deal of improvement by adaptation of the state-of-the-art methodologies that were successful in the blind tests.

If crystal structure prediction for OSCs can be realized, together with theoretical characterization techniques, it can provide valuable guidance for better OSC design. For example, one important area where crystal structure prediction has tremendous potential to facilitate OSC engineering is the optimization of side groups added for solubility [30]. If possible, these studies would tell the synthesis experts what substitutions would lead to what type of crystal structures and whether the substitutions enhance or degrade charge transport. This would facilitate the design process by rational design of the length, type, and positions of the side groups. Moreover, theoretical prediction and characterization can provide insights into the intrinsic limit of a material. The knowledge of whether the intrinsic limit of a semiconductor is reached and what type of changes in the structure can lead to improvement has the potential to facilitate the OSC design tremendously. In its most useful form, the crystal structure prediction can provide the thermodynamically stable structures a priori to synthesis, calculate the energy barriers between low lattice energy configurations, tell us what kind of laboratory conditions would lead to the better performing polymorph, or how to avoid the worse polymorphs with the manipulation of process conditions.

To conclude, we believe that we now have a good understanding of the structure–property relationships in OSCs and, by the adaptation of the state-of-the-art crystal structure prediction methods, there is a great potential to help explore new high performance OSCs for OFET applications. If crystal structure prediction for OSCs can be realized, the same tools can also promote discoveries of new organic photovoltaics, ferrroelectrics, or organic electronics in general.

Acknowledgements We thank Semion Saikin and Stéphanie Valleau for stimulating discussions and reading the manuscript. We acknowledge computing facilities at the High Performance Technical Center at the Faculty of Art and Science of Harvard University, XSEDE/Teragrid resources supported by National Science Foundation award number OCI-1053575, and software support from ChemAxon Ltd.

References

1. Schwoerer M, Wolf HC (2007) Organic molecular solids. Wiley-VCH, Weinheim
2. Katz HE (2004) Recent advances in semiconductor performance and printing processes for organic transistor-based electronics. Chem Mater 16(23):4748–4756. doi:10.1021/cm049781j
3. Tsumura A, Koezuka H, Ando T (1986) Macromolecular electronic device: field-effect transistor with a polythiophene thin film. Appl Phys Lett 49(18):1210–1212
4. Feng X, Marcon V, Pisula W, Hansen MR, Kirkpatrick J, Grozema F, Andrienko D, Kremer K, Müllen K (2009) Towards high charge-carrier mobilities by rational design of the shape and periphery of discotics. Nat Mater 8(5):421–426
5. Blouin N, Michaud A, Gendron D, Wakim S, Blair E, Neagu-Plesu R, Belletête M, Durocher G, Tao Y, Leclerc M (2007) Toward a rational design of poly(2,7-carbazole) derivatives for solar cells. J Am Chem Soc 130(2):732–742. doi:10.1021/ja0771989

6. Sokolov A, Atahan-Evrenk S, Mondal R, Akkerman HB, Sanchez-Carrera RS, Granados-Focil S, Schrier J, Mannsfeld SCB, Zoombelt AP, Bao Z, Aspuru-Guzik A (2011) From computational discovery to experimental characterization of a high hole mobility organic crystal. Nat Commun 2:437
7. Jun L, Yan Z, Huei Shuan T, Yunlong G, Chong-An D, Gui Y, Yunqi L, Ming L, Suo Hon L, Yuhua Z, Haibin S, Beng SO (2012) A stable solution-processed polymer semiconductor with record high-mobility for printed transistors. Sci Rep 2:754. doi:10.1038/srep00754
8. Marien H, Steyeart M, Heremans P (2013) Analog organic electronics, building blocks for organic smart sensor systems on foil. Analog circuits and signal processing. Springer, New York
9. Gelinck G, Heremans P, Nomoto K, Anthopoulos TD (2010) Organic transistors in optical displays and microelectronic applications. Adv Mater 22(34):3778–3798. doi:10.1002/adma.200903559
10. Heeger AJ (2010) Semiconducting polymers: the third generation. Chem Soc Rev 39(7):2354–2371. doi:10.1039/b914956m
11. Zschieschang U, Yamamoto T, Takimiya K, Kuwabara H, Ikeda M, Sekitani T, Someya T, Klauk H (2011) Organic electronics on banknotes. Adv Mater 23(5):654–658. doi:10.1002/adma.201003374
12. Olivares-Amaya R, Amador-Bedolla C, Hachmann J, Atahan-Evrenk S, Sanchez-Carrera RS, Vogt L, Aspuru-Guzik A (2011) Accelerated computational discovery of high-performance materials for organic photovoltaics by means of cheminformatics. Energy Environ Sci 4:4849–4861
13. Hachmann J, Olivares-Amaya R, Atahan-Evrenk S, Amador-Bedolla C, Sanchez-Carrera RS, Gold-Parker A, Vogt L, Brockway AM, Aspuru-Guzik A (2011) The Harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid. J Phys Chem Lett 2(17):2241–2251
14. Hachmann J, Olivares-Amaya R, Jinich A, Appleton AL, Blood-Forsythe MA, Seress LR, Roman-Salgado C, Trepte K, Atahan-Evrenk S, Er S, Shrestha S, Mondal R, Sokolov A, Bao Z, Aspuru-Guzik A (2013) Lead candidates for high-performance organic photovoltaics from high-throughput quantum chemistry – the Harvard clean energy project. Energy Environ Sci 7:698–704
15. Mei J, Diao Y, Appleton AL, Fang L, Bao Z (2013) Integrated materials design of organic semiconductors for field-effect transistors. J Am Chem Soc 135(18):6724–6746. doi:10.1021/ja400881n
16. Kanal IY, Owens SG, Bechtel JS, Hutchison GR (2013) Efficient computational screening of organic polymer photovoltaics. J Phys Chem Lett 4(10):1613–1623. doi:10.1021/jz400215j
17. O'Boyle NM, Campbell CM, Hutchison GR (2011) Computational design and selection of optimal organic photovoltaic materials. J Phys Chem C 115(32):16200–16210. doi:10.1021/jp202765c
18. Curtarolo S, Hart GLW, Nardelli MB, Mingo N, Sanvito S, Levy O (2013) The high-throughput highway to computational materials design. Nat Mater 12(3):191–201
19. Clancy P (2012) Chemical engineering in the electronics industry: progress towards the rational design of organic semiconductor heterojunctions. Curr Opin Chem Eng 1(2):117–122, 10.1016/j.coche.2012.01.001
20. Holliday S, Donaghey JE, McCulloch I (2013) Advances in charge carrier mobilities of semiconducting polymers used in organic transistors. Chem Mater 26:647–663. doi:10.1021/cm402421p
21. Wang C, Dong H, Hu W, Liu Y, Zhu D (2011) Semiconducting π-conjugated systems in field-effect transistors: a material odyssey of organic electronics. Chem Rev 112(4):2208–2267. doi:10.1021/cr100380z
22. Rivnay J, Mannsfeld SCB, Miller CE, Salleo A, Toney MF (2012) Quantitative determination of organic semiconductor microstructure from the molecular to device scale. Chem Rev 112(10):5488–5519. doi:10.1021/cr3001109

23. Troisi A (2011) Charge transport in high mobility molecular semiconductors: classical models and new theories. Chem Soc Rev 40(5):2347–2358. doi:10.1039/c0cs00198h
24. Wang L, Nan G, Yang X, Peng Q, Li Q, Shuai Z (2010) Computational methods for design of organic materials with high charge mobility. Chem Soc Rev 39(2):423–434. doi:10.1039/b816406c
25. Coropceanu V, Cornil J, Da Silva Filho DA, Olivier Y, Silbey R, Bredas J-L (2007) Charge transport in organic semiconductors. Chem Rev 107:926–952
26. Datta S, Grant DJW (2004) Crystal structures of drugs: advances in determination, prediction and engineering. Nat Rev Drug Discov 3(1):42–57
27. Bauer J, Spanton S, Henry R, Quick J, Dziki W, Porter W, Morris J (2001) Ritonavir: an extraordinary example of conformational polymorphism. Pharm Res 18(6):859–866. doi:10.1023/a:1011052932607
28. Della Valle RG, Brillante A, Venuti E, Farina L, Girlando A, Masino M (2004) Exploring the polymorphism of crystalline pentacene. Org Electron 5(1–3):1–6, 10.1016/j.orgel.2003.08.017
29. Jurchescu OD, Mourey DA, Subramanian S, Parkin SR, Vogel BM, Anthony JE, Jackson TN, Gundlach DJ (2009) Effects of polymorphism on charge transport in organic semiconductors. Phys Rev B 80(8):085201
30. Mei J, Bao Z (2014) Side chain engineering in solution-processible conjugated polymers for organic solar cells and field-effect transistors. Chem Mater 26(1):604–615. doi:10.1021/cm4020805
31. Sumrak JC, Sokolov AN, Macgillivray LR (2011) Crystal engineering organic semiconductors. In: Self-organized organic semiconductors. Wiley, New York, pp 1–19. doi:10.1002/9780470949122.ch1
32. Głowacki ED, Irimia-Vladu M, Kaltenbrunner M, Gsiorowski J, White MS, Monkowius U, Romanazzi G, Suranna GP, Mastrorilli P, Sekitani T, Bauer S, Someya T, Torsi L, Sariciftci NS (2013) Hydrogen-bonded semiconducting pigments for air-stable field-effect transistors. Adv Mater 25(11):1563–1569. doi:10.1002/adma.201204039
33. Stone AJ (2008) Intermolecular potentials. Science 321(5890):787–789. doi:10.1126/science.1158006
34. Klimes J, Michaelides A (2012) Perspective: advances and challenges in treating van der Waals dispersion forces in density functional theory. J Chem Phys 137(12):120901
35. Hongo K, Watson MA, Sánchez-Carrera RS, Iitaka T, Aspuru-Guzik A (2010) Failure of conventional density functionals for the prediction of molecular crystal polymorphism: a quantum Monte Carlo study. J Phys Chem Lett 1(12):1789–1794. doi:10.1021/jz100418p
36. Reilly AM, Tkatchenko A (2013) Seamless and accurate modeling of organic molecular materials. J Phys Chem Lett 4(6):1028–1033. doi:10.1021/jz400226x
37. Bardwell DA, Adjiman CS, Arnautova YA, Bartashevich E, Boerrigter SXM, Braun DE, Cruz-Cabeza AJ, Day GM, Della Valle RG, Desiraju GR, van Eijck BP, Facelli JC, Ferraro MB, Grillo D, Habgood M, Hofmann DWM, Hofmann F, Jose KVJ, Karamertzanis PG, Kazantsev AV, Kendrick J, Kuleshova LN, Leusen FJJ, Maleev AV, Misquitta AJ, Mohamed S, Needs RJ, Neumann MA, Nikylov D, Orendt AM, Pal R, Pantelides CC, Pickard CJ, Price LS, Price SL, Scheraga HA, van de Streek J, Thakur TS, Tiwari S, Venuti E, Zhitkov IK (2011) Towards crystal structure prediction of complex organic compounds – a report on the fifth blind test. Acta Crystallogr B 67(6):535–551. doi:10.1107/S0108768111042868
38. Day GM, Cooper TG, Cruz-Cabeza AJ, Hejczyk KE, Ammon HL, Boerrigter SXM, Tan JS, Della Valle RG, Venuti E, Jose J, Gadre SR, Desiraju GR, Thakur TS, van Eijck BP, Facelli JC, Bazterra VE, Ferraro MB, Hofmann DWM, Neumann MA, Leusen FJJ, Kendrick J, Price SL, Misquitta AJ, Karamertzanis PG, Welch GWA, Scheraga HA, Arnautova YA, Schmidt MU, van de Streek J, Wolf AK, Schweizer B (2009) Significant progress in predicting the crystal structures of small organic molecules - a report on the fourth blind test. Acta Crystallogr B 65(2):107–125. doi:10.1107/S0108768109004066

39. Neumann MA (2008) Tailor-made force fields for crystal-structure prediction. J Phys Chem B 112(32):9810–9829. doi:10.1021/jp710575h
40. Kazantsev AV, Karamertzanis PG, Adjiman CS, Pantelides CC (2011) Efficient handling of molecular flexibility in lattice energy minimization of organic crystals. J Chem Theory Comput 7(6):1998–2016. doi:10.1021/ct100597e
41. Neumann MA, Perrin M-A (2005) Energy ranking of molecular crystals using density functional theory calculations and an empirical van der Waals correction. J Phys Chem B 109(32):15531–15541. doi:10.1021/jp050121r
42. Price SL, Leslie M, Welch GWA, Habgood M, Price LS, Karamertzanis PG, Day GM (2010) Modelling organic crystal structures using distributed multipole and polarizability-based model intermolecular potentials. Phys Chem Chem Phys 12(30):8478–8490. doi:10.1039/c004164e
43. Silinsh EA, Capek V (1994) Organic molecular crystals: interaction, localization, and transport phenomena. American Institute of Physics, New York
44. Ortmann F, Bechstedt F, Hannewald K (2011) Charge transport in organic crystals: theory and modelling. Phys Status Solidi B 248(3):511–525. doi:10.1002/pssb.201046278
45. Cheng YC, Silbey RJ (2008) A unified theory for charge-carrier transport in organic crystals. J Chem Phys 128(11):114713. doi:10.1063.1.28948.0
46. Bao Z, Locklin JJ (2007) Organic field-effect transistors. CRC, Boca Raton
47. Madru M, Guillaud G, Sadoun MA, Maitrot M, Clarisse C, Contellec ML, André JJ, Simon J (1987) The first field effect transistor based on an intrinsic molecular semiconductor. Chem Phys Lett 142(1–2):103–105, 10.1016/0009-2614(87)87259-7
48. Sakanoue T, Sirringhaus H (2010) Band-like temperature dependence of mobility in a solution-processed organic semiconductor. Nat Mater 9(9):736–740. doi:10.1038/nmat2825
49. Takamiya M, Sekitani T, Ishida K, Someya T, Sakurai T (2013) Large area electronics with organic transistors. In: Cantatore E (ed) Applications of organic and printed electronics. Integrated circuits and systems. Springer, New York, pp 101–113. doi:10.1007/978-1-4614-3160-2_5
50. Katz HE, Bao Z (1999) The physical chemistry of organic field-effect transistors. J Phys Chem B 104(4):671–678. doi:10.1021/jp992853n
51. Horowitz G (2006) Organic transistors. In: Klauk H (ed) Organic electronics, materials, manufacturing and applications. Wiley-VCH, Weinheim, pp 3–32
52. Ito Y, Virkar AA, Mannsfeld S, Oh JH, Toney M, Locklin J, Bao Z (2009) Crystalline ultrasmooth self-assembled monolayers of alkylsilanes for organic field-effect transistors. J Am Chem Soc 131(26):9396–9404. doi:10.1021/ja9029957
53. Ukah NB, Granstrom J, Gari RRS, King GM, Guha S (2011) Low-operating voltage and stable organic field-effect transistors with poly (methyl methacrylate) gate dielectric solution deposited from a high dipole moment solvent. Appl Phys Lett 99(24):243302–243303
54. Zschieschang U, Kang MJ, Takimiya K, Sekitani T, Someya T, Canzler TW, Werner A, Blochwitz-Nimoth J, Klauk H (2012) Flexible low-voltage organic thin-film transistors and circuits based on C10-DNTT. J Mater Chem 22(10):4273–4277. doi:10.1039/c1jm14917b
55. Horowitz G (1998) Organic field-effect transistors. Adv Mater 10(5):365–377
56. Anthony JE, Brooks JS, Eaton DL, Parkin SR (2001) Functionalized pentacene: improved electronic properties from control of solid-state order. J Am Chem Soc 123:9482–9483. doi:10.1021/ja0162459
57. Yang YS, Yasuda T, Kakizoe H, Mieno H, Kino H, Tateyama Y, Adachi C (2013) High performance organic field-effect transistors based on single-crystal microribbons and microsheets of solution-processed dithieno[3,2-b:2′,3′-d]thiophene derivatives. Chem Commun 49(58):6483–6485. doi:10.1039/c3cc42114g
58. Podzorov V, Menard E, Borissov A, Kiryukhin V, Rogers JA, Gershenson ME (2004) Intrinsic charge transport on the surface of organic semiconductors. Phys Rev Lett 93(8):086602

59. Diao Y, Tee BCK, Giri G, Xu J, Kim DH, Becerril HA, Stoltenberg RM, Lee TH, Xue G, Mannsfeld SCB, Bao Z (2013) Solution coating of large-area organic semiconductor thin films with aligned single-crystalline domains. Nat Mater 12(7):665–671. doi:10.1038/nmat3650, http://www.nature.com/nmat/journal/v12/n7/abs/nmat3650.html - supplementary-information
60. Leclerc M, Morin J-F (eds) (2010) Design and synthesis of conjugated polymers. WILEY-VCH Verlag GmbH & Co.KGaA, Weinheim
61. Li J, Zhao Y, Tan HS, Guo Y, Di C-A, Yu G, Liu Y, Lin M, Lim SH, Zhou Y, Su H, Ong BS (2012) A stable solution-processed polymer semiconductor with record high-mobility for printed transistors. Sci Rep 2:754, http://www.nature.com/srep/2012/121018/srep00754/abs/srep00754.html - supplementary-information
62. Noriega R, Rivnay J, Vandewal K, Koch FPV, Stingelin N, Smith P, Toney MF, Salleo A (2013) A general relationship between disorder, aggregation and charge transport in conjugated polymers. Nat Mater 12(11):1038–1044
63. Bassler H, Kohler A (eds) (2012) Charge transport in organic semiconductors, vol 312. Top Curr Chem. Springer, Berlin
64. Brédas JL, Beljonne D, Coropceanu V, Cornil J (2004) Charge-transfer and energy-transfer processes in pi-conjugated oligomers and polymers: a molecular picture. Chem Rev 104 (11):4971–5003. doi:10.1021/Cr040084k
65. Ostroverkhova O, Cooke DG, Shcherbyna S, Egerton RF, Hegmann FA, Tykwinski RR, Anthony JE (2005) Bandlike transport in pentacene and functionalized pentacene thin films revealed by subpicosecond transient photoconductivity measurements. Phys Rev B 71(3): 035204
66. Karl N (2003) Charge carrier transport in organic semiconductors. Synth Met 133–134:649–657, 10.1016/S0379-6779(02)00398-3
67. Brédas JL, Calbert JP, da Silva Filho DA, Cornil J (2002) Organic semiconductors: a theoretical characterization of the basic parameters governing charge transport. Proc Natl Acad Sci U S A 99(9):5804–5809. doi:10.1073/pnas.092143399
68. Hatch RC, Huber DL, Höchst H (2009) HOMO band structure and anisotropic effective hole mass in thin crystalline pentacene films. Phys Rev B 80(8):081411
69. Ruhle V, Kirkpatrick J, Andrienko D (2010) A multiscale description of charge transport in conjugated oligomers. J Chem Phys 132(13):134103
70. Norton JE, Brédas JL (2008) Theoretical characterization of titanyl phthalocyanine as a p-type organic semiconductor: short intermolecular pi–pi interactions yield large electronic couplings and hole transport bandwidths. J Chem Phys 128(3):034701. doi:10.1063.1.28068.3
71. Senthilkumar K, Grozema FC, Bickelhaupt FM, Siebbeles LDA (2003) Charge transport in columnar stacked triphenylenes: effects of conformational fluctuations on charge transfer integrals and site energies. J Chem Phys 119(18):9809–9817
72. Kirkpatrick J (2008) An approximate method for calculating transfer integrals based on the ZINDO Hamiltonian. Int J Quantum Chem 108(1):51–56. doi:10.1002/qua.21378
73. Hannewald K, Stojanovic VM, Schellekens JMT, Bobbert PA, Kresse G, Hafner J (2004) Theory of polaron bandwidth narrowing in organic molecular crystals. Phys Rev B 69(7): 075211
74. Ferretti A, Ruini A, Molinari E, Caldas MJ (2003) Electronic properties of polymer crystals: the effect of interchain interactions. Phys Rev Lett 90(8):086401
75. Huang JS, Kertesz M (2004) Intermolecular transfer integrals for organic molecular materials: can basis set convergence be achieved? Chem Phys Lett 390(1–3):110–115. doi:10.1016/j.cplett.2004.03.141
76. Mikołajczyk M, Zaleśny R, Czyżnikowska Ż, Toman P, Leszczynski J, Bartkowiak W (2011) Long-range corrected DFT calculations of charge-transfer integrals in model metal-free phthalocyanine complexes. J Mol Model 17(9):2143–2149. doi:10.1007/s00894-010-0865-7
77. Sancho-Garcia JC, Horowitz G, Bredas JL, Cornil J (2003) Effect of an external electric field on the charge transport parameters in organic molecular semiconductors. J Chem Phys 119 (23):12563–12568

78. Kojima H, Mori T (2011) Dihedral angle dependence of transfer integrals in organic semiconductors with herringbone structures. Bull Chem Soc Jpn 84(10):1049–1056
79. Lee JY, Roth S, Park YW (2006) Anisotropic field effect mobility in single crystal pentacene. Appl Phys Lett 88(25):252106
80. Haddon RC, Siegrist T, Fleming RM, Bridenbaugh PM, Laudise RA (1995) Band structures of organic thin-film-transistor materials. J Mater Chem 5(10):1719–1724
81. Huang JS, Kertesz M (2005) Validation of intermolecular transfer integral and bandwidth calculations for organic molecular materials. J Chem Phys 122(23):234707. doi:10.1063.1.19256.1
82. Hotta C (2003) Classification of quasi-two dimensional organic conductors based on a new minimal model. J Phys Soc Jpn 72:840
83. Mori T, Mori H, Tanaka S (1999) Structural genealogy of BEDT-TTF-based organic conductors II. Inclined molecules: theta, alpha, and kappa phases. Bull Chem Soc Jpn 72 (2):179–197
84. Vehoff T, Baumeier B, Troisi A, Andrienko D (2010) Charge transport in organic crystals: role of disorder and topological connectivity. J Am Chem Soc 132(33):11702–11708. doi:10.1021/ja104380c
85. Marcus RA (1993) Electron transfer reactions in chemistry. Theory and experiment. Rev Mod Phys 65(3):599–610
86. Reimers JR (2001) A practical method for the use of curvilinear coordinates in calculations of normal-mode-projected displacements and Duschinsky rotation matrices for large molecules. J Chem Phys 115(20):9103–9109
87. McMahon DP, Troisi A (2010) Evaluation of the external reorganization energy of polyacenes. J Phys Chem Lett 1(6):941–946. doi:10.1021/jz1001049
88. Norton JE, Brédas JL (2008) Polarization energies in oligoacene semiconductor crystals. J Am Chem Soc 130(37):12377–12384. doi:10.1021/Ja8017797
89. Duhm S, Xin Q, Hosoumi S, Fukagawa H, Sato K, Ueno N, Kera S (2012) Charge reorganization energy and small polaron binding energy of rubrene thin films by ultraviolet photoelectron spectroscopy. Adv Mater 24(7):901–905. doi:10.1002/adma.201103262
90. Kera S, Hosoumi S, Sato K, Fukagawa H, Nagamatsu S-I, Sakamoto Y, Suzuki T, Huang H, Chen W, Wee ATS, Coropceanu V, Ueno N (2013) Experimental reorganization energies of pentacene and perfluoropentacene: effects of perfluorination. J Phys Chem C 117(43): 22428–22437. doi:10.1021/jp4032089
91. da Silva Filho DA, Coropceanu V, Fichou D, Gruhn NE, Bill TG, Gierschner J, Cornil J, Brédas JL (2007) Hole-vibronic coupling in oligothiophenes: impact of backbone torsional flexibility on relaxation energies. Philos Trans R Soc A 365(1855):1435–1452. doi:10.1098/rsta.2007.2025
92. Martinelli NG, Olivier Y, Athanasopoulos S, Ruiz-Delgado MC, Pigg KR, da Silva DA, Sánchez-Carrera RS, Venuti E, Della Valle RG, Brédas JL, Beljonne D, Cornil J (2009) Influence of intermolecular vibrations on the electronic coupling in organic semiconductors: the case of anthracene and perfluoropentacene. ChemPhysChem 10(13):2265–2273. doi:10.1002/cphc.200900298
93. Nan G, Li Z (2012) Influence of lattice dynamics on charge transport in the dianthra [2,3-b:2′,3′-f]-thieno[3,2-b]thiophene organic crystals from a theoretical study. Phys Chem Chem Phys 14(26):9451–9459. doi:10.1039/c2cp40857k
94. Troisi A, Orlandi G (2006) Dynamics of the intermolecular transfer integral in crystalline organic semiconductors. J Phys Chem A 110(11):4065–4070. doi:10.1021/Jp055432g
95. Sánchez-Carrera RS, Atahan S, Schrier J, Aspuru-Guzik A (2010) Theoretical characterization of the air-stable, high-mobility dinaphtho[2,3-b:2′,3′-f]thieno[3,2-b]-thiophene organic semiconductor. J Phys Chem C 114(5):2334–2340. doi:10.1021/jp910102f
96. Sánchez-Carrera RS, Paramonov P, Day GM, Coropceanu V, Brédas J-L (2010) Interaction of charge carriers with lattice vibrations in oligoacene crystals from naphthalene to pentacene. J Am Chem Soc 132(41):14437–14446. doi:10.1021/ja1040732

97. Coropceanu V, Sánchez-Carrera RS, Paramonov P, Day GM, Brédas JL (2009) Interaction of charge carriers with lattice vibrations in organic molecular semiconductors: naphthalene as a case study. J Phys Chem C 113(11):4679–4686. doi:10.1021/Jp900157p
98. Davydov SA (1962) Theory of molecular excitons (trans: M. K, M. O). McGraw-Hill, New York
99. Pope M, Swenberg CE (1999) Electronic processes in organic crystals and polymers, 2nd edn. Oxford University Press, New York
100. Takimiya K, Shinamura S, Osaka I, Miyazaki E (2011) Thienoacene-based organic semiconductors. Adv Mater 23(38):4347–4370. doi:10.1002/adma.201102007
101. Salzmann I, Duhm S, Heimel G, Oehzelt M, Kniprath R, Johnson RL, JrP R, Koch N (2008) Tuning the ionization energy of organic semiconductor films: the role of intramolecular polar bonds. J Am Chem Soc 130(39):12870–12871. doi:10.1021/ja804793a
102. Kazuo T, Tatsuya Y, Hideaki E, Takafumi I (2007) Design strategy for air-stable organic semiconductors applicable to high-performance field-effect transistors. Sci Technol Adv Mater 8(4):273
103. Tang ML, Reichardt AD, Wei P, Bao Z (2009) Correlating carrier type with frontier molecular orbital energy levels in organic thin film transistors of functionalized acene derivatives. J Am Chem Soc 131:5264–5273. doi:10.1021/ja809659b
104. Kobayashi H, Kobayashi N, Hosoi S, Koshitani N, Murakami D, Shirasawa R, Kudo Y, Hobara D, Tokita Y, Itabashi M (2013) Hopping and band mobilities of pentacene, rubrene, and 2,7-dioctyl[1]benzothieno[3,2-b][1]benzothiophene (C8-BTBT) from first principle calculations. J Chem Phys 139:014707
105. Watanabe M, Chang YJ, Liu S-W, Chao T-H, Goto K, IslamMd M, Yuan C-H, Tao Y-T, Shinmyozu T, Chow TJ (2012) The synthesis, crystal structure and charge-transport properties of hexacene. Nat Chem 4(7):574–578, http://www.nature.com/nchem/journal/v4/n7/abs/nchem.1381.html - supplementary-information
106. Hutchison GR, Ratner MA, Marks TJ (2005) Hopping transport in conductive heterocyclic oligomers: reorganization energies and substituent effects. J Am Chem Soc 127(7):2339–2350. doi:10.1021/ja0461421
107. Halik M, Klauk H, Zschieschang U, Schmid G, Ponomarenko S, Kirchmeyer S, Weber W (2003) Relationship between molecular structure and electrical performance of oligothiophene organic thin film transistors. Adv Mater 15(11):917–922. doi:10.1002/adma.200304654
108. Misra M, Andrienko D, Br B, Faulon J-L, von Lilienfeld OA (2011) Toward quantitative structure–property relationships for charge transfer rates of polycyclic aromatic hydrocarbons. J Chem Theory Comput 7(8):2549–2555. doi:10.1021/ct200231z
109. Faulon J-L (1994) Stochastic generator of chemical structure. 1. Application to the structure elucidation of large molecules. J Chem Inf Comput Sci 34(5):1204–1218. doi:10.1021/ci00021a031
110. Kuo M-Y, Liu C-C (2009) Molecular design toward high hole mobility organic semiconductors: tetraceno[2,3-c]thiophene derivatives of ultrasmall reorganization energies. J Phys Chem C 113(37):16303–16306. doi:10.1021/jp9065423
111. Kwon O, Coropceanu V, Gruhn NE, Durivage JC, Laquindanum JG, Katz HE, Cornil J, Bredas JL (2004) Characterization of the molecular parameters determining charge transport in anthradithiophene. J Chem Phys 120:8186–8194. doi:10.1063.1.16896.6
112. Kuo M-Y, Chen H-Y, Chao I (2007) Cyanation: providing a three-in-one advantage for the design of n-type organic field-effect transistors. Chemistry 13(17):4750–4758. doi:10.1002/chem.200601803
113. Chen H-Y, Chao I (2006) Toward the rational design of functionalized pentacenes: reduction of the impact of functionalization on the reorganization energy. ChemPhysChem 7(9):2003–2007. doi:10.1002/cphc.200600266

114. Sancho-Garcia JC, Perez-Jimenez AJ, Olivier Y, Cornil J (2010) Molecular packing and charge transport parameters in crystalline organic semiconductors from first-principles calculations. Phys Chem Chem Phys 12(32):9381–9388. doi:10.1039/b925652k
115. Pola S, Kuo C-H, Peng W-T, Islam MM, Chao I, Tao Y-T (2012) Contorted tetrabenzocoronene derivatives for single crystal field effect transistors: correlation between packing and mobility. Chem Mater 24(13):2566–2571. doi:10.1021/cm301190c
116. Desiraju GR, Gavezzotti A (1989) Crystal structures of polynuclear aromatic hydrocarbons. Classification, rationalization and prediction from molecular structure. Acta Crystallogr B 45 (5):473–482. doi:10.1107/S0108768189003794
117. Glowacki ED, Leonat L, Irimia-Vladu M, Schwodiauer R, Ullah M, Sitter H, Bauer S, Sariciftci NS (2012) Intermolecular hydrogen-bonded organic semiconductors–Quinacridone versus pentacene. Appl Phys Lett 101(2):023304–023305
118. da Silva Filho DA, Kim EG, Brédas JL (2005) Transport properties in the rubrene crystal: electronic coupling and vibrational reorganization energy. Adv Mater 17(8):1072–1076. doi:10.1002/adma.200401866
119. Haas S, Stassen AF, Schuck G, Pernstich KP, Gundlach DJ, Batlogg B, Berens U, Kirner HJ (2007) High charge-carrier mobility and low trap density in a rubrene derivative. Phys Rev B 76(11):115203
120. McGarry KA, Xie W, Sutton C, Risko C, Wu Y, Young VG, Brédas J-L, Frisbie CD, Douglas CJ (2013) Rubrene-based single-crystal organic semiconductors: synthesis, electronic structure, and charge-transport properties. Chem Mater 25(11):2254–2263. doi:10.1021/cm400736s
121. Anthony JE (2006) Engineered pentacenes. In: Klauk H (ed) Organic electronics: materials, manufacturing and applications. Wiley-VCH, Weinheim, FRG
122. Anthony JE (2008) The larger acenes: versatile organic semiconductors. Angew Chem Int Ed 47(3):452–483
123. Giri G, Verploegen E, Mannsfeld SCB, Atahan-Evrenk S, Kim DH, Lee SY, Becerril HA, Aspuru-Guzik A, Toney MF, Bao Z (2011) Tuning charge transport in solution-sheared organic semiconductors using lattice strain. Nature 480:504–508
124. Zhang L, Fakhouri SM, Liu F, Timmons JC, Ran NA, Briseno AL (2011) Chalcogenoarene semiconductors: new ideas from old materials. J Mater Chem 21(5):1329–1337. doi:10.1039/c0jm02522d
125. Tucker NM, Briseno AL, Acton O, Yip H-L, Ma H, Jenekhe SA, Xia Y, Jen AKY (2013) Solvent-dispersed benzothiadiazole-tetrathiafulvalene single-crystal nanowires and their application in field-effect transistors. ACS Appl Mater Interfaces 5(7):2320–2324. doi:10.1021/am3025036
126. Anthony JE (2007) Induced pi-stacking in acenes. In: Muller TJJ, Bunz UHF (eds) Functional organic materials. Wiley-VCH, Weinheim, p 511
127. Curtis MD, Cao J, Kampf JW (2004) Solid-state packing of conjugated oligomers: from π-stacks to the Herringbone structure. J Am Chem Soc 126(13):4318–4328. doi:10.1021/ja0397916
128. Kang MJ, Yamamoto T, Shinamura S, Miyazaki E, Takimiya K (2010) Unique three-dimensional (3D) molecular array in dimethyl-DNTT crystals: a new approach to 3D organic semiconductors. Chem Sci 1(2):179–183. doi:10.1039/c0sc00156b
129. Reese C, Roberts ME, Parkin SR, Bao Z (2009) Tuning crystalline solid-state order and charge transport via building-block modification of oligothiophenes. Adv Mater 21(36): 3678–3681. doi:10.1002/adma.200900836
130. Akkerman HB, Mannsfeld S, Kaushik A, Verploegen E, Burnier L, Zoombelt A, Saathoff J, Hong S, Atahan-Evrenk S, Liu X, Aspuru-Guzik A, Toney M, Clancy P, Bao Z (2013) Effects of odd-even side chain length of alkyl-substituted diphenyl-bithiophenes on first monolayer thin film packing structure. J Am Chem Soc 135(30):11006–11014

131. Liu J, Zhang Y, Phan H, Sharenko A, Moonsin P, Walker B, Promarak V, Nguyen T-Q (2013) Effects of stereoisomerism on the crystallization behavior and optoelectrical properties of conjugated molecules. Adv Mater 25(27):3645–3650. doi:10.1002/adma.201300255
132. Troisi A, Orlandi G (2005) Band structure of the four pentacene polymorphs and effect on the hole mobility at low temperature. J Phys Chem B 109(5):1849–1856. doi:10.1021/jp0457489
133. Bussac MN, Picon JD, Zuppiroli L (2004) The impact of molecular polarization on the electronic properties of molecular semiconductors. Europhys Lett 66(3):392
134. Topham BJ, Soos ZG (2011) Ionization in organic thin films: electrostatic potential, electronic polarization, and dopants in pentacene films. Phys Rev B 84(16):165405
135. Minder NA, Ono S, Chen Z, Facchetti A, Morpurgo AF (2012) Band-like electron transport in organic transistors and implication of the molecular structure for performance optimization. Adv Mater 24(4):503–508. doi:10.1002/adma.201103960
136. Chang Y-f L, Z-y AL-j, J-p Z (2011) From molecules to materials: molecular and crystal engineering design of organic optoelectronic functional materials for high carrier mobility. J Phys Chem C 116(1):1195–1199. doi:10.1021/jp208063h
137. Reck G, Schulz BW (2006) Benzo(e)pyrene (CSD-CEQGEL)
138. Marcon V, Raos G (2004) Molecular modeling of crystalline oligothiophenes: testing and development of improved force fields. J Phys Chem B 108(46):18053–18064. doi:10.1021/Jp047128d
139. Della Valle RG, Venuti E, Brillante A, Girlando A (2006) Inherent structures of crystalline tetracene. J Phys Chem A 110(37):10858–10862. doi:10.1021/jp0611020
140. Venuti E, Della Valle RG, Brillante A, Masino M, Girlando A (2002) Probing pentacene polymorphs by lattice dynamics calculations. J Am Chem Soc 124(10):2128–2129. doi:10.1021/ja0166949
141. Della Valle RG, Venuti E, Brillante A, Girlando A (2003) Inherent structures of crystalline pentacene. J Chem Phys 118(2):807–815
142. Della Valle RG, Venuti E, Brillante A, Girlando A (2008) Are crystal polymorphs predictable? The case of sexithiophene. J Phys Chem A 112(29):6715–6722. doi:10.1021/jp801749n
143. Williams DE, Starr TL (1977) J Comput Chem 1:13
144. Spek AL (2003) Platon. J Appl Cryst 36:7
145. Woodley SM, Catlow R (2008) Crystal structure prediction from first principles. Nat Mater 7(12):937–946
146. Bazterra VE, Ferraro MB, Facelli JC (2002) Modified genetic algorithm to model crystal structures. I. Benzene, naphthalene and anthracene. J Chem Phys 116(14):5984–5991
147. Kim S, Orendt AM, Ferraro MB, Facelli JC (2009) Crystal structure prediction of flexible molecules using parallel genetic algorithms with a standard force field. J Comput Chem 30 (13):1973–1985. doi:10.1002/jcc.21189
148. Baur WH, Kassner D (1992) The perils of Cc: comparing the frequencies of falsely assigned space groups with their general population. Acta Crystallogr B 48(4):356–369. doi:10.1107/S0108768191014726
149. Padmaja N, Ramakumar S, Viswamitra MA (1990) Space-group frequencies of proteins and of organic compounds with more than one formula unit in the asymmetric unit. Acta Crystallogr A 46(9):725–730. doi:10.1107/S0108767390004512
150. Price S (2013) Why don't we find more polymorphs? Acta Crystallogr B 69(4):313–328. doi:10.1107/S2052519213018861
151. Mellot-Draznieks C (2007) Role of computer simulations in structure prediction and structure determination: from molecular compounds to hybrid frameworks. J Mater Chem 17(41): 4348–4358. doi:10.1039/b702516p
152. Sanchez-Carrera RS, Atahan-Evrenk S, Schrier J, Aspuru-Guzik A (2010) Theoretical characterization of the air-stable, high-mobility dinaphtho[2,3-b:2′3′-f]thieno[3,2-b]-thiophene organic semiconductor. J Phys Chem C 114(5):2334–2340

153. Uno M, Tominari Y, Yamagishi M, Doi I, Miyazaki E, Takimiya K, Takeya J (2009) Moderately anisotropic field-effect mobility in dinaphtho[2,3-b:2(′),3(′)-f]thiopheno[3,2-b] thiophenes single-crystal transistors. Appl Phys Lett 94(22):223308. doi:10.1063.1.31531.9
154. Accelrys (2006) Materials studio
155. Clancy P (2011) Application of molecular simulation techniques to the study of factors affecting the thin-film morphology of small-molecule organic semiconductors. Chem Mater 23(3):522–543. doi:10.1021/cm102231b
156. Minemawari H, Yamada T, Matsui H, Tsutsumi JY, Haas S, Chiba R, Kumai R, Hasegawa T (2011) Inkjet printing of single-crystal films. Nature 475(7356):364–367, http://www.nature.com/nature/journal/v475/n7356/abs/nature10313.html - supplementary-information
157. Ebata H, Izawa T, Miyazaki E, Takimiya K, Ikeda M, Kuwabara H, Yui T (2007) Highly soluble [1]Benzothieno[3,2-b]benzothiophene (BTBT) derivatives for high-performance, solution-processed organic field-effect transistors. J Am Chem Soc 129(51):15732–15733

Top Curr Chem (2014) 345: 139–180
DOI: 10.1007/128_2013_486
© Springer-Verlag Berlin Heidelberg 2013
Published online: 28 November 2013

Data Mining Approaches to High-Throughput Crystal Structure and Compound Prediction

Geoffroy Hautier

Abstract Predicting unknown inorganic compounds and their crystal structure is a critical step of high-throughput computational materials design and discovery. One way to achieve efficient compound prediction is to use data mining or machine learning methods. In this chapter we present a few algorithms for data mining compound prediction and their applications to different materials discovery problems. In particular, the patterns or correlations governing phase stability for experimental or computational inorganic compound databases are statistically learned and used to build probabilistic or regression models to identify novel compounds and their crystal structures. The stability of those compound candidates is then assessed using ab initio techniques. Finally, we report a few cases where data mining driven computational predictions were experimentally confirmed through inorganic synthesis.

Keywords Ab initio computations · Crystal structure prediction · Data mining · High-throughput computing

Contents

G. Hautier (✉)
Université Catholique de Louvain, Institute of Condensed Matter and Nanosciences (IMCN), Chemin des étoiles 8, bte L7.03.01, 1348 Louvain-la-Neuve, Belgium
e-mail: geoffroy.hautier@uclouvain.be

1 Introduction

First principles or ab initio computations aim at computing materials properties (e.g., thermodynamic stability, conductivity, light absorbance) from the fundamental laws of quantum physics. Following the emergence of ab initio techniques and especially of density functional theory (DFT) [1], the field has seen a combination of theoretical developments, standard codes developments (e.g., [2–4]), and increase in computational power. Materials science is even moving to a new paradigm where computations are not only used to explain experimental observations but also to *predict* new materials and their properties [5].

One emerging route towards computational discovery of materials is to use high-throughput computing. High-throughput computing consists of evaluating material properties on thousands of different materials to identify the best performing compounds and to understand trends from large datasets [5, 6]. This approach has already been used in various fields such as catalysis [7], Li-ion batteries [8, 9], scintillators [10], photocatalytic water splitters [11–13], thermoelectric materials [14, 15], mercury sorbents [16], organic photovoltaics [17], and topological insulators [18]. High-throughput infrastructures have reached such a maturity that large sets of computations are nowadays stored in computational databases such as

the materials project [19, 20] and others [21, 22] that can be accessed through web interfaces. With the data repository and analysis tools they provide, materials scientists now have access to an unprecedented amount of data [23].

Many high-throughput studies have concentrated on evaluating properties on known compounds extracted from databases such as the Inorganic Crystal Structure Database (ICSD) [24] or on a limited structural framework (e.g., perovskites [11]). While those studies are of great value, they face some limitations. Databases are often not up to date, i.e., they do not have the latest reported structures in the literature. Also, many inorganic compounds are known to exist at a given stoichiometry but their crystal structure has not been determined from powder diffraction data. Finally, compounds of greatest interest for a specific application might not have been synthesized yet. This is especially the case for multicomponent systems (e.g., ternaries and quaternaries) or less common chemistries.

Finding new compounds and determining their crystal structure before synthesis is called crystal structure prediction. Since 1988, when Nature's editor John Maddox called our inability to properly perform crystal structure prediction one of "continuous scandal in physical science," the field has greatly evolved [25–27]. Among the different approaches to structure prediction, data mining has been developed in parallel with high-throughput computational searching. Indeed, in contrast to other approaches, data mining typically compromises on the exhaustivity of the search in favor of less computational time and an access to much larger chemical spaces to explore. The idea behind data mined compound prediction is very simple and has been driving solid state chemistry for centuries: nature is not random and there are patterns that one could learn from observing phase stability. The novelty lies in the use of quantitative mathematical approaches from the fields of machine learning or statistical learning.

In this chapter we will start by presenting how thermodynamical phase stability can be evaluated from DFT computations (Sect. 2). The different techniques and accuracy of approximations will be outlined. The general idea behind data mining driven structure prediction will be presented in Sect. 3 and specific examples of methods and algorithms will be explained in detail in Sects. 4, 5, and 6. Finally, we will present in Sect. 7 a few selected examples of successful data mining compound predictions where the computational suggestion was followed by successful experimental verification.

2 Phase Stability Evaluation Through Ab Initio Computing

An important factor determining the existence of inorganic compounds is their thermodynamical phase stability. To evaluate whether a compound is thermodynamically stable, one needs to compare its (free) energy with the (free) energy of other competing phases. This step is essential for the compound prediction problem and DFT computations are routinely used to perform such an analysis. In this section we will overview the standard thermodynamic constructions along with the different approximations involved and assess their accuracy.

2.1 Low Temperature Stability: The Convex Hull Construction

Assessing thermodynamical phase stability in a chemical system requires the comparison of the free energy of the different phases present [28, 29]. An isothermal, isobaric and closed system requires the use of the Gibbs free energy as thermodynamic potential. For a binary component system with N_A atoms of A and N_B atoms of B, at temperature T and pressure p, the Gibbs free energy G is expressed as

$$G(N_A, N_B, T, p) = E(N_A, N_B, T, p) + pV(N_A, N_B, T, p) - TS(N_A, N_B, T, p), \quad (1)$$

where V is the volume, S the entropy, and E the energy.

The first approximation we will make is to assume that the pV term is small. This approximation is valid when only solid phases are involved in the phase equilibrium. In addition, we will work at zero temperature. No entropic effects need to be taken into account then. Entropic effects can be modeled but this would require a more important computational budget as all relevant excitations (vibrational, configurational, and electronic) would need to be considered [30–32].

Under these approximations, the relevant thermodynamic potential is the energy. The energy normalized by the total number of particles in the system ($N = N_A + N_B$) : $\overline{E}(x_A, x_B)$ and fractions instead of amounts: $x_A = \frac{N_A}{N}$ and $x_B = \frac{N_B}{N}$ will be used. The normalized energy is usually expressed in meV/atom.

Solving the Kohn–Sham equation in the DFT framework can directly provide an approximation to this energy. Ab initio computations can therefore associate an energy to any compound present in a given chemical system. In the specific case of zero temperature and negligible volume effects, phase stability can then be directly computed from a simple set of DFT ionic relaxations on all the phases of interest. Let us illustrate this with the example of a simple binary A-B chemical system. In this system, computations have been performed for compounds at a composition A_2B, AB_2, and AB in different crystal structures designated respectively by α_1, α_2, β_1, β_2, β_3, and γ. The elemental phases have also been computed and, as a convention, all energies will be expressed as formation energies from the elements. Figure 1 plots the formation energy for the different phases computed in function of the fraction of B. From this plot, a very simple construction called the *convex hull* can be performed. The construction consists of finding a convex envelope containing all the points in the plot. This envelop called the convex hull (or hull) is plotted in green in Fig. 1. The phases present on this convex hull are the most stable phases or *ground states* for the system studied. For instance, α_2 is thermodynamically unstable and will decompose to form α_1. The phase γ will decompose into two phases: α_1 and β_2 (as γ is above the tie line formed by α_1 and β_2).

This construction can be performed in any dimension and thus on multi-component systems such as ternaries, quaternaries, etc.

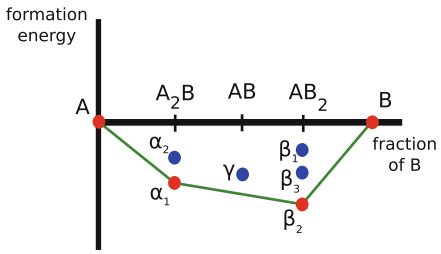


Fig. 1 Convex hull construction for an A-B system. The *points* represent different phases. The *line* is the convex hull. The *points on the line* are the most stable phases or ground states and *points above the line* are unstable phases according to the construction

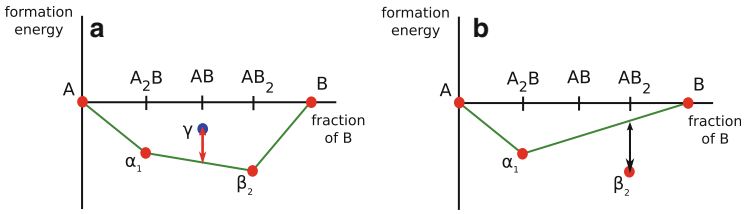


Fig. 2 Illustration of different measure of stability from the convex hull construction. The energy above the hull is illustrated for the unstable phase γ by the *double arrow* in (**a**). The inverse distance to the hull is represented for the stable phase β_2 by the *double arrow* in (**b**). Reprinted with permission from [33]. Copyright 2012 American Chemical Society

Different measures of (in)stability can be defined using this convex hull construction:

- *Energy above the hull* (*or distance to the hull*)
  For an unstable phase, the energy above the hull consists in the energy separating the phase from its decomposition tie-line (see red double arrow in Fig. 2a). It is equivalent to the opposite of the energy associated with the decomposition reaction from the phase to the stable products. It is a positive number and usually expressed in meV/atom. Stable phases have by definition an energy above the hull equals to zero.
- *Inverse energy above the hull* (*or inverse distance to the hull*)
  This quantity is defined only for stable phases. It is computed by removing the phase of interest from the convex hull and constructing a new convex hull. The distance to the new convex hull for the phase is then computed and called the inverse energy above the hull. It is equivalent to the opposite of the energy of formation of the phase of interest from the phases that would be stable if it did not exist. It is a positive number and expressed in meV/atom. A large inverse distance to hull represents a high stability of the predicted structure. The inverse energy above the hull is represented for the phase β_2 in Fig. 2b.

Convex hull constructions and the analysis of computed phase diagrams can be performed using the pymatgen package [34].

2.2 Stability for Open Systems

Oxides are very important compounds technologically and are better studied with an open instead of close thermodynamical system approach. A ternary system composed of particles of A, B, and oxygen will be used here as an example. In the previous section we assumed that the relevant thermodynamic variables are the amount of constituents (N_A, N_B, and N_O), the temperature T, and the pressure p. In reality, very often during oxide synthesis, the amount of oxygen present in the system is not directly controlled and the system is an open system to oxygen. In this case, the relevant thermodynamic potential is the Legendre transform of the Gibbs free energy with respect to the oxygen amount: the oxygen grand potential φ:

$$\varphi(N_A, N_B, \mu_O, T, p) = G - \mu_O N_O. \tag{2}$$

Normalizing the grand canonical potential by $N = N_A + N_B$ and using factions of A, B, x_A, and x_B, we get

$$\overline{\varphi}(N_A, N_B, \mu_O, T, p) = \frac{G - \mu_O N_O}{N}. \tag{3}$$

This is a situation very similar to that in the previous section except that the Gibbs free energy is replaced by the oxygen grand potential. Here, the effect of volume and temperature can be approximated by assuming that the dominant volume and entropy factors come from the gaseous oxygen and that the entropy and volume factors from the solid phase can be neglected. This approximation has been successfully used by Ong et al. for the study of the Li-Fe-P-O phase diagram [35]. The normalized grand canonical potential is then

$$\overline{\varphi}(N_A, N_B, \mu_O, T, p) = \frac{E - \mu_O N_O}{N}. \tag{4}$$

Only the μ_O term has a pressure and temperature dependence. Practically, a convex hull construction using the normalized grand canonical potential at a fixed μ_O can be performed to obtain the stable phases in specific conditions. The oxygen chemical potential can be linked to the oxidizing or reducing nature of the environment. Ways to increase the oxygen chemical potential (i.e., to be more oxidizing) are to decrease the temperature or increase the oxygen partial pressure. In contrast, the oxygen chemical potential can be decreased (i.e., be more reducing) by increasing the temperature or lowering the oxygen partial pressure.

It follows from this analysis that any oxide compound exists in an oxygen potential window with a maximal and minimal oxygen chemical potential. Any environment setting a chemical potential lower than the minimal oxygen chemical potential would be too reducing for the compound to form while any environment setting a higher chemical potential than the maximal oxygen chemical potential would be too oxidizing.

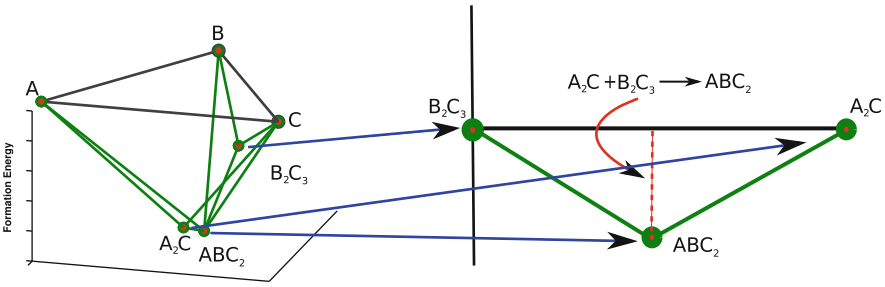


Fig. 3 Convex hull of a typical A-B-C system. The stability of the ternary ABC_2 phase will depend directly on the reaction energies from the binaries, not from the elements

2.3 Accuracy of DFT(+U) in Determining Phase Stability

Curtarolo et al. performed one of the first large scale studies of the performance of DFT on phase stability [36]. The authors focused on binary metals. They computed a large number of competing crystal structure prototypes in 80 binary metal systems and they studied how often the experimentally observed ground state was in agreement with the computed one. DFT successfully found the actual ground state in at least 90% of the cases.

For oxides and other insulators or semiconductors, the typical errors from standard DFT in oxides on the elemental formation energies can be quite large and up to hundreds of meV/atom [37]. However, for multicomponent compounds, phase stability will not depend directly on the elemental formation energy but more often on the reaction energies between multicomponent phases. Figure 3 illustrates this by presenting the convex hull of an A-B-C system. The stability of the ABC_2 phase does not depend directly on the $A + B + 2C \rightarrow ABC_2$ reaction (i.e., the formation energy from the elements) but will depend on the $A_2C + B_2C_3 \rightarrow ABC_2$ reaction (dashed red line). For instance, determining whether a ternary oxide is stable or not will depend on its reaction energy from the binary oxides. A recent study showed that those reaction energies are significantly better described by DFT than by elemental reaction energies due to cancellation of errors when comparing chemically similar phases [38]. Comparing computed to experimental reaction energies, an error distribution centered on 0 and with a standard deviation around 25 meV/atom was found. When analyzing compound prediction results, this error bar should be kept in mind.

For metal oxides with partially occupied *d* orbitals (i.e., FeO, Mn_3O_4, etc.), DFT is known to perform poorly because of a self-interaction error present in the typical functionals used in DFT. The DFT+U method is one way of circumventing this issue by effectively localizing *d* electrons and providing a more physically accurate picture of the bonding in oxides [39, 40]. On the other hand, in metals the electron delocalization induced by pure DFT is actually close to the real metallic bonding state and applying a U correction would only cause the model to deviate from the reality. We stand therefore in a situation in which, for transition metals, DFT reproduces

sufficiently well the energy in metallic systems but in oxides, only DFT+U does. As computations with two different Hamiltonians (DFT and DFT+U) cannot be directly compared, it is impossible to compute energies and then evaluate phase stability when compounds of different natures are involved, such as oxides and metals. To treat this situation, Jain et al. developed an approach relying on an energy shift of the DFT energies [41]. This shift is based on a calibration on experimental binary oxides formation energies from the metal. After applying this shift to DFT computed phases, all computed data can be compared and used to assess phase stability. A similar approach has been proposed by Stevanovic et al. [42].

3 Data Mining Compound and Crystal Structure Prediction

Section 2 showed how the phase stability of compounds is assessed using DFT. However, the most challenging part of the compound prediction problem lies in the efficient selection of compound candidates to test for stability. Nowadays this selection is typically performed following one of two approaches: optimization or data mining-based.

3.1 Optimization Approaches

Optimization-based methods consider that finding the most stable crystal structure (at a given composition) can be mapped to the mathematical problem of finding the values of the structural degrees of freedom (i.e., lattice parameters and atomic positions) minimizing the (free) energy. The search for a global minimum on the energy landscape is, however, far from simple as the energy function (or landscape) is very large, complex, and presents many local minima [43].

One popular way of simplifying this problem has been to reduce the number of degrees of freedom by working on a fixed crystal lattice, only allowing different decorations of the underlying crystal structure framework. For instance, we can study any ordering on a face-centered cubic lattice at a composition AB and find possibly a rock salt ground state. This approach is usually coupled with the use of a simplified Hamiltonian fitted on a limited set of computations performed on selected orderings through the cluster expansion technique [44–46]. Identifying new phases on a fixed lattice has been especially useful in alloy theory [47–49], but close-packed oxides have also been studied through cluster expansion [50].

However, when the underlying lattice is not known, researchers must rely on advanced optimization techniques such as simulated annealing or genetic algorithms to explore the rugged energy landscape. Simulated annealing (and the related basin hopping) [51, 52] rely on applying perturbations to a starting configuration. Those perturbation are accepted or not depending on how the energy is

lowered, offering a way to scan the energy landscape efficiently in search of a global minimum. Genetic algorithms, on the other hand, are inspired by the biological process of evolution and the idea of survival of the fittest [53–57].

Optimization methods have been used to study many different chemistries, often with empirical potentials. However, a growing number of studies are now being performed purely on first-principles computations (e.g., the Na-N [58], W-N [59], Fe-B [60] chemical systems). New phases proposed by optimization approaches include new high-pressure phases of boron [59], $CaCO_3$ [55, 61], and FeB_4 [62] as well as a new metastable polymorph of LiBr [63]. The optimization approach to structure prediction is very appealing but suffers from very extensive requirements in terms of computational budget, especially when multicomponent systems are explored. For instance, finding the ground state of $MgSiO_3$ by a genetic algorithm required around 1,000 energy evaluations [56].

3.2 Data Mining Approaches

The optimization approach assumes no previous knowledge (except for the energy model). On the other hand, solid state chemists have been for long using empirical or heuristic rules to rationalize and sometimes predict crystal structures. A very well known example of such a set of rules is the Pauling rules relating stability to atomic factors (such as ionic size, charge) and structural factors (such as the number of edges or facets shared by cation-anion polyhedra) [64].

Another common heuristic approach consists of building structure maps [65–67]. Structure maps rely on the existence of common *crystal structure prototypes*. Different compounds can form similar arrangement of atoms called prototypes. Traditionally, these structure prototypes are named after the formula and/or name of the mineral from one of the compounds forming this structure. For example, the "NaCl" or "rocksalt" structure prototype is formed not only by NaCl but also by CoO, AgBr. etc. (see Fig. 4).

Structure maps are constructed by plotting for what values of atomic factors certain crystal structure prototypes form. These atomic factors can, for instance, be ionic radii or chemical scales such as the Mendeleev number in Pettifor maps. If the factors are relevant, the structure types will cluster in different regions of the structure map.

Empirical rules such as the Pauling rules are not really predictive and are mainly used to rationalize the existence of already characterized crystals. While structure maps can be used as a predictive tool as shown by Morgan et al. [68], they present limitations due to their focus on specific factors such as size or electronegativity and tend to be available only for very well populated stoichiometries.

Inspired by the success of empirical rules, researchers have been developing data mining or machine learning techniques that learn from previous computations or experiments and make informed guesses about likely crystal structure candidates [69]. The approach relies greatly on the recent developments in data mining,

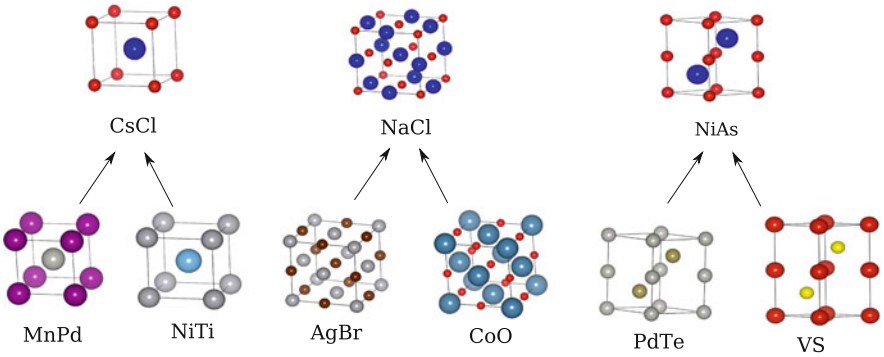


Fig. 4 Some examples of compounds and their crystal structure prototypes

machine learning, and statistical learning [70]. While we will focus on inorganic compounds in this chapter, data mining approaches are also used more and more in the fields of organic chemistry (see for instance [71, 72]).

Sections 4, 5, and 6 will present in more detail some data mining approaches to compound prediction. They all rely on the use of a database of experimental or computed data that is used to fit a probabilistic or regression model. This data mined model can propose likely compound and crystal structure candidates that are tested for stability with DFT.

4 Linear Regression Based Approaches to Data Mining Crystal Structure Prediction

The work from Curtarolo et al. pioneered the use of data mining approaches in combination with ab initio computations [73]. The authors focused on the correlations existing between the energy of crystal structure prototypes in a binary system.

4.1 The Principal Component Analysis Model

Curtarolo et al. built a database of 114 crystal structure prototypes in 55 binary metallic systems. They computed the energy of each of those compounds using DFT.

The information included in this database can be expressed as a series of 55 vectors E_i (1 for each binary system) with 114 dimensions:

$$E_i = (E_{i1}, E_{i2}, \ldots, E_{in}) \tag{5}$$

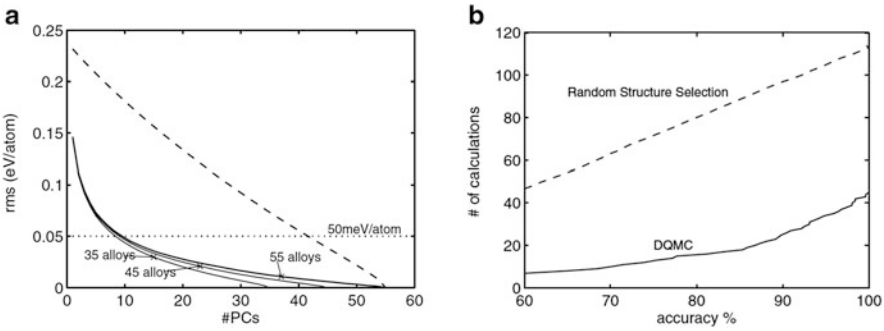


Fig. 5 Root mean squared error in function of the dimension reduction (**a**) and (**b**) number of computations as a function of the number of ground states accurately predicted. The *dashed line* indicates picking the structures randomly and the *plain lines* indicate the data mining driven approach. Reprinted figures with permission from [73]. Copyright 2003 by the American Physical Society

If the energies are not distributed randomly in the 114 dimension (i.e., if there are correlations between energies in different alloys and crystal structures), we can represent the energy vectors in a subspace of lower dimension than the full 114 dimensions. This dimension reduction can be formally performed with the commonly used principal component analysis (PCA).

PCA starts by expressing the vector E_i as an expansion on a subspace of smaller dimension:

$$E_i = \sum_{j=1}^{d} \alpha_{ij} e_i + \varepsilon_i(d), \tag{6}$$

where ε_i is the error on the alloy i. PCA then finds the basis set $\{e_i\}$ minimizing the squared sum of errors $\sum_i \varepsilon_i^T \varepsilon_i$. This new basis set consists of a new set of axes in the 114 dimension space that are adequate to represent our set of alloy energies in reduced dimensions.

Reducing the dimension naturally induces an error compared to the full database in the 114 dimensions. The smaller the dimension reduction (the larger d), the smaller the error induced by dimensional reduction. This is illustrated in Fig. 5a which shows the root mean squared error depending on the number of dimensions. Only nine dimensions (nine alloys) are necessary to obtain the energy of an alloy in a specific crystal structure within an error of 50 meV/atom.

4.2 Prediction Procedure

The correlations indicated by the PCA can be used to accelerate the prediction of new phases. Using these correlations, the amount of ab initio computations to perform can be reduced dramatically. A data mining driven structure prediction

procedure consists of three stages: *prediction*, *suggestion*, *calculations*. Given a previously computed library of crystal structure prototypes in different alloys, we can use the PCA to predict the energies of crystal structures not computed yet in a given alloy. Using these data mined predicted energies we can identify the structures that are the farthest below the convex hull or the closest to the hull. This limited set of candidates are then computed by DFT. The new DFT results are added to the database and a new series of prediction, suggestion, calculations is performed until a convergence to a stable solution is reached.

Figure 5b compares the number of calculations required to reach a certain percentage of ground states accurately predicted in both the random selection (dashed line) and data mining driven case (plain line). The data mining approach performs significantly better.

This technique has been used to perform searches of new borides [74, 75] or rhodium alloys [76].

5 Data Mining Approach Based on Correlations Between Crystal Structure Prototypes

The approach based on PCA presented in Sect. 4 is of great interest but requires a database of computed energies for known (often stable) compounds and for hypothetical compounds (often unstable) and their crystal structures. Such a database is unfortunately not available for most areas of chemistry. On the other hand, experimental crystal structure databases such as the ICSD are widely available, giving access to observed inorganic compounds. In 2006, Fischer et al. proposed an approach based on correlations between observed crystal structures that do not require any previous computational data [77]. Instead of a regression problem (i.e., predicting continuous quantities such as energies), a classification problem is tackled: predicting whether a given crystal structure is likely to be stable or not (without modeling how stable it will be). We will present here the algorithm in detail and its application on a high-throughput large scale search for ternary oxides [78].

5.1 General Principle of the Algorithm

Crystalline inorganic compounds have a limited set of crystal structure prototypes (see Fig. 4). The basic idea behind the algorithm is to consider that the presence of a given crystal structure prototype in a chemical system can be correlated to factors such as the elements in this chemical system and the crystal structures co-existing at other compositions. For instance, the crystal structure prototype of $LaMn_2O_5$ forms very often with Mn. A strong correlation exists between the presence of this crystal

structure prototype in a chemical system and manganese. Likewise, the $FeSb_2O_6$ and Sb_2O_5 crystal structure prototypes are also strongly correlated. From this observation one can think about using partial information about a chemical system (e.g., the presence of Mn or of the Sb_2O_5 prototype) to infer the crystal structures likely to form. In the following sections we will discuss how this basic idea is implemented mathematically. The data abstraction and variables will be introduced along with the probabilistic model rigorously integrating all those correlations.

5.2 Data Abstraction

We will assume that a prototype label has been assigned to all the compounds in the database. This prototyping step can be fully automated by using, for instance, the algorithm proposed by Hundt et al. [79]. After transformation of the raw database to a prototyped database, the data are in the form of a composition-crystal structure prototype pair for each compound.

For the sake of simplicity we will use discrete composition variables in our model. Compositions are continuous variables and, to project this continuous problem to a discrete one, we will consider any composition to be present in a composition bin. For instance, the composition bins could be AB, A_2B, AC_2, etc. for the binaries and ABC, ABC_2, etc. for the ternaries. Each of these composition bins c_i is associated with a variable x_{c_i} indicating what crystal structure is present at this composition. For example, if c_i represents the composition AB_2C_4 then x_{c_i} may have values such as *spinel*, *olivine*, etc. The condition $x_{c_i} = \text{no structure}$ value indicates the absence of a compound at the given composition. In addition, variables representing the system's constituents (e.g., E_i=Ag, Cu, Na, etc.) are defined. With these definitions, any chemical system of C constituents and n compositions can be represented by a vector $X = (x_{c_1}, x_{c_2}, \ldots, x_{c_n}, x_{E_1}, x_{E_2}, \ldots, x_{E_c})$ where the composition space is discretized by using n composition bins.

In this formalism, any information from the database on a chemical system can be represented by an instance of the vector $\mathbf{X}$ (see Fig. 6). Any prototyped crystal structure database D can then be represented as a collection of N $\mathbf{X}_i$ instances, $\mathsf{D} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N\}$.

5.3 Probabilistic Function and New Compound Discovery Procedure

The probability density $p(\mathbf{X})$ provides information as to what crystal structures tend to coexist in a chemical system. Based on the available information at known compositions in a system, this probability density can be used to assess if another composition (c_j) is likely to be compound-forming. Mathematically, this is evaluated by computing the probability of forming a compound:

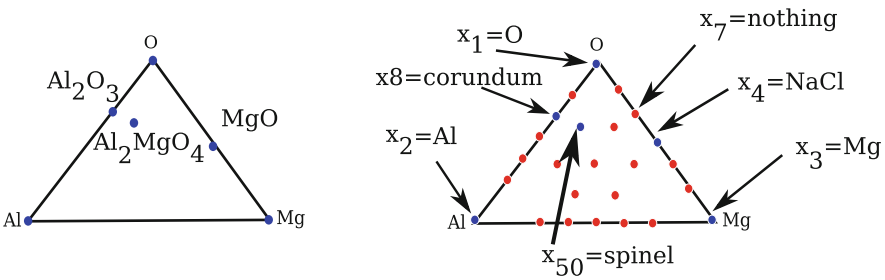


Fig. 6 An example of how the information on the Al-Mg-O chemical system is projected onto the composition variables. All *dots* indicate composition bins. *Red dots* are composition bins without any known compound and *blue dots* are composition bins with a known compound crystallizing in a specific prototype marked by an *arrow*

$$p_{\text{compound}}(c_j) = 1 - p\left(x_{c_j} = \text{nostructure} \middle| x_{c_1}, x_{c_2}, \ldots, x_{c_{j-1}}, x_{c_{j+1}}, \ldots, x_{c_n}, \ldots, x_{E_1}, x_{E_2}, \ldots, x_{E_c}\right). \tag{7}$$

In addition, when a composition c_j of interest is targeted, the probability density can be used to suggest the most likely crystal structures by evaluating the following:

$$p\left(x_{c_j} \middle| x_{c_1}, x_{c_2}, \ldots, x_{c_{j-1}}, x_{c_{j+1}}, \ldots, x_{c_n}, \ldots, x_{E_1}, x_{E_2}, \ldots, x_{E_c}\right). \tag{8}$$

For the different values of x_{c_j} (i.e., for the different crystal structure prototypes known at this composition), a list of the l most likely crystal structure candidates can be established. These candidate crystal structures can then be tested for stability by an accurate energy model such as DFT. The procedure for compound discovery is summarized in Fig. 7.

We should stress that, in contrast to most optimization techniques, this approach can not only suggest likely crystal structures for a given composition but also suggest which compositions are likely to form stable compounds. This is very important, especially for multi-component systems (ternaries or quaternaries), as the compositional space is larger than for binary compounds.

5.4 Approximated Probabilistic Function

While very useful for structure prediction, this probability function is extremely complex. In the case of ternary oxides, our model requires 183 variables. With roughly 100 crystal structure prototypes possible per variable, this probability function is defined on a domain of around 10^{366} values!

For all practical purpose this probability function needs to be approximated. The way the approximation is made here is to use an approach known in statistical

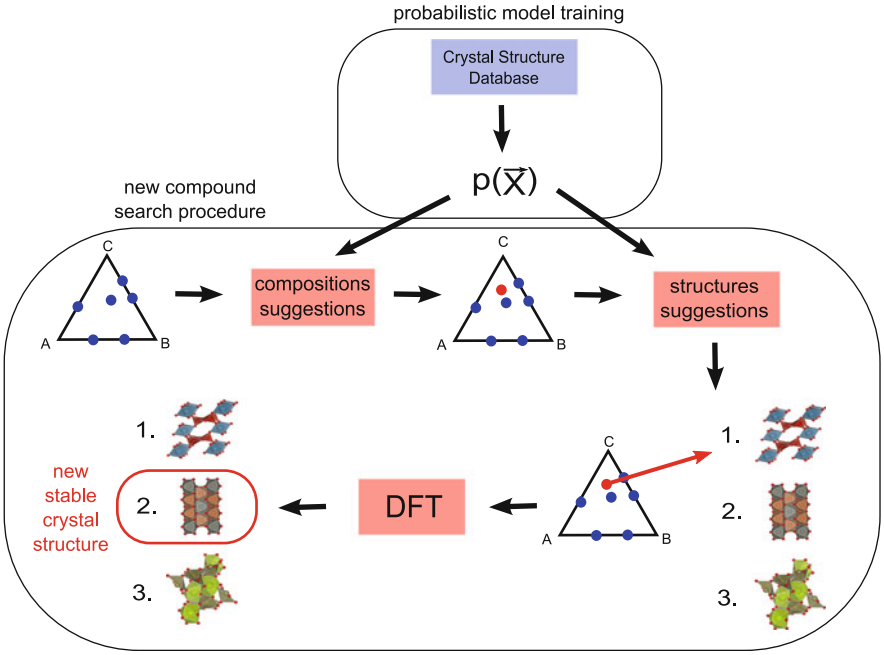


Fig. 7 Data-mining driven compound discovery procedure. A probabilistic model is built from a crystal structure database. In any system A-B-C, this model is used to identify the new compositions (*red dots*) most likely to form a compound. For those compositions, the most likely crystal structures are proposed using the same probabilistic model. These structure candidates are then tested for stability by an accurate energy model as DFT

mechanics as the cumulant expansion [80]. The cumulant expansion can be presented starting with the identity

$$p(\mathbf{X}) = \prod_i g_i(x_{c_i}) \prod_{j<k} g_{jk}\left(x_{c_j}, x_{c_k}\right) \prod_{l<m<n} g_{lmn}(x_{c_l}, x_{c_m}, x_{c_n}) \ldots \tag{9}$$

Following this expression, $p(\mathbf{X})$ can be seen as a product of independent variables with corrections from pair, triplet, etc., correlations. The cumulant terms can be defined recursively. Starting with a one variable probability function, we trivially have

$$g_i(x_{c_i}) = p(x_{c_i}); \tag{10}$$

with a two variables probability function we have

$$p\left(x_{c_i}, x_{c_j}\right) = p(x_{c_i}) p\left(x_{c_j}\right) g_{ij}\left(x_{c_i}, x_{c_j}\right), \tag{11}$$

which implies that

$$g_{ij}\left(x_{c_i}, x_{c_j}\right) = \frac{p\left(x_{c_i}, x_{c_j}\right)}{p(x_{c_i})p\left(x_{c_j}\right)}. \tag{12}$$

The general form for a cumulant over the variable X_α is

$$g_\alpha(x_\alpha) = \frac{p(x_\alpha)}{\Pi_{\beta\subset\alpha} g_\beta\left(x_\beta\right)}, \tag{13}$$

for which the products at the denominator extends over all subsets of α.

So far, no approximation has been introduced. The approximation will consist of truncating the cumulant expansion, considering that all the cumulants beyond pairs (triplets, quadruplets etc. . .) are equal to 1 so that

$$p(\mathbf{X}) = \frac{1}{Z} \prod_i p(x_{c_i}) \prod_{j<k} \frac{p\left(x_{c_i}, x_{c_j}\right)}{p(x_{c_i})p\left(x_{c_j}\right)}, \tag{14}$$

where Z is a normalization constant or partition function.

5.5 *Estimating the Probabilistic Function from Available Data*

Having decided on the form of an approximated probability function (14), we still need to estimate the values of these function parameters. Using a database D, we will search for the values $p\left(x_{c_i}, x_{c_j}\middle|\mathsf{D}\right)$ and $p\left(x_{c_i}\middle|\mathsf{D}\right)$ in best agreement with the data. One can see this process – called parameter estimation – as a fit of the model to the available data.

We will present two common ways of estimating the parameters of a probabilistic model from the data: the maximum likelihood and the Bayesian approach. For pedagogical purposes we will first present derivations for the single variable case and will generalize later on the multi-variable case [81].

5.5.1 Single Variable Multinomial Parameter Estimation by Maximum Likelihood

Let us assume a random variable X that can take on n possible values $x \in \{v_1, v_2, \ldots, v_q\}$. Assuming we have a database D of N observed values for $\mathsf{D} = \{x_1, x_2, \ldots, x_N\}$, we would like to infer the probability function $p\left(x\middle|\mathsf{D}\right)$. For each of the possible q values of X we assign a parameter with the value of the probability function. We then have q parameters θ_{v_i} with $p(x = v_i) = \theta_{v_i}$. All these parameters can be for notation purpose regrouped in one vector $\boldsymbol{\theta}$.

It is very common to approach the parameter estimation using the maximum likelihood approach [82]. The best estimate for the parameter is the one maximizing the (log)-likelihood of the data l:

$$l(\mathsf{D},\boldsymbol{\theta}) = \log p\left(\mathsf{D}\,|\boldsymbol{\theta}\right) = \log p\left(x_1, x_2, \ldots, x_N|\boldsymbol{\theta}\right) = \sum_{t=1}^{N} \log p\left(x_t|\boldsymbol{\theta}\right)$$
$$= \sum_{x} n(x) \log \theta_x \tag{15}$$

This derivation has been performed assuming that all the x_i observations are independent. $n(x)$ indicates the number of times the value x is observed in the data D. Maximizing the likelihood function in (15), under the constraint that $\sum_x \theta_x = 1$, leads to

$$\theta_x^{\mathrm{ML}} = \frac{n(x)}{\sum_{x'} n(x')}. \tag{16}$$

The maximum likelihood estimate of the probability for a given value to be drawn is therefore the frequency at which this value appeared in the data set.

5.5.2 Single Variable Multinomial Parameter Bayesian Estimation

In the simple maximum likelihood approach presented in the previous section, there is one set of values for the $\boldsymbol{\theta}$ parameters. Another approach, called Bayesian estimation, considers that assigning a *unique* value for a parameter is too rigid and argues that one should be interested in discovering instead the probability distribution of the parameter $p\left(\boldsymbol{\theta}|\mathsf{D}\right)$. As an illustration, if one is observing a coin toss leading to 1,001 heads and 999 tails, a maximum likelihood approach would find out that the probability for heads should be 0.5005. A Bayesian approach, in contrast, will argue that from this information one cannot rule out the possibility that the value of the parameter is 0.5 for example. From this information the Bayesian approach would rather propose a $p\left(\boldsymbol{\theta}|\mathsf{D}\right)$ peaked on 0.5005 but allowing some spread and non-zero values for values close to 0.5005. A very complete presentation of the Bayesian approach to probability can be found in Jaynes [83].

In the Bayesian approach, the probability for a value x to be observed is now computed by integrating on all possible values of $\boldsymbol{\theta}$ weighted by their probability:

$$p\left(x|\mathsf{D}\right) = \int p\left(x|\boldsymbol{\theta},\mathsf{D}\right) p(\boldsymbol{\theta},\mathsf{D}) d\boldsymbol{\theta}. \tag{17}$$

The parameters θ_x are now defined as

$$\theta_x = p\left(x\middle|\boldsymbol{\theta}, \mathsf{D}\right). \tag{18}$$

The parameter estimation process consists in finding $p\left(\boldsymbol{\theta}\middle|\mathsf{D}\right)$. Using Bayes' rule of probability, we can show that

$$p\left(\boldsymbol{\theta}\middle|\mathsf{D}\right) = p\left(\mathsf{D}\middle|\boldsymbol{\theta}\right)\frac{p(\boldsymbol{\theta})}{p(\mathsf{D})} \tag{19}$$

$$= p\left(x_1, x_2, \ldots, x_N\middle|\boldsymbol{\theta}\right)\frac{p(\boldsymbol{\theta})}{p(x_1, x_2, \ldots, x_N)} \tag{20}$$

$$= \lambda \prod_x \theta_x^{n(x)} p(\boldsymbol{\theta}) \tag{21}$$

With $\lambda = \dfrac{1}{p(x_1, x_2, \ldots, x_N)}$.

A new quantity appeared during this derivation: $p(\boldsymbol{\theta})$. This is called the *prior* on the parameters. This represents the a priori belief the observer had before any observation was actually done. In the multinomial case, a common prior used for convenience is the Dirichlet distribution:

$$p(\boldsymbol{\theta}) = \beta(\boldsymbol{\alpha}) \prod_x \theta_x^{\alpha_x - 1}, \tag{22}$$

where $\beta(\boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_x \alpha_x\right)}{\prod_x \Gamma(\alpha_x)}$ and Γ is the Gamma function. Plugging the Dirichlet prior (22) in the expression of the posterior (20), we get

$$p\left(\boldsymbol{\theta}\middle|\mathsf{D}\right) = \lambda\beta(\boldsymbol{\alpha}) \prod_x \theta_x^{n(x)+\alpha_x - 1}. \tag{23}$$

As we can see, using the Dirichlet prior with a multinomial distribution leads to a multinomial distribution as posterior. This very convenient behavior makes the Dirichlet distribution the so-called conjugate prior of a multinomial distribution.

The last piece of our problem not yet solved is the value of λ. We can use the normalization condition $\int p(\boldsymbol{\theta}|\mathsf{D})d\boldsymbol{\theta} = 1$. Applying this constraint, it can be shown that

$$p\left(\boldsymbol{\theta}\middle|\mathsf{D}\right) = \Gamma\left(\sum_{x'} n\left(x'\right) + \alpha_{x'}\right) \prod_x \frac{\theta_x^{n(x)+\alpha_x - 1}}{\Gamma(n(x) + \alpha_x)} \tag{24}$$

$$= C(n, \boldsymbol{\alpha}) \prod_{x'} \theta_{x'}^{n\left(x'\right)+\alpha_{x'}}, \tag{25}$$

where the part of the expression involving the Gamma function has been regrouped for clarity in $C(n, \boldsymbol{\alpha})$. Now that we have found the expression for $p\left(\boldsymbol{\theta}\middle|\mathsf{D}\right)$, we can evaluate the probability to observe a value v_i for the variable X:

$$p(x = v_i) = \int \theta_{v_i} p(\boldsymbol{\theta}, \mathsf{D}) d\theta \tag{26}$$

$$= C(n, \boldsymbol{\alpha}) \int \theta_{v_i} \prod_{x'} \theta_{x'}^{n(x')+\alpha_{x'}} d\theta \tag{27}$$

$$= \frac{n(v_i) + \alpha_{v_i}}{\sum_{x'} n(x') + \alpha_{x'}}. \tag{28}$$

This final expression can be compared to that obtained using the maximum likelihood (16). The way the prior influences the result is by adding extra counts α_x to the evaluation of the probability. We can see that if there is an important amount of data available the probability will be driven mainly by the frequency of counts. On the other hand, if there are very few data points, the prior will drive the probability.

While we have chosen the Dirichlet prior, we still have to choose what parameters $\boldsymbol{\alpha}$ to use. There is no unique answer to that question. This choice would depend on the prior belief we have in the outcome. In the case of no prior information being available [84, 85], there is a common choice of prior called the minimum information uniform Dirichlet prior, where $\boldsymbol{\alpha}$ is chosen as

$$\alpha_x = \frac{1}{q} \tag{29}$$

where q represents the number of possible values for X.

5.5.3 Generalization to Multiple Variables

The results presented in the two previous sections can be generalized for multiple variables. Let us say that we have two variables X and Y and we want to estimate $p(x, y|\mathsf{D})$. D refers to a set of N observations $\mathsf{D} = \{(x, y)_1, (x, y)_2, \ldots, (x, y)_N\}$. If there are q possible values for X and r values possible for Y, then there are qr possible values for the pair (X, Y). Results from the single variable case can then be directly used with a multinomial defined on qr values. Then the maximum likelihood is

$$\theta_{x,y}^{\mathrm{ML}} = \frac{n(x, y)}{N}; \tag{30}$$

the Bayesian estimate is

$$p(x = v_i, y = w_j|\mathsf{D}) = \frac{n(v_i, w_j) + \alpha_{v_i, w_j}}{N + \sum_{x,y} \alpha_{x,y}}; \tag{31}$$

and the minimum information Dirichlet prior is

$$\alpha_x = \frac{1}{qr}. \tag{32}$$

5.6 *Searching for Unknown Ternary Oxides Using Data Mining Compound Prediction*

Ternary oxides are important for many technologies. The model presented here has been used to search for new ternary oxides. We estimated a cumulant expansion probabilistic model (14) using the oxide experimental data available in the ICSD [24] and the Bayesian estimation procedure presented in Sect. 5.5. The 2006 version of the ICSD was searched for duplicate compounds. After this analysis, 616 unique binary and 4,747 ternary oxides compounds were identified. These compounds were grouped by crystal structure prototype. Both duplicate checks and prototyping were performed using Hundt et al.'s algorithm [79]. Composition bins were binned into the 30 most common binary compositions and the 120 most common ternary compositions. Any compound not fitting perfectly in one of these bins was binned in the closest composition bin. Adding the 3 element variables, 183 variables were used in total in the probability model.

5.6.1 New Ternary Oxides Predictions

We then searched for new compounds in 2,211 A-B-O systems with A and B taken from H, Li, Be, B, C, N, F, Na, Mg, Al, Si, P, S, Cl, K, Ca, Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Ga, Ge, As, Se, Br, Rb, Sr, Y, Zr, Nb, Mo, Ag, Cd, In, Sn, Sb, Te, I, Cs, Ba, La, Hf, Ta, W, Pt, Hg, Tl, Pb, Bi, Ce, Pr, Nd, Sm, Eu, Gd, Dy, Ho, Er, Tm, Yb, and Lu. In these systems we used the procedure described in Fig. 7 and searched for compositions where no ternary oxide is given in the ICSD but for which the probability for forming a compound (7) is higher than a certain threshold. This threshold represents a compromise between the computational budget required and the rate of discovery expected. The value of the threshold we chose suggested 1,261 possible compositions and exhibited a 45% true positive rate during cross-validation. At these selected compositions, the most likely crystal structures were determined from the data mined probability density using (8). The number of suggested crystal structures at each composition corresponds to the list length that gave 95% accuracy in cross-validation. This corresponds to a total of 5,546 crystal structures whose energy needed to be calculated with ab initio DFT. All existing binary, ternary, and element structures in the ICSD were also calculated so that relative phase stability can be assessed (using the thermodynamical convex hull construction presented in Sect. 2). Hence, a new structure is stable when its energy is lower than any combination of energies of compounds in the system weighted to the same composition.

From the 1,261 compositions suggested by the model, the ab initio computations confirmed 355 to be stable against every compound known in the ICSD.

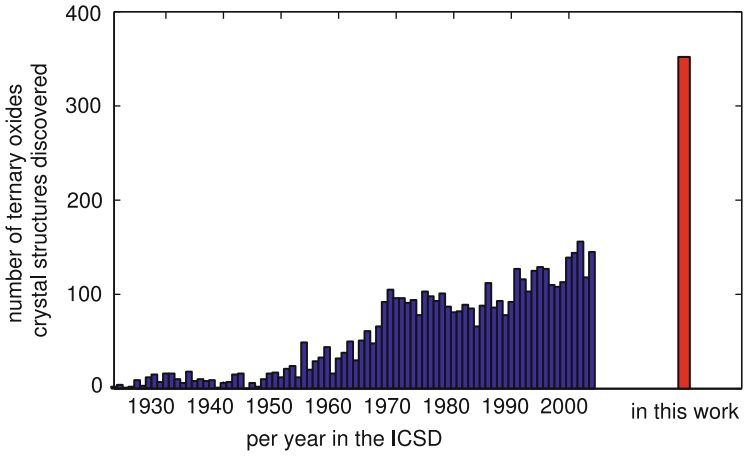


Fig. 8 New ternary oxide discovery per year according to the ICSD. The *bars* from 1930 to 2005 indicate the number of new ternary oxides discovered per year. They are compared to the number of new compounds discovered in this work

This represents 1 new stable compound predicted per 16 DFT computations. A fully exhaustive search (i.e., computing all possible structure prototypes in any composition bin) in the 2,211 A-B-O systems of interest would be prohibitive and require 5,428,287 computations. Even restricting such an exhaustive search to the crystal structure prototypes present in the selected 1,261 compositions bins would need substantially more computations (183,007) than the 5,546 needed while using the machine learned model.

To put this number of 355 new compounds predicted in perspective, we compared it to the number of experimentally discovered and characterized ternary oxides. We identified the earliest date of publication for any ternary oxide compound present in the ICSD. We did not take into account multiple reports of the same compound and compounds with partial occupancies. Figure 8 indicates in blue how many new ternary oxide compounds were discovered each year according to the ICSD from 1930 to 2005. The red bar shows how many new compounds have been discovered in this work. The experimental discovery rate for ternary oxides is around 100 per year since the 1970s. The 355 new compounds suggested were obtained with about 55 days of computing on 400 Intel Xeon 5140 2.33-Ghz cores. Those numbers show the potential for accelerating new compound discovery through combining data mining with DFT computations.

Details and discussion on the results are available in Hautier et al. [78] and details of all the new compounds are available on a web site [86].

6 Data Mined Ionic Substitution Model

In Sect. 6 we present a compound prediction algorithm based on correlations between crystal structures co-existing in a same chemical system. This algorithm was used in combination with high-throughput DFT computations to discover new ternary oxides.

While, in theory, this algorithm can be used to make predictions in chemical systems with any number of components, there are practical limitations to its application, for instance, to the prediction of quaternary compounds. Indeed, the data available for quaternaries is sparser than for ternaries, making the extraction of informative correlations more difficult. More specifically, as the model presented in the previous section is based on correlations between crystal structure prototypes, it shows predictive limits for the crystal structure prototypes appearing only once in the database. Those unique crystal structure prototypes do not have enough occurrences for the model to capture useful correlations. The problem associated with unique prototypes is already present in ternary compounds but tends to be even more critical in the quaternary space. In the ICSD, 20% of the ternary crystal structure prototypes are unique but up to 50% are unique in the case of quaternary prototypes.

In the coming section we will show how a different data mining approach can be used to make predictions in sparser regions. A probabilistic model can be built to assess the likelihood for ionic species to substitute for each other while retaining the crystal structure [87]. We describe the mathematical model and its training on an experimental crystal structures database. The model predictive power is then evaluated by cross-validation and the emerging chemical substitution rules are analyzed.

6.1 Ionic Substitution Approach to New Compound Discovery

Chemical knowledge often drives researchers to postulate new compounds based on substitution of elements or ions from another compound. For instance, when the first superconducting pnictide oxide $LaFeAsO_{1-x}F_x$ was discovered, crystal chemists started to synthesize many other isostructural new compounds by substituting lanthanum with other rare earth elements such as samarium [88].

A formalization of this substitution approach exists in the Goldschmidt rules of substitution, stating that the ions closest in radius and charge are the easiest to substitute for each other [89]. While those rules have been widely used to rationalize a posteriori experimental observations, they lack a real quantitative predictive power.

The data mining ionic substitution approach follows this substitution idea but proposes a mathematical and quantitative framework around it. The basic principle is to learn from an experimental database how likely the substitution of certain ions

in a compound will lead to another compound with the same crystal structure. Mathematically, the substitution knowledge is embedded in a substitution probability function. This probability function can be evaluated to assess quantitatively if a given substitution from a known compound is likely to lead to another stable compound. For instance, in the simple case of the $LaFeAsO_{1-x}F_x$ compound we expect the probability function to indicate a high likelihood of substitution between La^{3+} and Sm^{3+} and thus a high likelihood of existence for the $SmFeAsO_{1-x}F_x$ compound in the same crystal structure as $LaFeAsO_{1-x}F_x$ but with Sm on the La sites.

This method follows an approach used in the field of machine translation [90]. The aim of machine translation is to develop models able to translate texts from one language to another. Therefore, one approach is to build probabilistic models that evaluate the probability for a word in one language to correspond to another word in another language. In the case of our ionic substitution model, the approach is similar but it is a correspondence between ionic species instead of words that is sought.

6.2 The Probabilistic Model

We present here the different variables and the mathematical form of the substitution probabilistic model.

Let us represent a compound formed by n different ions by an n component vector:

$$\mathbf{X} = (X_1, X_2, \ldots X_n). \tag{33}$$

Each of the X_j variables are defined on the domain Ω of existing ionic species:

$$\Omega = \left\{Fe^{2+}, Fe^{3+}, Ni^{2+}, La^{3+}, \ldots\right\}. \tag{34}$$

The quantity of interest to assess the likelihood of an ionic substitution is the probability p_n for two n-component compounds to exist in nature in the same crystal structure. If X_j and X'_j respectively indicate the ions present at the position j in the crystal structure common to two compounds, then one needs to determine

$$p_n\left(\mathbf{X}, \mathbf{X}'\right) = p_n\left(X_1, X_2, \ldots, X_n, X'_1, X'_2, \ldots, X'_n\right). \tag{35}$$

Knowing such a probability function allows one to assess how likely any ionic substitution is. For example, by computing $p_4(Ni^{2+}, Li^{1+}, P^{5+}, O^{2-}|Fe^{2+}, Li^{1+}, P^{5+}, O^{2-})$, one can evaluate how likely Fe^{2+} in a lithium transition metal phosphate is to be substituted by Ni^{2+}. In this specific example, this value is expected to be high as Ni^{2+} and Fe^{2+} are both transition metals with similar charge

and size. Actually, $LiNiPO_4$ and $LiFePO_4$ both form in the same olivine-like structure. On the other hand, the substitution of Fe^{2+} by Sr^{2+} would be less likely and $p_4(Sr^{2+}, Li^{1+}, P^{5+}, O^{2-}|Fe^{2+}, Li^{1+}, P^{5+}, O^{2-})$ should have a low value. We must point out that the probability function does not have any crystal structure dependence. The fact that the compound targeted for substitution forms an olivine structure does not influence the result of the evaluated probability. This is an approximation in our approach.

The probability function $p_n(\mathbf{X},\mathbf{X}')$ is a multivariate function defined in a high-dimensional space and cannot be estimated directly. For all practical purposes, this function needs to be approximated. We follow here an approach successfully used in other fields such as machine translation and, based on the use of binary indicators f, so-called *feature functions*.[91] These feature functions are mathematical representations of important aspects of the problem. The only mathematical requirement for a feature function is to be defined on the domain of the probability function $(\mathbf{X},\mathbf{X}')$ and return 1 or 0 as a result. They can be as complex as required by the problem. For an ionic substitution model, one could choose, for example, as a feature function:

$$f\left(\mathbf{X},\mathbf{X}'\right) = \begin{cases} 1 & \text{if } Ca^{2+} \text{ substitutes for } Ba^{2+} \text{ in the presence of } O^{2-} \\ 0 & \text{else} \end{cases} \tag{36}$$

The relevant feature functions are commonly defined by experts from prior knowledge. If our chosen set of feature functions are informative enough, we expect to be able to approximate the probability function by a weighted sum of those feature functions:

$$p_n\left(\mathbf{X},\mathbf{X}'\right) \approx \frac{e^{\sum_i \lambda_i f_i^{(n)}\left(\mathbf{X},\mathbf{X}'\right)}}{Z}. \tag{37}$$

Here λ_i indicates the weight given to the feature $f_i^{(n)}(\mathbf{X},\mathbf{X}')$ in the probabilistic model. Z is a partition function ensuring the normalization of the probability function. The exponential form chosen in (37) follows a commonly used convention in the machine learning community [92].

The model presented is extremely general and can be adjusted by using whatever feature function is considered relevant. A first assumption made is to consider that the feature functions do not depend on the number n of ions in the compound. Simply put, we assume that the ionic substitution rules are independent of the compound's number of components (binary, ternary, quaternary, etc.).

Therefore we will omit any reference to n in the probability and feature functions. Equation (37) then becomes

$$p_n\left(\mathbf{X},\mathbf{X}'\right) \approx \frac{\mathrm{e}^{\sum_i \lambda_i f_i\left(\mathbf{X},\mathbf{X}'\right)}}{Z}. \tag{38}$$

While the feature functions could be more complex, only simple binary substitutions are considered in this work. This means that the likelihood for two ions to substitute for each other is independent of the nature of the other ionic species present in the compound. Mathematically, this translates into the assumption that the relevant feature functions are simple binary features of the form

$$f_k^{a,b}\left(\mathbf{X},\mathbf{X}'\right) = \begin{cases} 1 & X_k = a \quad \text{and} \quad X'_k = b \\ 0 & \text{else} \end{cases} \tag{39}$$

Each pair of ions a and b present in the domain Ω is assigned a set of feature functions with corresponding weights $\lambda_k^{a,b}$ indicating how likely the ions a and b can substitute in position k. For instance, one of the feature functions will be related to the Ca^{2+} to Ba^{2+} substitution:

$$f_k^{\mathrm{Ca}^{2+},\mathrm{Ba}^{2+}}\left(\mathbf{X},\mathbf{X}'\right) = \begin{cases} 1 & X_k = \mathrm{Ca}^{2+} \quad \text{and} \quad X'_k = \mathrm{Ba}^{2+} \\ 0 & \text{else} \end{cases} \tag{40}$$

The magnitude of the weight $\lambda_k^{\mathrm{Ca}^{2+},\mathrm{Ba}^{2+}}$ associated with this feature function indicates how likely this binary substitution is to happen.

Finally, the features weights should satisfy certain constraints so that any permutations of the components do not change the result of the probability evaluation. Those symmetry conditions are

$$\lambda_k^{a,b} = \lambda_k^{b,a}, \tag{41}$$

and

$$\lambda_k^{a,b} = \lambda_l^{a,b}. \tag{42}$$

6.3 Training of the Probability Function

While the mathematical form for our probabilistic model is now well established, the model parameters (the weights $\lambda_k^{a,b}$) still need to be evaluated. Those weights are estimated from the information present in an experimental crystal structure database.

From any experimental crystal structure database, structural similarities can be obtained using structure comparison algorithms [79, 93]. For instance, $CaTiO_3$ and $BaTiO_3$ both form cubic perovskite structures with Ca and Ba on equivalent sites. This translates in our mathematical framework as a specific assignment for the variables vector $(\mathbf{X},\mathbf{X}') = (Ca^{2+},Ti^{4+},O^{2-},Ba^{2+},Ti^{4+},O^{2-})$. We will follow the

convention in probability theory, designing specific values of the random variable vector ($\mathbf{X},\mathbf{X}'$) by lower case letters ($\mathbf{x},\mathbf{x}'$). An entire crystal structure database D will lead to m assignments $(\mathbf{X},\mathbf{X}') = (\mathbf{x},\mathbf{x}')^t$ with $t = 1, \ldots, m$

$$\mathsf{D}=\left\{\left(\mathbf{X},\mathbf{X}'\right)=\left(\mathbf{x},\mathbf{x}'\right)^1,\left(\mathbf{X},\mathbf{X}'\right)=\left(\mathbf{x},\mathbf{x}'\right)^2,\ldots,\left(\mathbf{X},\mathbf{X}'\right)=\left(\mathbf{x},\mathbf{x}'\right)^{m-1},\left(\mathbf{X},\mathbf{X}'\right)=\left(\mathbf{x},\mathbf{x}'\right)^m\right\}. \tag{43}$$

Coming back to our analogy to machine translation, probabilistic translation models are estimated from databases of texts with their corresponding translation. The analogue to the translated texts database in our substitution model is the crystal structure database.

Using these assignments obtained from the database, we follow the commonly used maximum-likelihood approach to find the adequate weights from a database [82]. The weights maximizing the likelihood to observe the training data are considered as the best estimates to use in the model. For notation purposes we will represent the set of weights by a weight vector $\boldsymbol{\lambda}$.

From those m assignments, the log-likelihood l of the observed data D can be computed as

$$l(\mathsf{D},\boldsymbol{\lambda}) = \sum_{t=1}^{m} \log p\left(\left(\boldsymbol{x},\boldsymbol{x}'\right)^t|\boldsymbol{\lambda}\right) \tag{44}$$

$$= \sum_{t=1}^{m}\left[\sum_i \lambda_i f_i\left(\left(\boldsymbol{x},\boldsymbol{x}'\right)^t\right) - \log Z(\boldsymbol{\lambda})\right] \tag{45}$$

The feature weights maximizing the log-likelihood of observing the data D ($\boldsymbol{\lambda}_{\mathbf{ML}}$) are obtained by solving

$$\boldsymbol{\lambda}_{\mathbf{ML}} = \arg\max_{\boldsymbol{\lambda}} l(\mathsf{D},\boldsymbol{\lambda}). \tag{46}$$

There is a last caveat in the training of this probability function. Any ionic pair never observed in the data set could theoretically have any weight value. All those unobserved ionic pair weights will be set to a common value α. As these ionic pairs should be unlikely, a low value of α (for instance $\alpha = 10^{-5}$ in the rest of this work) will be used.

6.4 Compound Prediction Process

When the substitution probabilistic model in (37) has been trained, it can be used to predict new compounds and their structures from a database of existing compounds. The procedure to predict a compound formed by species a, b, c, and d is presented

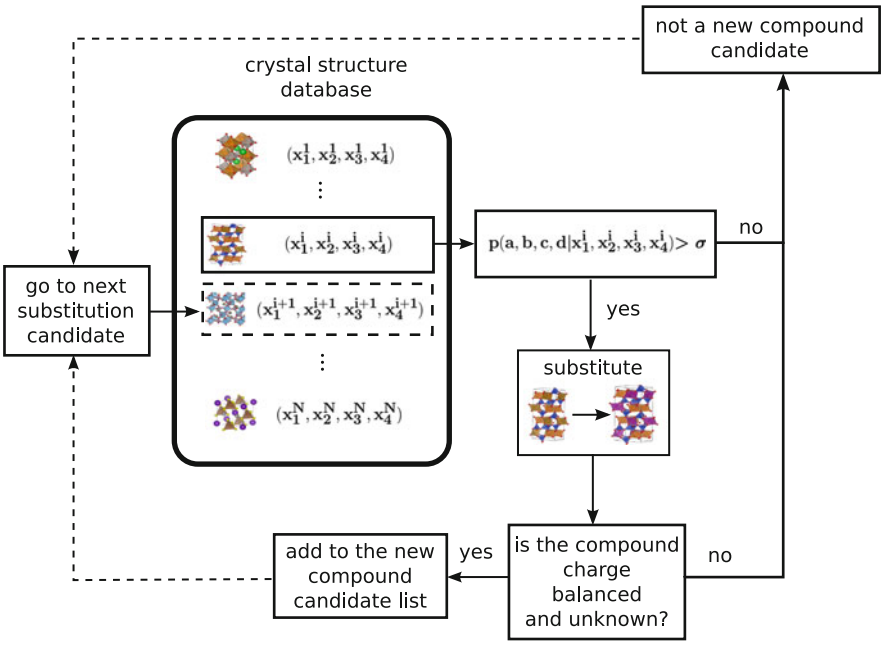


Fig. 9 Procedure to predict new compounds formed by the *a*, *b*, *c*, and *d* species using the substitutional probabilistic model. Reprinted with permission from [87]. Copyright 2011 American Chemical Society

in Fig. 9. For each compound containing $(x_i^1, x_i^2, x_i^3, x_i^4)$ as ionic species, the probability to form a new compound by substitution of *a*, *b*, *c*, and *d* for x_i^1, x_i^2, x_i^3, and x_i^4 is evaluated by computing $p(a, b, c, d|x_i^1, x_i^2, x_i^3, x_i^4)$. If this probability is higher than a given threshold σ, the substituted structure is considered. If this new compound candidate is charge balanced and previously unknown, it can be added to our list of new compound candidates. If not, the algorithm goes to the next $i + 1$ compound in the crystal structure database. The substitutions proposed by the model do not have to be isovalent. However, all suggested compounds have to be charge balanced.

At the end of the new compound prediction process, a list of new compounds candidates in the *a*, *b*, *c*, *d* chemistry is available. This list should be tested in a second step for stability vs all already known compounds by accurate ab initio techniques such as DFT (see Sect. 2).

6.5 *Analysis of the Model*

A binary feature model based on the ternary and quaternary ionic compounds present in the inorganic crystal structure database (ICSD, [24]) has been built. In this work we consider a compound to be ionic if it contains one of the following

anions: O^{2-}, N^{3-}, S^{2-}, Se^{2-}, Cl^{-}, Br^{-}, I^{-}, F^{-}. Only ordered compounds (i.e., compounds without partially occupied sites) are considered. Crystal structure similarity was found using Hundt et al.'s algorithm [79] and used to obtain the database D of m assignments ((43) necessary to train the model. A binary feature model was fitted on this data set using a maximum likelihood procedure.

6.5.1 Cross-Validation on Quaternary ICSD Compounds

The procedure to discover new compounds using the probabilistic model was presented in Sect. 6.4. Using this procedure, we evaluated the predictive power of this approach by performing a cross-validation test [70]. Cross-validation consists in removing part of the data available (the test set) and training the model on the remaining data set (the training set). The model built in this way is then used to predict back the test set and evaluate its performance. We divided the quaternary ordered and ionic chemical systems from the ICSD in three equal-sized groups. We performed three cross-validation tests using all compounds in one of the groups as test set and the remaining quaternary and ternary compounds as training set. This extensive cross-validation tested 2,967 compounds in total. The cross-validation tests excluded compounds forming in prototypes unique to one compound, as our substitution strategy by definition cannot predict compounds in such unique prototypes. We also only considered substitution leading to charge balanced compounds.

Figure 10 indicates the false positive and true positive rates for a given threshold σ. The true positive rate (TP_{rate}) indicates the fraction of existing ICSD compound that are indeed found back by the model (i.e., true hits):

$$TP_{rate}(\sigma) = \frac{TP(\sigma)}{P}, \tag{47}$$

where P is the number of existing compounds considered during our cross-validation test and $TP_{(\sigma)}$ is the number of those existing compounds found by our model with a given threshold σ (i.e., the number of true positives). The false positive rate (FP_{rate}) indicates the fraction of compounds not existing in the ICSD and suggested by the model (i.e., false alarms):

$$FP_{rate}(\sigma) = \frac{FP(\sigma)}{N}, \tag{48}$$

where P is the number of compounds of proposed compounds non-existing in the ICSD but considered during cross-validation and $TP(\sigma)$ is the number of those non-existing compounds proposed by our model with a given threshold σ (i.e., the number of false positives).

High threshold values will lead to fewer false alarms but will imply fewer true hits. On the other hand lower threshold values give more true hits at the expense of

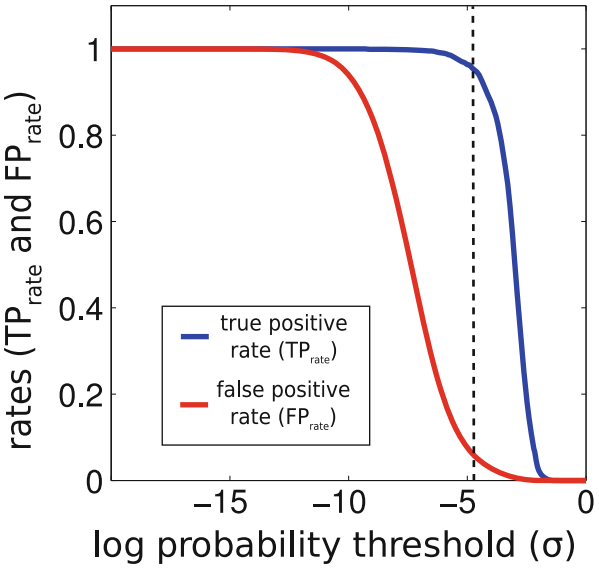


Fig. 10 True positive rate (TP_{rate}, *blue line*) and false positive rate (FP_{rate}, *red line*) in function of the probability threshold (σ) logarithm during cross-validation. Reprinted with permission from [87]. Copyright 2011 American Chemical Society

generating more false alarms. In practice, an adequate threshold is found by compromising between these two situations.

The clear separation between the two curves in Fig. 10 shows that the model is indeed predictive and can effectively distinguish between the substitutions leading to an existing compound and those leading to non-existing ones. Moreover, Fig. 10 can be used to estimate a value of probability threshold for a given true positive rate. For instance, the threshold required to find back 95% of the existing compounds during cross-validation is indicated in Fig. 10 by a dashed line.

6.5.2 Ionic Pair Substitution Analysis

The tendency for a pair of ions to substitute for each other can be estimated by computing the pair correlation:

$$g_{ab} = \frac{p\left(X_1 = a, X'_1 = b\right)}{p(X_1 = a)p(X_1 = b)} \tag{49}$$

$$= \frac{p\left(X_1 = a, X'_1 = b\right)}{\sum_j p\left(X_1 = a, X'_1 = x'_j\right) \sum_j p\left(X_1 = b, X'_1 = x'_j\right)} \tag{50}$$

$$= \frac{\frac{1}{Z} e^{\lambda_1^{a,b}}}{\frac{1}{Z} \sum_j e^{\lambda_1^{a,x'_j}} \frac{1}{Z} \sum_j e^{\lambda_1^{b,x'_j}}} \tag{51}$$

where a and b are two different ions and the sum represent a summation on all the possible values x'_j of the variable X'_1, i.e., a sum over all possible ionic species.

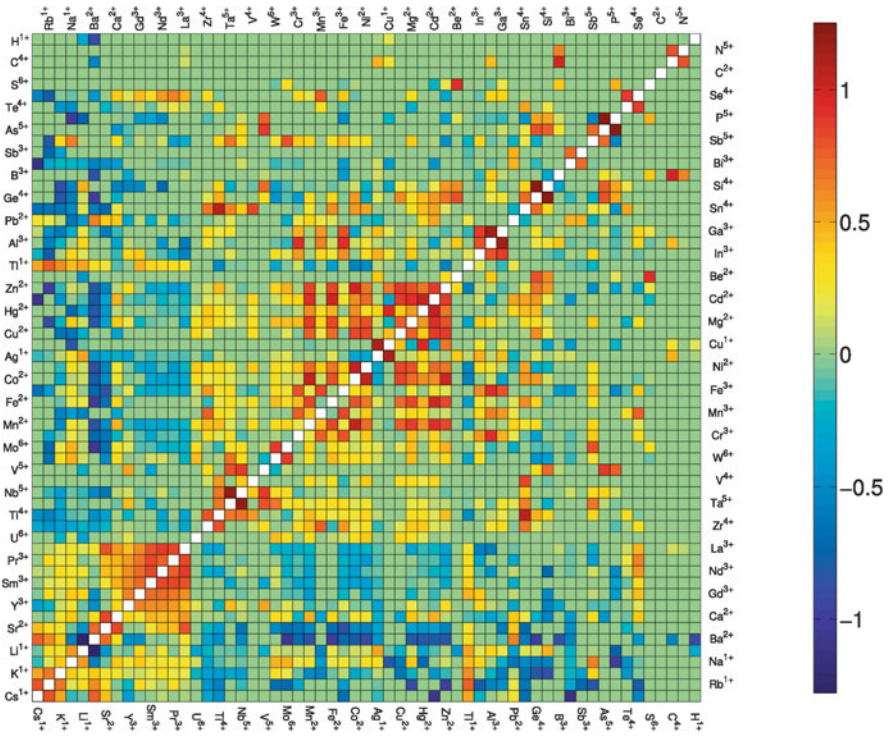


Fig. 11 Logarithm (base 10) of the pair correlation g_{ab} for each ion couple a, b. Equation (49) was used to evaluate the pair correlation g_{ab}. The ions are sorted according to their element's Mendeleev number. Only the 60 most common ions in the ICSD are presented in this graph. These correlation coefficients were obtained by training our probabilistic model on the ICSD. Positive values indicate a tendency to substitute while negative values, in contrast, show a tendency not to substitute. The symmetry of the pair correlation ($g_{ab} = g_{ba}$) is reflected in the symmetry of the matrix. Reprinted with permission from [87]. Copyright 2011 American Chemical Society

This pair correlation measures the increased probability to observe two ions at equivalent positions in a particular crystal structure over the probability to observe each of these ions in nature. Two ions which substitute well for each other will have a pair correlation higher than 1 ($g_{ab} > 1$) while ions which rarely substitute will have a pair correlation lower than one ($g_{ab} < 1$). The pair correlation is therefore a useful quantitative measure of the tendency for two ions to substitute for each other.

Figure 11 plots the logarithm (base10) of this pair correlation for the 60 most common cations in the ICSD (the pair correlation for all the ionic pairs is presented in supplementary information). Positive values indicate a tendency to substitute while negative values show a tendency not to substitute. The ions are sorted by their element Mendeleev number [65]. This ordering relates to their position in the periodic table. Therefore, the different ions are automatically clustered by chemical classes (alkali, alkali earth, rare earth, transition metals, and main group elements).

Different "blocks" of strong substitutional tendency are observed. For instance, the rare earth elements tend to substitute easily to each other. The similar charges (usually +3) and ionic size for those rare earth elements explain this strong substitution tendency.

The alkali elements also form a strongly substituting group. Only the ions with the largest size difference (Cs with Na or Li) do not substitute easily.

While transition metals in general tend to substitute easily for each other, two subgroups of strong pair correlation can be observed: the early transition metals (Zr^{4+}, Ti^{4+}, Ta^{5+}, Nb^{5+}, V^{4+}, V^{5+}, W^{6+}, Mo^{6+}) and late transition metals (Cr^{3+}, Mn^{2+}, Mn^{3+}, Fe^{2+}, Fe^{3+}, Co^{2+}, Ni^{2+}, Cu^{2+}, Hg^{2+}, Cd^{2+}, Zn^{2+}). This separation into two groups could be explained by a charge effect. The early transition metals have higher common oxidation states (+4 to +6) than the late ones (+2 to +3). Two notable exceptions to the general strong substitution tendency between transition metals are Ag^{1+} and Cu^{1+}. While substituting strongly for each other, those two ions do not substitute for any other transition metal. Indeed, electronic structure factors drive both ions to form very unusual linear environments [94].

On the other hand, the main group elements do not have a homogeneous strong substitution tendency across the entire chemical class. Only smaller subgroups such as Ga^{3+}, Al^{3+}, and In^{3+} or Si^{4+}, Ge^{4+}, and Sn^{4+} can be observed.

Regions of unfavorable substitutions are also present. Transition metals do not likely substitute for alkali or alkali earth metals. Only the smallest ions: Li^{1+}, Na^{1+}, and Ca^{2+} exhibit mild substitution tendencies for some transition metals. In addition, transition metals are very difficult to substitute for rare earths. Only Y^{3+} (and Sc^{3+} not shown in the figure) can substitute moderately with both rare earth and transition metals, indicating their ambivalent nature at the edge of these two very different chemistries.

Rare earth compounds do not substitute with main group elements with the surprising exception of Se^{4+}. Se^{4+} can occupy the high coordination sites that rare earth elements take in the very common Pnma perovskite structure formed by $MgSeO_3$, $CoSeO_3$, $ZnSeO_3$, $CrLaO_3$, $InLaO_3$, $MnPrO_3$, etc...

The oxidation state of an element can have a significant impact on whether an element will substitute for others. The two main oxidation states for antimony, Sb^{3+} and Sb^{5+}, behave very differently. The rather large +3 ion substitutes mainly with Pb^{2+} and Bi^{3+}, while the smaller +5 ion substitute preferentially with transition metals Mo^{6+}, Cr^{3+}, Fe^{3+}, etc.

Some ions tend to form very specific structures and local environments. Those ions will substitute only with very few others. For instance, C^{4+} almost only substitutes with B^{3+}. Both ions share a very uncommon tendency to form planar polyanions such as CO_3^{2-} and BO_3^{3-}. Hydrogen is an even more extreme example with no favorable substitution from H^{1+} (with the exception of a mild substitution with Cu^{1+}) to any other ion, in agreement with its very unique nature.

6.6 Limits and Strengths of the Model

The substitution model makes several simplifying assumptions. The absence of dependence on the number of components implies that, for instance, the substitution rules do not change if the compounds are ternaries or quaternaries. If Fe^{2+} is established to substitute easily for Ni^{2+} in ternary compounds, the same substitution should be likely in quaternaries.

In addition, the substitution rules do not depend on structural factors. In reality, how easy a chemical substitution is will depend somewhat on the specific structure. Some crystal structure sites will accommodate for instance a wider range of ions with different size without major distortion. Perovksites are a good example of structures where the specific size tolerance factor is established (see for instance Zhang et al. [95]). In some ways our model is "coarse grained" over structures.

The second major assumption is the use of binary features only. This implies that the substitution model only focuses on two substituted ions at a given site and does not take into account the "context" such as the other elements present in the crystal structure. Here again, a more accurate description will require this context to be taken into account. For instance, two cations might substitute in oxides but not in sulfides.

Those simplifying assumptions are, however, very useful in the sense that they allow the model to capture rules from data dense regions and use them to make predictions in data sparse regions. The substitution rules learned from ternary chemical systems can be used to predict compounds in the much less populated quaternary space. Likewise, substitution rules learned from very common crystal structure prototypes can be learned and used to make predictions in uncommon crystal structures. It is this capacity for this simpler model to make predictions in sparser data regions which constitutes its main advantage vs more powerful models such as that presented in Sect. 5.

Of course, our model could be refined in many ways. The most straightforward way to add structural factors would be to introduce a dependence on the ion local environment. The features could also be extended to go beyond binary features. Interesting work in feature selection has shown that complex features can be built iteratively from the data by combining very simple basic features [92].

The ionic substitution model has been used to search with high-throughput computing for novel multicomponent oxides and polyanionic systems (e.g., phosphates) in the field of Li-ion batteries [8, 38, 96, 97]. The technique has also been used recently to explore the field of oxynitrides for water splitting. The lack of knowledge of oxynitride chemistry justified relying heavily on data mining driven compound prediction [13].

7 From Computer to Synthesis: Examples of Successful Compound Prediction Through Data Mining

The ultimate success of a compound prediction technique is to lead to an experimental synthesis of the predicted phase. The theoretical approaches presented in this review chapter have already led to several successful syntheses of compounds suggested through computation. We will outline briefly (and not exhaustively) some of those successful predictions and describe their context.

7.1 Assigning a Structure to a Powder Diffraction Pattern

There are a significant number of compounds present in powder diffraction databases (e.g., the PDF4+ database [98]) that do not have any crystal structure assigned. This is an important issue, especially for computational materials science, as ab initio computations need a material's crystal structure to evaluate any property. Structure assignment from powder diffraction data, for instance by Rietveld refinement, needs a structural guess of the crystal structure that data mining crystal structure prediction algorithms can provide. In the large scale search for ternary oxides presented in Sect. 5, 355 compounds not present in the ICSD were suggested [78]. Of those 355 compounds, 64 compositions are present in a powder diffraction database but without any structural data associated with the ICSD. Figure 12 compares the simulated vs the experimental powder diffraction spectrum present in the PDF database for two predicted compounds: $MgMnO_3$ and $CoRb_2O_3$ (00-024-0736 [99] and 00-027-0515 [100]). Not only did the algorithm identify successfully the stoichiometries absent from the ICSD 2006 database (without data from the PDF database) but the computed and experimental patterns are in good agreement (if one takes into account the overestimation of the lattice constant by a few percent present with DFT computations in the generalized gradient approximation). Only one peak in the 50° region does not match the powder diffraction pattern for $MgMnO_3$.

These two examples show that a purely data mining driven approach based on no human intervention can successfully assign crystal structure to powder diffraction patterns.

7.2 $SnTiO_3$

Among the compounds without any data available (even powder diffraction data), the large scale data mined ternary oxide search presented in Sect. 5 found $SnTiO_3$ to be a stable stoichiometry with an ilmenite structure being the most stable phase. This $SnTiO_3$ ilmenite prediction is of technological interest as $SnTiO_3$ perovskite has been predicted through ab initio computation to be a good candidate Pb-free

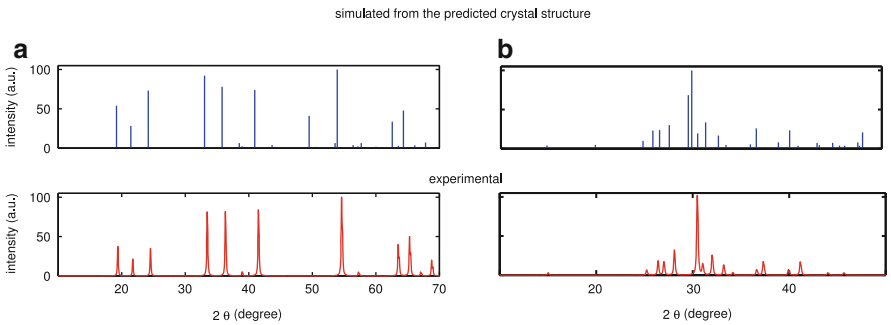


Fig. 12 Comparison between the predicted (*above*) and the experimental (from PDF4+ database, *below*) powder diffraction patterns for $MgMnO_3$ (**a**) and $CoRb_2O_3$ (**b**)

ferroelectric material [101]. Unfortunately, the interesting piezoelectric properties are only present for the perovskite structure. The synthesis of $SnTiO_3$ had been unsuccessful at the time of publication of the paper on ternary oxides but was reported very shortly after by Fix et al. [102]. The experimental results very clearly confirm the computed prediction of an ilmenite phase. Not only is this example a success of computational prediction but it illustrates how important it is to study the stability of the phases that are used to make materials properties prediction in the ab initio literature.

7.3 $Li_9V_3(P_2O_7)_3(PO_4)_2$

Finding novel cathodes for Li-ion batteries is of great importance for energy storage [103–105]. Using the possibility to predict important battery properties by ab initio computations (voltage, Li-ion diffusion, stability when charged) [106, 107], a high-throughput computational search for new cathode materials has been performed by Ceder et al. This project made extensive use of some of the data mining based compound prediction approaches that have been previously described.

During this high-throughput study, an entirely novel phase – $Li_9V_3(P_2O_7)_3(PO_4)_2$ – was predicted by the ionic substitution approach suggesting that a substitution of Fe^{3+} to V^{3+} in $Li_9Fe_3(P_2O_7)_3(PO_4)_2$ leads to a compound lying low in energy [8, 108]. This example shows how unusual structures, beyond the common spinels, rock salt, ilmenite etc., can also be suggested by data mining approaches and lead to technologically relevant materials.

We should note that an independent report on this phase by Kuang et al. [109] had appeared in the literature. However, the patent anteriority date from the Ceder team (before Kuang et al.'s publication) clearly confirms the true predictive nature of the result.

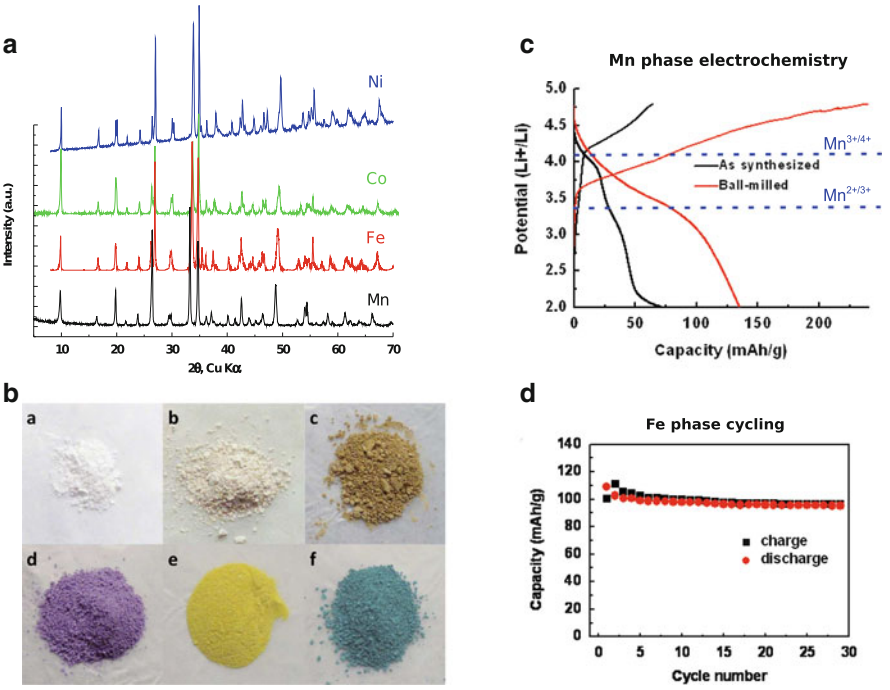


Fig. 13 XRD patterns (**a**) and powders (**b**) of first-time synthesized $Na_3M(CO_3)(PO_4)$ with M = Mn, Ni, Fe, Co, etc. The electrochemical activity (voltage vs capacity) of the Mn-based Li version $Li_3Mn(CO_3)(PO_4)$ (**c**) and the cyclability of the $Li_3Fe(CO_3)(PO_4)$ phase (**d**). Adapted with permission from [33] and [110]. Copyright 2012 American Chemical Society

7.4 *Sidorenkite*

The high-throughput cathode project also led to the identification of an even more exotic class of materials: the sidorenkite carbonophosphates [33, 38, 110]. Carbonophosphates had only been known as rare minerals but were identified by high-throughput computations to form very promising lithium-ion battery cathodes. The predicted compounds were then synthesized by hydrothermal reaction followed by ion exchange as suggested by computational phase stability analysis. Some carbonophosphates have shown electrochemical activity and very good cyclability as Li-ion battery cathode (see Fig. 13c, d).

7.5 *$LiCoPO_4$*

Compound prediction can also push for the reinvestigation of chemical systems that were believed to be very well known. In their high-throughput phosphate analysis, Hautier et al. made the surprising observation that data mining and DFT suggested

a polymorph of the well studied $LiCoPO_4$ olivine structure [8]. While $LiCoPO_4$ olivine incorporates Co coordinated by octahedra of oxygen, the new predicted polymorph shows the structure of $LiZnPO_4$ based on tetrahedral Co. The prediction was confirmed by Jähne et al. when they reported on the first synthesis of tetrahedral $LiCoPO_4$ in the structure that was suggested computationally [111].

8 Conclusion and Future Avenues

Materials science is moving more and more towards computationally oriented materials design. Compound and crystal structure prediction is a critical step in this new paradigm. Current DFT techniques are mature enough to model the phase stability reasonably well and different approaches to compound predictions have been developed. Among them, data mining offers high-throughput-friendly, efficient methods that have already been used in several fields from Li-ion batteries to oxynitrides for water splitting. We not only presented these methods in details but also reported on several successes where computational predictions were confirmed by experimental synthesis.

In the future, the development of large databases of freely available computed data such as the Materials Project will surely help in providing large data sets to be used for fitting more efficient data mining crystal structure prediction models. We can expect an improvement in the predictive power of data mining based techniques as the models are refined and the data sets become larger.

However, the main limitation of data mining techniques is their inability to predict (in contrast to optimization techniques such as genetic algorithms) crystal structures that have never been observed before. Combination of optimization and data mining approaches could offer a solution to this problem, aiming at keeping the low computational budget of knowledge-based methods while approaching the exhaustivity of the optimization approaches.

We hope the many compound prediction techniques available and the current understanding of the accuracy of phase stability prediction will in the future make phase stability a more central part of the computational materials design process. Too often new phases with exceptional computed properties are proposed without assessing their phase stability.

Finally, while computations can be truly predictive to determine the existence of an inorganic phase, the step between computational compound prediction and finding the most appropriate synthesis route is still very empirical. A better fundamental understanding of the different synthesis approaches (solid state reaction, hydrothermal, etc.) needing a joint effort from experimentalists and theorists would be of great value here.

References

1. Kohn W, Sham L (1965) Self-consistent equations including exchange and correlation effects. Phys Rev 140(4A):1131–1138
2. ABINIT (2004). http://www.abinit.org/. Accessed 1 July 2013
3. Vienna ab initio simulation package (VASP). http://www.vasp.at/. Accessed 1 July 2013
4. Quantum Espresso (2012). http://www.quantum-espresso.org/. Accessed 1 July 2013
5. Hautier G, Jain A, Ong SP (2012) From the computer to the laboratory: materials discovery and design using first-principles calculations. J Mater Sci 47(21):7317–7340
6. Curtarolo S, Hart GLW, Nardelli MB, Mingo N, Sanvito S, Levy O (2013) The high-throughput highway to computational materials design. Nat Mater 12(3):191–201
7. Greeley J, Jaramillo TF, Bonde J, Nørskov JK, Chorkendorff IB (2006) Computational high-throughput screening of electrocatalytic materials for hydrogen evolution. Nat Mater 5(11):909–913
8. Hautier G, Jain A, Ong SP, Kang B, Moore C, Doe R, Ceder G (2011) Phosphates as lithium-ion battery cathodes: an evaluation based on high-throughput ab initio calculations. Chem Mater 23:3495–3508
9. Mueller T, Hautier G, Jain A, Ceder G (2011) Evaluation of tavorite-structured cathode materials for lithium-ion batteries using high-throughput computing. Chem Mater 23:3854–3862
10. Setyawan W, Gaume RM, Lam S, Feigelson RS, Curtarolo S (2011) High-throughput combinatorial database of electronic band structures for inorganic scintillator materials. ACS Comb Sci 13(4):382–390
11. Castelli IE, Olsen T, Datta S, Landis DD, Dahl S, Thygesen KS, Jacobsen KW (2012) Computational screening of perovskite metal oxides for optimal solar light capture. Energy Environ Sci 5(2):5814
12. Jain A, Castelli IE, Hautier G, Bailey DH, Jacobsen KW (2013) Performance of genetic algorithms in search for water splitting perovskites. J Mater Sci 48:6519–6534
13. Wu Y, Lazic P, Hautier G, Persson K, Ceder G (2013) First principles high throughput screening of oxynitrides for water-splitting photocatalysts. Energy Environ Sci 6:157–168
14. Madsen GKH (2006) Automated search for new thermoelectric materials: the case of LiZnSb. J Am Chem Soc 128(37):12140–12146
15. Wang S, Wang Z, Setyawan W, Mingo N, Curtarolo S (2011) Assessing the thermoelectric properties of sintered compounds via high-throughput ab-initio calculations. Phys Rev X 1(2):021012
16. Jain A, Seyed-Reihani SA, Fischer CC, Couling DJ, Ceder G, Green WH (2010) Ab initio screening of metal sorbents for elemental mercury capture in syngas streams. Chem Eng Sci 65(10):3025–3033
17. Olivares-Amaya R, Amador-Bedolla C, Hachmann J, Atahan-Evrenk S, Sánchez-Carrera RS, Vogt L, Aspuru-Guzik A (2011) Accelerated computational discovery of high-performance materials for organic photovoltaics by means of cheminformatics. Energy Environ Sci 4:4849–4861
18. Yang K, Setyawan W, Wang S, Buongiorno Nardelli M, Curtarolo S (2012) A search model for topological insulators with high-throughput robustness descriptors. Nat Mater 11(7):614–619
19. Materials project. http://www.materialsproject.org. Accessed 1 July 2013
20. Jain A, Hautier G, Moore CJ, Ping Ong S, Fischer CC, Mueller T, Persson KA, Ceder G (2011) A high-throughput infrastructure for density functional theory calculations. Comp Mater Sci 50:2295–2310
21. AFLOWLIB: http://www.aflowlib.org. Accessed 1 July 2013
22. "The Electronic Structure Project", http://gurka.fysik.uu.se/ESP/. Accessed 1 July 2013
23. Service RF (2012) Materials scientists look to a data-intensive future. Science 335:1434–1435

24. Inorganic Crystal Structure Database (ICSD), http://www.fiz-karlsruhe.de/icsd.html, Accessed 1 July 2013
25. Maddox J (1988) Crystals from first principles. Nature 335:201
26. O'Keeffe M (2010) Aspects of crystal structure prediction: some successes and some difficulties. Phys. Chem. Chem. Phys. 12:10–15
27. Woodley SM, Catlow R (2008) Crystal structure prediction from first principles. Nat Mater 7(12):937–946
28. Callen HB (1985) Thermodynamics and an introduction to thermostatistics. Wiley, New York
29. Chandler D (1987) Introduction to modern statistical mechanics. Oxford University Press, Oxford
30. Ceder G, Ven A, Marianetti C, Morgan D (2000) First-principles alloy theory in oxides. Modelling Simul. Mater. Sci. Eng. 8:311–321
31. Van De Walle A, Ceder G (2000) First-principles computation of the vibrational entropy of ordered and disordered Pd_3V. Phys Rev B 61(9):5972–5978
32. Zhou F, Maxisch T, Ceder G (2006) Configurational electronic entropy and the phase diagram of mixed-valence oxides: the case of Li_xFePO_4. Phys Rev Lett 97:155704
33. Chen H, Hautier G, Ceder G (2012) Synthesis, computed stability and crystal structure of a new family of inorganic compounds: carbonophosphates. J Am Chem Soc 134(48):19619–19627
34. Ong SP, Richards WD, Jain A, Hautier G, Kocher M, Cholia S, Gunter D, Chevrier VL, Persson KA, Ceder G (2013) Python materials genomics (pymatgen): a robust, open-source python library for materials analysis. Comp Mater Sci 68:314–319
35. Ong SP, Wang L, Kang B, Ceder G (2008) Li-Fe-P-O_2 phase diagram from first principles calculations. Chem Mater 20(5):1798–1807
36. Curtarolo S, Morgan D, Ceder G (2005) Accuracy of methods in predicting the crystal structures of metals: a review of 80 binary alloys. CALPHAD 29(3):163–211
37. Lany S (2008) Semiconductor thermochemistry in density functional calculations. Phys Rev B 78(24):245207
38. Hautier G, Ong SP, Jain A, Moore CJ, Ceder G (2012) Accuracy of density functional theory in predicting formation energies of ternary oxides from binary oxides and its implication on phase stability. Phys Rev B 85:155208
39. Dudarev SL, Savrasov SY, Humphreys CJ, Sutton AP (1998) Electron-energy-loss spectra and the structural stability of nickel oxide: an LSDA+U study. Phys Rev B 57(3):1505–1509
40. Zhou F, Cococcioni M, Marianetti CA, Morgan D, Ceder G (2004) First-principles prediction of redox potentials in transition-metal compounds with LDA+U. Phys Rev B 70:235121
41. Jain A, Hautier G, Ong SP, Moore CJ, Fischer CC, Persson KA, Ceder G (2011) Formation enthalpies by mixing GGA and GGA+*U* calculations. Phys Rev B 84:045115
42. Stevanović V, Lany S, Zhang X, Zunger A (2012) Correcting density functional theory for accurate predictions of compound enthalpies of formation: fitted elemental-phase reference energies. Phys Rev B 85:115104
43. Oganov AR, Valle M (2009) How to quantify energy landscapes of solids. J Chem Phys 130(10):104504
44. Ceder G (1993) A derivation of the Ising model for the computation of phase diagrams. Comp Mater Sci 1(2):144–150
45. Ducastelle F (1991) Order and phase stability in alloys, volume 3 (cohesion and structure). North Holland, Amsterdam
46. Sanchez JM, Ducastelle F, Gratias D (1984) Generalized cluster description of multicomponent systems. Physica A 128:334–350
47. Blum V, Zunger A (2004) Structural complexity in binary bcc ground states: the case of bcc Mo-Ta. Phys Rev B 69(2):20103
48. Hart GLW (2009) Verifying predictions of the $L1_3$ crystal structure in Cd-Pt and Pd-Pt by exhaustive enumeration. Phys Rev B 80(1):014106

49. Sanati M, Wang L, Zunger A (2003) Adaptive crystal structures: CuAu and NiPt. Phys Rev Lett 90(4):045502
50. Van Der Ven A, Aydinol MK, Ceder G (1998) First-principles evidence for stage ordering in Li_xCoO_2. J Electrochem Soc 145(6):2149
51. Wales DJ, Doye JPK (1997) Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. J Phys Chem A 101(28):5111–5116
52. Wales DJ, Scheraga HA (1999) Global optimization of clusters, crystals, and biomolecules. Science 285(5432):1368–1372
53. Abraham NL, Probert MIJ (2006) A periodic genetic algorithm with real-space representation for crystal structure and polymorph prediction. Phys Rev B 73(22):224104
54. Bush TS, Catlow CRA, Battle PD (1995) Evolutionary programming techniques for predicting inorganic crystal structures. J Mater Chem 5(8):1269–1272
55. Oganov AR, Glass CW (2006) Crystal structure prediction using ab initio evolutionary techniques: principles and applications. J Chem Phys 124(24):244704
56. Oganov AR, Glass CW (2008) Evolutionary crystal structure prediction as a tool in materials design. J Phys Condens Matter 20(6):064210
57. Trimarchi G, Zunger A (2007) Global space-group optimization problem: finding the stablest crystal structure without constraints. Phys Rev B 75(10):104113
58. Zhang X, Zunger A, Trimarchi G (2010) Structure prediction and targeted synthesis: a new $Na_{(n)}N_2$ diazenide crystalline structure. J Chem Phys 133(19):194504
59. Oganov AR, Chen J, Gatti C, Ma Y, Ma Y, Glass CW, Liu Z, Yu T, Kurakevych OO, Solozhenko VL (2009) Ionic high-pressure form of elemental boron. Nature 457 (February):863–868
60. Kolmogorov A, Shah S, Margine E, Bialon A, Hammerschmidt T, Drautz R (2010) New superconducting and semiconducting Fe-B compounds predicted with an ab initio evolutionary search. Phys Rev Lett 105(21):217003
61. Ono S, Kikegawa T, Ohishi Y (2007) High-pressure transition of $CaCO_3$. Am Mineral 92(7):1246–1249
62. Gou H, Dubrovinskaia N, Bykova E, Tsirlin AA, Kasinathan D, Richter A, Merlini M, Hanfland M, Abakumov AM, Batuk D, Van Tendeloo G, Nakajima Y, Kolmogorov AN, Dubrovinsky L (2013) Discovery of a superhard iron tetraboride superconductor. Phys Rev Lett 111:157002
63. Liebold-Ribeiro Y, Fischer D, Jansen M (2008) Experimental substantiation of the "energy landscape concept" for solids: synthesis of a new modification of LiBr. Angew Chem Int Edit 47(23):4428–4431
64. Pauling L (1929) The principles determining the structure of complex ionic crystals. J Am Chem Soc 51:1010–1026
65. Pettifor DG (1990) Structure maps in alloy design. J Chem Soc Faraday Trans 86 (8):1209–1213
66. Pettifor DG (2003) Structure maps revisited. J Phys Condens Matter 15:13–16
67. Villars P (1983) A three-dimensional structural stability diagram for 998 binary AB intermetallic compounds. J Less Common Met 92(2):215–238
68. Morgan D, Rodgers J, Ceder G (2003) Automatic construction, implementation and assessment of Pettifor maps. J Phys Condens Matter 15:4361–4369
69. Ceder G, Morgan D, Fischer C, Tibbetts K, Curtarolo S (2006) Data-mining-driven quantum mechanics for the prediction of structure. MRS Bull 31:981–985
70. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction. 2nd edn. (Springer Series in Statistics), Springer, chap 4, pp 80–113
71. von Lilienfeld OA (2013) First principles view on chemical compound space: gaining rigorous atomistic control of molecular properties. Int J Quantum Chem 113(12):1676–1689
72. Rupp M, Tkatchenko A, Müller KR, von Lilienfeld OA (2012) Fast and accurate modeling of molecular atomization energies with machine learning. Phys Rev Lett 108:058301

73. Curtarolo S, Morgan D, Persson K, Rodgers J, Ceder G (2003) Predicting crystal structures with data mining of quantum calculations. Phys Rev Lett 91(13):135503
74. Kolmogorov AN, Curtarolo S (2006) Prediction of different crystal structure phases in metal borides: a lithium monoboride analog to MgB_2. Phys Rev B 73(18):180501
75. Kolmogorov AN, Curtarolo S (2006) Theoretical study of metal borides stability. Phys Rev B 74(22):224507
76. Levy O, Chepulskii RV, Hart GLW, Curtarolo S (2009) The new face of rhodium alloys: revealing ordered structures from first principles. J Am Chem Soc 132(2):833–837
77. Fischer CC, Tibbetts KJ, Morgan D, Ceder G (2006) Predicting crystal structure by merging data mining with quantum mechanics. Nat Mater 5(8):641–646
78. Hautier G, Fischer CC, Jain A, Mueller T, Ceder G (2010) Finding nature's missing ternary oxide compounds using machine learning and density functional theory. Chem Mater 22(12):3762–3767
79. Hundt R, Schön JC, Jansen M (2006) CPMZ-an algorithm for the efficient comparison of periodic structures. J Appl Crystallogr 39:6–16
80. Morita T (1957) Cluster variation method of cooperative phenomena and its generalization I. J Phys Soc Jpn 12(7):753–755
81. Fischer CC (2007) A machine learning approach to crystal structure prediction. PhD thesis, Massachusetts Institute of Technology
82. Eliason SR (1993) Maximum likelihood estimation: logic and practice. Sage Publications, Inc, Newberry Park
83. Jaynes ET (2003) Probability theory: the logic of science. Cambridge University Press, Cambridge
84. Buntine W (1991) Theory refinement on Bayesian networks. In: Proceedings of the seventh conference on uncertainty in artificial intelligence, Citeseer 91:52–60
85. Lynch RSJ, Willett PK (2003) Adaptive Bayesian classification using noninformative Dirichlet priors. IEEE Trans Syst Man Cybern 33(3):2812–2815
86. Ternary oxides predictions. http://ceder.mit.edu/ternaryoxides, accessed: 01 July 2013
87. Hautier G, Fischer C, Ehrlacher V, Jain A, Ceder G (2011) Data mined ionic substitutions for the discovery of new compounds. Inorg Chem 50:656–663
88. Johrendt D, Pöttgen R (2008) Pnictide oxides: a new class of high-T_C superconductors. Angew Chem Int Edit 47(26):4782–4784
89. Goldschmidt V (1926) Die gesetze der krystallochemie. Naturwissenschaften 14:477–485
90. Brown PF, Della Pietra SA, Della Pietra VJ, Mercer RL (1993) The mathematics of statistical machine translation: parameter estimation. Comput Linguist 19:263–312
91. Berger A, Della Pietra VJ, Della Pietra SA (1996) A maximum entropy approach to natural language processing. Comput Linguist 22(1):39–72
92. Della Pietra SA, Della Pietra VJ, Lafferty J (1997) Inducing features of random fields. IEEE Trans Pattern Anal Mach Intell 19(4):1–13
93. Parthé E, Gelato L (1984) The standardization of inorganic crystal-structure data. Acta Crystallogr A 40:169–183
94. Gaudin E, Boucher F, Evain M (2001) Some factors governing Ag^+ and Cu^+ Low coordination in chalcogenide environments. J Solid State Chem 160(1):212–221
95. Zhang H, Li N, Li K, Xue D (2007) Structural stability and formability of ABO3-type perovskite compounds. Acta Crystallogr Sec B 63:812–818
96. Jain A, Hautier G, Moore CJ, Kang B, Lee J, Chen H, Twu N, Ceder G (2012) A computational investigation of $Li_9M_3(P_2O_7)_2(PO_4)_2$ (M=V, Mo) as cathodes for Li ion batteries. J Electrochem Soc 159(5):A622–A633
97. Ma X, Hautier G, Jain A, Doe R, Ceder G (2013) Improved capacity retention for $LiVO_2$ by Cr substitution. J Electrochem Soc 160(2):A279–A284
98. International centre for diffraction data. PDF4+ database. http://www.icdd.com/products/pdf4.htm. Accessed 1 July 2013
99. Chamberland B, Sleight AW, Weiher JF (1970) Preparation and characterization of $MgMnO_3$ and $ZnMnO_3$. J Solid State Chem 1(3–4):512–514

100. Jansen M, Hoppe R (1974) Neue oxocobaltate (IV):$Cs_2[CoO_3]$, $Rb_2[CoO_3]$ und $K_2[CoO_3]$. Z Anorg Allg Chem 408:75–82
101. Matar S, Baraille I, Subramanian M (2009) First principles studies of $SnTiO_3$ perovskite as potential environmentally benign ferroelectric material. Chem Phys 355(1):43–49
102. Fix T, Sahonta SL, Garcia V, MacManus-Driscoll JL, Blamire MG (2011) Structural and dielectric properties of $SnTiO_3$, a putative ferroelectric. Crystal Growth Des 11:1422–1426
103. Ellis BL, Lee KT, Nazar LF (2010) Positive electrode materials for Li-ion and Li-batteries. Chem Mater 22(3):691–714
104. Goodenough JB, Kim Y (2010) Challenges for rechargeable Li batteries. Chem Mater 22(3):587–603
105. Whittingham MS (2004) Lithium batteries and cathode materials. Chem Rev 104(10):4271–4302
106. Ceder G, Hautier G, Jain A, Ong SP (2011) Recharging lithium battery research with first-principles methods. MRS Bull 36(3):185–191
107. Meng YS, Arroyo-de Dompablo ME (2013) Recent Advances in First Principles Computational Research of Cathode Materials for Lithium-Ion Batteries, Acc Chem Res, 46 (5):1171–1180
108. Ceder G, Jain A, Hautier G, Kim JC, Kang B, Daniel R (2013) Mixed phosphate-diphosphate electrode materials and methods of manufacturing same US8399130 B2
109. Kuang Q, Xu J, Zhao Y, Chen X, Chen L (2011) Layered monodiphosphate $Li_9V_3(P_2O_7)_3(PO_4)_2$: a novel cathode material for lithium-ion batteries. Electrochim Acta 56(5):2201–2205
110. Chen H, Hautier G, Jain A, Moore C, Kang B, Doe R, Wu L, Zhu Y, Tang Y, Ceder G (2012) Carbonophosphates: a new family of cathode materials for Li-ion batteries identified computationally. Chem Mater 24(11):2009–2016
111. Jähne C, Neef C, Koo C, Meyer HP, Klingeler R (2013) A new $LiCoPO_4$ polymorph via low temperature synthesis. J Mater Chem A 1(8):2856

Top Curr Chem (2014) 345: 181–222
DOI: 10.1007/128_2013_489
© Springer-Verlag Berlin Heidelberg 2014
Published online: 11 February 2014

Structure and Stability Prediction of Compounds with Evolutionary Algorithms

Benjamin C. Revard, William W. Tipton, and Richard G. Hennig

Abstract Crystal structure prediction is a long-standing challenge in the physical sciences. In recent years, much practical success has been had by framing it as a global optimization problem, leveraging the existence of increasingly robust and accurate free energy calculations. This optimization problem has often been solved using evolutionary algorithms (EAs). However, many choices are possible when designing an EA for structure prediction, and innovation in the field is ongoing. We review the current state of evolutionary algorithms for crystal structure and composition prediction and discuss the details of methodological and algorithmic choices. Finally, we review the application of these algorithms to many systems of practical and fundamental scientific interest.

Keywords Structure prediction · Genetic algorithm · Ccrystal structure · Energy landscape · Heuristic optimization · Phase diagram

Contents

B.C. Revard, W.W. Tipton, and R.G. Hennig (✉)
Department of Materials Science and Engineering, Cornell University, Ithaca, 14853 New York, USA
e-mail: rhennig@cornell.edu.

1 Introduction

Many of the most crucial technological challenges today are essentially materials problems. Materials with specific properties are needed but unknown, and new materials must be found or designed. In some cases experiments can be performed to search for and characterize new materials [1], but these methods can be expensive and difficult. Thus, computational approaches can be complementary or advantageous. Theoretical prediction of many materials properties is possible once the atomic structure of a material is known, but structure prediction remains a challenge. However, a number of new methods have been proposed in recent years to address this problem [2–8]. These techniques are often faster and less expensive than experimental work, they dispense with the need to work with sometimes toxic chemicals, and they can be used to explore materials systems under conditions that are still inaccessible to experiment, such as very high pressures.

Unless kinetically constrained, materials tend to form structures that are in thermodynamic equilibrium, i.e., have the lowest Gibbs free energy. Thus, in order to predict a material's structure, we must find the arrangement of atoms that minimizes the Gibbs free energy, given by

$$G = U - TS + pV$$

Here, U is the internal energy, p the pressure, V the volume, T the temperature, and S the entropy. The entropy is comprised of three contributions: electronic, vibrational, and configurational. The vibrational and configurational components are expensive to calculate, and much of the error introduced by neglecting the

entropy vanishes when taking energy differences [9, 10]. For these reason, the entropy is often neglected, effectively constraining the search to the $T = 0$ regime. That is, the enthalpy $H = U + pV$ is frequently used to approximate the Gibbs free energy. Finite temperature effects can be included as a post-processing step once particularly promising structures have been identified. We note that, at high temperatures, anharmonic contributions to the vibrational entropy can stabilize phases that are mechanically or dynamically unstable at low temperature [11]. However, in order to search for stable materials at low temperature and fixed composition, the function we need to minimize, known as the objective function, is the enthalpy per atom.

A thermodynamic ensemble is not always used as the objective function. Bush et al. devised an objective function based on Pauling's valence rules and only performed energy calculations on the best structures identified thereby [2, 12]. Although this approach is computationally efficient, it is limited to ionic materials and is not as reliable as a direct search over the correct thermodynamic quantity.

1.1 Potential Energy Landscape

Given the atomic structure, there exist efficient methods for approximating the enthalpy. A complete description of a crystal structure includes six lattice parameters and $3N - 3$ atomic coordinates, where N is the number of atoms in the unit cell. Thus, the function we seek to minimize can be thought of as a surface in a $3N + 3$-dimensional space. These surfaces are referred to as energy landscapes. The lowest enthalpy structure is located at the deepest, or global, minimum of the energy landscape. In this way, the physical problem of predicting a material's atomistic structure is expressed as a mathematical optimization problem. In order to understand the search for the global minimum of an energy landscape, it is helpful to examine some general properties of energy landscapes of materials, as follows:

- Much of the configuration space corresponds to structures with unphysical small interatomic distances. These areas of the configuration space can be neglected.
- The energy landscape is effectively partitioned into basins of attraction by the use of a local optimization routine. The local optimizer takes any two structures in the same basin into the local minimum located at the bottom of the basin.
- The number of local minima on the energy landscape scales exponentially with the dimensionality of the search space, i.e., with the number of atoms in the cell [13]. Venkatesh et al. calculated the number of local minima as a function of system size for clusters containing up to 14 Lennard–Jones particles, illustrating this exponential trend [14].
- Deeper basins tend to occupy larger volumes in the multidimensional space. Specifically, a power law distribution describes the relationship between the depth of a basin and its hyper-volume. Combined with our capability for local minimization, this greatly simplifies the search for the optimum structure [15].

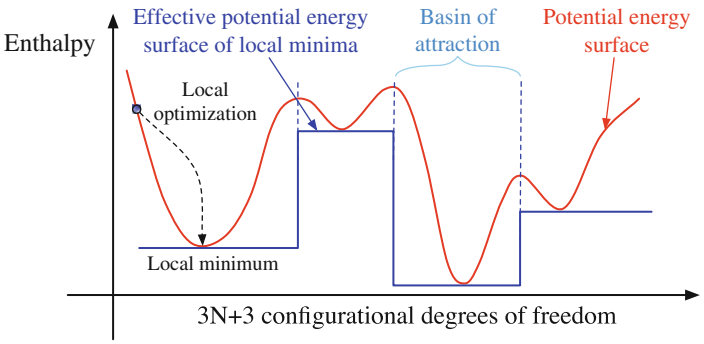


Fig. 1 Potential energy surface. The use of local optimization simplifies the search problem by dividing the continuous solution space into basins of attraction

- The barrier to reach a neighboring basin is usually low if that basin has a deeper minimum than the current basin. This is a consequence of the Bell–Evans–Polanyi principle [7].
- Low-energy minima in the landscape usually correspond to symmetrical structures [7].
- Low-lying minima are usually located near each other on the energy landscape. This tendency gives the landscape an overall structure that can be exploited while searching for the global minimum [16].

No analytical form exists for the enthalpy as a function of atomic configuration. We can only sample the enthalpy and its derivatives at discrete points on the energy landscape using methods such as density functional theory (DFT). Thus, one often resorts to heuristic search methods. One such class of methods that has proven successful is the evolutionary algorithm. This approach draws inspiration from biological evolution. Efficient local optimization utilizes the derivative information and is very beneficial in the solution of the optimization problem (see Sect. 4). Figure 1 illustrates how local optimization transforms the continuous potential energy landscape into a discrete set of basins of attractions, which dramatically simplifies the search space.

1.2 Evolutionary Algorithms

In nature, genetic information is carried in organisms. It is maintained in a population's gene pool if it is passed on from parents to offspring. New information can be introduced through mutation events, but these are rare (and usually lethal). The success that an organism has in passing on its genes is called the organism's fitness.

The fitness of an organism is not universal but depends on its environment. Many species which are very successful in their native habitats would do poorly in other

environments. More subtly, there is variance of traits within a single species. In some cases these differences can lead to a difference in the organisms' fitness. The genes of low-fitness individuals are less likely to be passed on, so traits of the high-fitness individuals are likely to be more common in subsequent generations. In this way, populations (but not individuals) evolve to be well suited to their environment. This assumes, of course, that relevant traits are passed on, to varying degrees, from parents to offspring. The correlation between a trait in a parent and that in an offspring is known as the heritability of a trait. In order for environmental pressure to cause quick evolution of a trait, that trait must have high heritability.

Evolutionary algorithms leverage the power of this process to "evolve" solutions to optimization problems. Initial efforts to apply evolutionary algorithms to the structure prediction problem were aimed at finding the lowest energy conformation of large organic molecules [17–21]. Evolutionary search techniques were also successfully applied to atomic clusters [22, 23], and soon the method was extended to 3D periodic systems [2, 12, 24].

It has been observed that evolutionary algorithms (EAs) are well suited to the structure prediction problem for several reasons [25]. First, they can efficiently find the global minimum of multidimensional functions. EAs require little information and few assumptions about the lowest energy structure, which is advantageous when searching for structures about which little is known a priori. Finally, if designed correctly, an EA can take advantage of the structure of the energy landscape discussed in Sect. 1.1.

The evolutionary approach to structure prediction is modeled after the natural process. Each crystal structure is analogous to a single organism. In nature, the fitness of an organism is based on how well its phenotype is suited to its environment and, in particular, how successful it is in reproducing. In an evolutionary algorithm, fitnesses are assigned to the organisms based on their objective function values, and they are allowed to reproduce based on those fitnesses. Pressures analogous to those which force species to adapt to their environments will thus lead to crystal structures with lower energies.

Organisms in an EA are often grouped into generations. The algorithm proceeds by creating successive generations. The methods by which an offspring generation is made from parents are called variation operations or variations and include operations that are analogous to genetic mutation and crossover. A single offspring organism can be created from either one or two parents, depending on which variation is used. Every offspring organism must meet some minimum standards to be considered viable, analogous to the "growing up" process in nature. This is known as the development process. The algorithm terminates when some user-defined stopping criteria are met (see Sect. 2.13).

Improvements are made to the biological analogy when possible. In particular, we would rather not let the optimal solution worsen from one algorithmic iteration to the next. To prevent this, a promotion operator is used to advance some of the best organisms from one generation directly to the next. Also, mutations in nature are usually detrimental. When searching for structures, one might try to use mutation variations that are likely to introduce valuable new information to the gene pool.

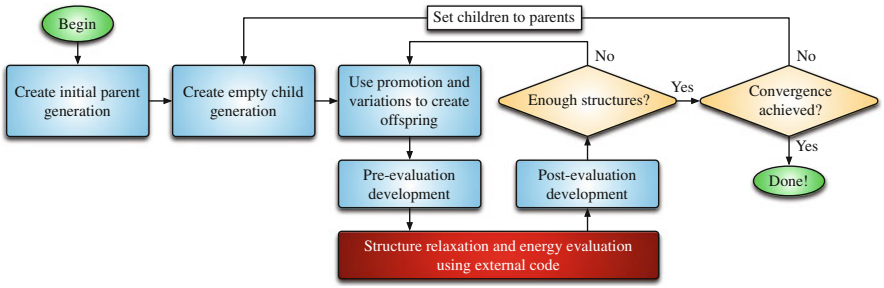


Fig. 2 Outline of evolutionary algorithm for structure prediction

Figure 2 outlines how a typical evolutionary algorithm for structure prediction proceeds, such as that implemented by Tipton et al. [26, 27]. The EA starts by creating an initial population (Sect. 2.2) and calculating the fitness of each organism in it (Sect. 2.3). Organisms are then selected to act as parents (Sect. 2.4) or to be promoted to the next generation (Sect. 2.5). The parents create offspring structures via mating (Sect. 2.6) or mutation (Sect. 2.7). The offspring are developed (Sect. 2.9) and the energy of each offspring organism is calculated using some external energy code, followed by a post-evaluation development step. If successful, the offspring structures are then added to the next generation, and their fitnesses are calculated. Unless the EA has converged (Sect. 2.13), the current children become parents in the next generation.

2 Details of the Method

2.1 Representation of Structures

A total of $3N + 3$ dimensions describe the atomic coordinates and lattice vectors of a crystal structure. Additionally, the number of atoms, N, itself must be determined for ab initio structure predictions. However, these degrees of freedom are not all truly independent. Alternate choices of lattice vectors provide infinitely many ways to represent the same crystal structure, as illustrated in Fig. 3. Additionally, for molecular crystals, the dimensionality of the search space is effectively reduced since the molecular units typically stay intact in these crystal structures. This is due to the separation of energy scales, with strong intramolecular covalent interactions and much weaker intermolecular van der Waals interactions. In this case, structure search algorithms can take advantage of this trait by treating complete molecules, instead of individual atoms, as indivisible structural building blocks [28, 29]. Since the solution space is somewhat more complicated than it has to be, the task of searching that space is also more complicated than necessary. This difficulty may be addressed by attempting to standardize the way structures are represented in the computer.

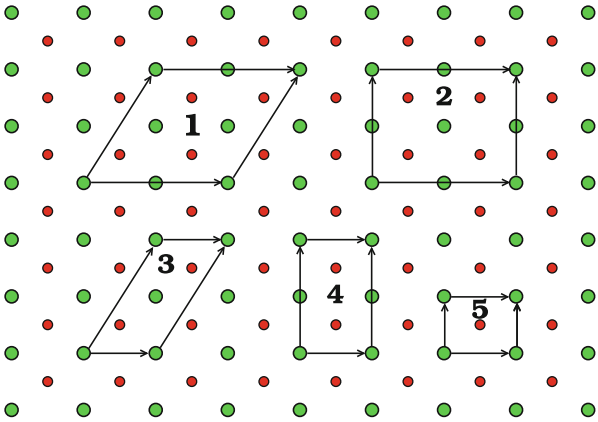


Fig. 3 Five alternate representations of a single physical crystal are shown. Cells 2 and 4 are Niggli reduced versions of cells 1 and 3, respectively. They are also 2 × 2 and 1 × 2 supercells, respectively, of the primitive cell 5

2.1.1 Standardization of Representation

Two techniques are employed to standardize the representation of structures. The first and most widely used method is to impose hard constraints on the structures. These constraints include minimum interatomic distances and lattice parameter magnitudes. Limits on the maximum interatomic distances and lattice vector magnitudes are sometimes enforced as well [26, 27, 30, 31]. In addition, most authors constrain the range of angles between the lattice vectors [26, 27, 30–33]. If the algorithm varies the number of atoms per cell, this value is also constrained [27]. A restriction on the total volume of the cell is an additional possibility [30, 31, 34].

Physical considerations must be taken into account when choosing the constraints. For the constraint of the minimum interatomic distances, choosing 80% of the typical bond length or of the sum of the covalent radii of the two atoms under consideration has been proposed [30, 33]. The minimum lattice length has been chosen by adding the typical bond length and the diameter of the largest atom in the system [30, 32, 33]. Bahmann et al. set the maximum lattice vector length to the sum of the covalent diameters of all atoms in the cell [30]. Several authors limit angles between lattice vectors to lie between 60° and 120° [31, 32], although a more liberal range of 45–135° has also been used [30, 33]. Ji et al. fix the volume of the cell during the search [34], and Lonie et al.'s algorithm can be set to use either a fixed cell volume or to constrain the volume to a user-specified range [31]. In the work of Bahmann et al. the cell volume is constrained to the range defined by the volume of the close-packed structure and four times that value [30].

Additional constraints may be used when one wishes to limit the search to a particular geometry. For example, Bahmann et al. restrict the allowable atomic positions and increase the maximum allowable lattice length in one direction to facilitate a search for two-dimensional structures [30]. Woodley et al. use an EA to search for nanoporous materials by incorporating "exclusion zones," or regions in which atoms are forbidden to reside [35].

Several advantages are gained by using these constraints. Many energy models behave poorly when faced with geometries with very small interatomic distances, so enforcing this constraint from the start helps prevent failed structural relaxations and energy calculations. As mentioned in Sect. 1.1, large regions of the potential energy surface correspond to unphysical structures, and constraints help limit the search to regions that do contain physical minima.

Additionally, they help to ensure that structures are represented similarly. Removing as much redundancy as possible from the space of solutions makes the problem easier without limiting our set of possible answers or introducing any a priori assumptions as to the form of the solution. On the other hand, it is more dangerous to remove merely unlikely regions of the space from consideration, since doing so would bring into question both the validity of results and the claim to first-principles structure prediction.

The second method used to help standardize structure representation involves transforming the cells of all organisms to a unique and physically compact representation when possible. One way to do this is the Niggli cell reduction [27, 36]. There is a Niggli cell for any lattice that is both unique and has the shortest possible lattice lengths. Figure 3 illustrates the representation problem and the Niggli cell reduction. A similar transformation is used by Lonie et al. and Oganov et al. [31, 37]. In addition to simplifying the space that must be searched, removing redundancy by standardizing the representation of structures usually helps to increase the quality of the offspring produced by the mating variation, as is discussed in Sect. 2.6.

2.2 Initial Population

If no experimental data are available for the system under study, then the organisms in the initial population are generated randomly, subject to the constraints discussed above. The initial population generated in this way should sample the entire potential energy landscape within the constraints. If experimental data is available, such as from X-ray diffraction analysis, it can be used to seed the initial population with likely organisms. If one is predicting an entire phase diagram (see Sect. 3), the correct elemental and binary phases may already be known experimentally and can be used in the initial population. The use of pre-existing knowledge has the potential to decrease significantly the time needed to find the global minimum

When searching for molecular crystals, one typically places coherent molecular units instead of individual atoms into the structure [28]. Zhu et al. made an additional modification to the generation of the initial population to facilitate their study of molecular solids [29]. Instead of placing molecules at completely random locations within the cell, structures are built from randomly selected space groups. The authors found that this provides the algorithm with a more diverse initial population and improves the success of the search.

2.3 Fitness

The fitness of an organism is the property on which evolutionary pressure acts, and it depends on the value of the objective function. It is defined so that better solutions have higher fitness, and thus minimizing the energy means maximizing the fitness. It is usually defined as a linear function of the objective function, relative to the other organisms in the population. Exponential and hyperbolic fitness functions have also been used as an alternative way to introduce more flexibility into the selection algorithm (see Sect. 2.4) [38, 39]. In one frequently-applied scheme, an organism with a formation energy per atom, E_f, is assigned a fitness

$$f = \frac{E_f - E_f^{min}}{E_f^{max} - E_f^{min}},$$

where E_f^{max} and E_f^{min} are the highest and lowest formation energies per atom, respectively, in the generation [26, 27, 31, 38]. In this case, the organism with the lowest energy in the generation is assigned a fitness of 1, and the organism with the highest energy has a fitness of 0. An alternative approach is to rank the organisms within a generation by their objective function values. The fitness of an organism is then defined as its rank [32, 34]. For cases when the stoichiometry and number of atoms in the cell is fixed, the fitness can simply be defined as the negative of the energy of the organism [30, 33].

2.4 Selection

The selection method determines which organisms will act as parents. Generally, structures with higher fitnesses are more likely to reproduce. The selection method is a crucial component of the search because it is the only way that the algorithm applies pressure on the population to improve towards the global minimum. Three commonly used strategies are elitist (or truncated) selection, roulette wheel selection, and tournament selection. In elitist selection the top several organisms are allowed to reproduce with equal probability while the rest are prevented from mating [39]. In roulette wheel selection, a random number d between the fitnesses of the best and worst organisms is generated for each organism, and if d is less than the fitness of the organism, it is allowed to reproduce [38]. In this way, it is possible for any organism except the worst one to reproduce, but it is more likely for organisms of higher fitness. Finally, in tournament selection, all of the organisms in the parent generation are randomly divided into small groups, usually pairs, and the best member of each group is allowed to reproduce.

Tipton et al. employ another approach to selection which is essentially a generalization of the three outlined above. Organisms are selected on the basis of a probability distribution over their fitnesses [26, 27]. Two parameters are used to

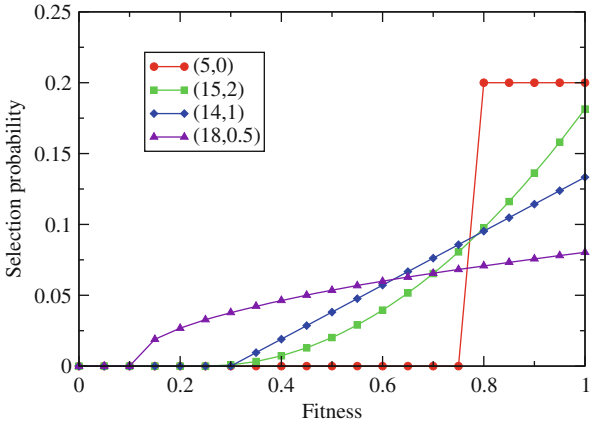


Fig. 4 Tuning the selection probability distribution (N; p) described by the number of potential parents N and power law p allows Tipton et al. to adjust the aggressiveness with which an EA seeks to converge

describe the distribution: the number of potential parents and an exponent which determines the shape of the probability distribution. The number of parents specifies how many of the best organisms in the current generation have nonzero probabilities of acting as parents. The exponent describes a power law. This method allows fine-grained control over the trade-off between convergence speed and the probability of finding the ground state. An aggressive distribution that puts a lot of pressure on the population to improve leads to faster convergence, but the algorithm is more likely to converge to only a local minimum. On the other hand, a less aggressive distribution will probably take more time to converge, but the algorithm has a better chance of finding the global minimum because a higher degree of diversity is maintained in the population. Several choices of selection probability distribution are illustrated in Fig. 4.

Authors employ these strategies in a variety of ways. One approach is to use elitist selection to remove some fraction of the parent generation, and then grow the resulting group back to its original size by creating offspring organisms from the remaining parent organisms with equal probability [30, 33, 34] or with a linear or quadratic probability distribution over their fitnesses [32]. Abraham et al. use roulette wheel selection. When the number of offspring organisms equals the number of parent organisms, either roulette wheel or elitist selection are used on the combined pool of structures to determine which organisms will make up the next generation. Elitist selection was found to be preferable in the final step [38]. Lonie et al. employ a linear probability distribution over the fitnesses to select organisms to act as parents. A continuous workflow is used instead of a generational scheme, so that offspring organisms are immediately added to the breeding population when they are created [31].

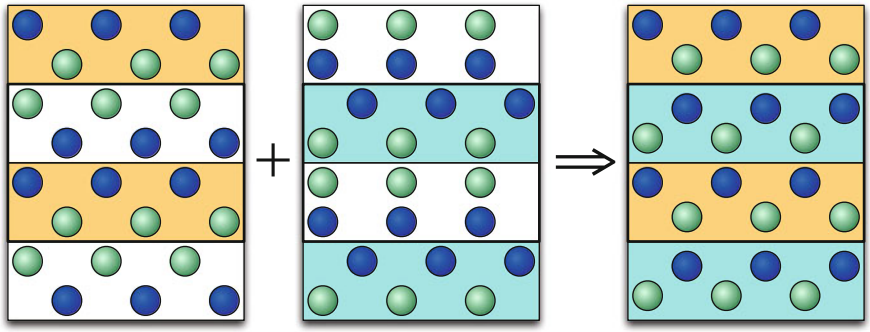


Fig. 5 Schematic illustration of how the mating operator in the evolutionary algorithm combines two crystal structures. For clarity, 1 × 2 supercells of the child and parent structures are shown

2.5 *Promotion*

A new generation is created from the structures in the previous one by applying selection in conjunction with promotion and variation. The promotion operation places some of the organisms in the old generation directly into the new generation without undergoing any changes. This is used to ensure that good genetic material is maintained in the population. Many authors use elitist selection to choose which organisms to promote. Lyakhov et al. refined selecting for promotion by only promoting structures whose fingerprints were significantly different (see Sect. 2.11 for fingerprinting) [40]. This was done to prevent loss of population diversity due to promoting similar organisms.

2.6 *Mating*

The goal of the mating variation is to combine two parents and preserve their structural characteristics in a single offspring organism. In its most basic form, mating consists of slicing parent organisms (cells) into two sections each and then combining one from each parent to produce an offspring organism. This is illustrated in Fig. 5.

It is important that the mating operation be designed so that traits which are important to the energy minimization problem have high heritability. The most energetically-important interactions in materials come from species located close to one another. This suggests that there is some amount of spatial separability in the energy-minimization problem, with the energy depending primarily on the local structure. The mating variation works by exploiting this feature of the problem. The slicing mating variation maintains much of the local structure of each parent in a very direct way, and we will now detail this operation.

2.6.1 Slice Plane Location and Orientation

After two parent organisms have been selected, the next step is choosing the planes along which to slice them. In order to mate organisms that do not have identical lattices, a fractional space representation is used. The positions of atoms in a cell are expressed in the coordinate system of the cell's lattice vectors. As a result, the fractional coordinates of all atoms within the cell have values within the interval [0;1).

Authors have used various techniques to choose the orientation and location of the slice plane. In one method, a lattice vector A and a fractional coordinate s along A are randomly chosen [31, 32]. All atoms in one of the parent organisms with a fractional coordinate greater than s along A and all atoms in the second parent with fractional coordinate less than or equal to s are copied to the new child. Restricting the range of allowed values of s can be used to specify the minimum contribution by each parent to the offspring organism [31].

Alternatively, one may randomly select two planes that are parallel to a randomly chosen facet of the cell. Atoms that lie between these planes are then exchanged between the two parents. By "exchanged" we mean that each atom in the offspring has the same species-type and fractional coordinates as in the corresponding parent. This approach is equivalent to performing a translation operation on the atoms in the cell and then using the single slice plane method outlined above. Both of these methods choose slice planes that are parallel in real space to one of the cell facets of the parent structures [33, 34, 38]. Tipton et al. selects the two slice planes slightly differently: the fractional coordinate corresponding to the center of the sandwiched slab is randomly selected. The width of the sandwiched slab is then randomly chosen from a Gaussian distribution, and the two slice planes are placed accordingly [26, 27]. Another approach is simply to fix the locations of the two slice planes. For example, Abraham et al. specify that the two cuts be made at fractional coordinates of 1/4 and 3/4 along the chosen lattice vector [38].

Abraham et al. introduced a periodic slicing operation [38]. In this case, the value s described above becomes a cell-periodic function of the fractional coordinates along the cell lattice vectors other than A. A sine curve is often used, with the amplitude and wavelength drawn from uniform distributions. The wavelength is commonly constrained to be larger than the typical interatomic distance and smaller than the dimensions of the cell. The amplitude should also be small enough to ensure that no portion of the slice exceeds the boundaries of the cell. Abraham et al. found that periodic slicing improved the mean convergence time of their algorithm over planar slicing [38].

Constraining the degree of contribution of each parent by, for example, stipulating a minimum parental contribution can help prevent the mating operation from reproducing one or other of the parent structures essentially unchanged. Once the contribution of each parent has been determined, lattice vectors must be chosen for the offspring structure. Frequently, a randomly weighted average of the parents' lattice vectors is assigned to the offspring [31–33]. Simply averaging the lattice vectors of the parent organisms, i.e., fixing the weight at 0.5, is another common choice [26, 27, 34].

2.6.2 Number of Atoms and Stoichiometry of Offspring

An offspring organism produced via mating as described so far may have a different number of atoms or a different composition than its parents. This presents a difficulty if one wishes to perform a search with a fixed cell size or at a single composition. The simplest way to deal with these issues is simply to reject all offspring organisms that do not meet the desired constraints [33, 38]. Alternatively, nonconforming offspring can be made acceptable by the addition or removal of atoms. It may be best to add and remove atoms from locations near the slice plane [34]. These corrections minimize disruption to the structure transmitted from the parent organisms. Glass et al. use a slightly different approach: atoms to be removed or added are selected randomly from the discarded fragments of one of the parent organisms [31, 32]. Atomic order parameters have also been used to decide which excess atoms to remove (see Sect. 2.11) [40]. Those with the lower degrees of local order are more likely to be removed.

2.6.3 Modifications to the Mating Variation

Several additional modifications of the mating variation have been explored. The first involves shifting all the atoms in a cell by the same amount before mating [32, 41]. These shifts may happen with different probabilities along the axis where the cut is made and an additional random axis. This removes any bias caused by the implicit correlation between the coordinate s on the axis A in one crystal with the coordinate s on the axis A in the other. A similar effect may be obtained by selecting a random vector and shifting all atoms by this vector prior to making the cut [31]. In practice, these shifts help repeat good local structures to other parts of the cell.

In a further innovation, the parent organisms are subjected to random rotations and reflections prior to mating. This procedure removes bias toward any given orientation [31]. Additionally, an order parameter may be used to inform the choice of contribution from each parent (see Sect. 2.11). Several trial slabs of equal thickness are cut from the parents at random locations, and the slabs with the highest degree of order parameters are passed to the offspring [40, 41].

The simple slicing mating operator is not appropriate for molecular crystals, since it does not respect the integrity of the molecular units. Zhu et al. adapted the mating variation to search for molecular crystals [29]. In their scheme, each molecule is treated as an indivisible unit, and the location of the geometric center of a molecule is used to determine its location for the purposes of the mating operator.

2.6.4 Shortcomings of the Slicing Mating Variation

The mating variation acts directly on the particular representation of a structure in the computer. Since it is performed in fractional space, the mating variation can be applied to any two parent organisms, regardless of their cell shapes. However, the offspring structure may not always be successful. An offspring organism that has little in common with either parent can be produced if their representations (in particular the lattice parameters or number of atoms in the cells) are sufficiently different. As a result, the offspring will often have low fitness. Thus the mating variation is most successful when the parents are represented similarly because this increases the heritability of important traits.

The constraints and cell transformations discussed in Sect. 2.1 combat this issue through standardization of structure representation. Another method to increase the similarity of representation prior to mating is to use a supercell of one of the parent structures during mating [26, 27]. If one of the parent structures contains more than twice as many atoms as the other, a supercell of the smaller parent is used in the mating process. This technique ensures that both parent organisms are approximately the same size before mating, which aids in the creation of successful offspring.

Lyakhov et al. use an additional technique to help increase the viability of the offspring. If the distance between the parent organisms' fingerprints (see Sect. 2.10) exceeds a user-specified value, the would-be parents are not allowed to mate. The rationale behind this stipulation is that if two parents are from different funnels in the energy landscape, then their offspring would likely be located somewhere between those funnels and therefore have low fitness [40, 41].

2.6.5 Other Mating Operations

Not all evolutionary algorithms employ the previously described slicing method for mating. Bahmann et al. use a general recombination operation instead, where an offspring structure is produced by combining the lattice vectors and atomic positions of the two parent organisms [30]. This can be done in two ways: intermediate recombination takes a weighted average of the parents' values, and discrete recombination takes some values from each parent without changing them. Smith et al. used binary strings to represent structures on a fixed lattice, with each character in the string indicating the type of atom at a point on the lattice [24]. Mating was carried out by splicing together the strings of two parent structures. Jóhannesson et al. used a similar approach to search for stable alloys of 32 different metals [117]. Although all of these methods combine traits from each of the parents, they may not be as successful in passing the important local structural motifs of parents to the offspring.

2.7 Mutation

The goal of the mutation operation is to introduce new genetic material into the population. Its utility lies in its ability to explore the immediate vicinity of promising regions of the potential energy surface that have been found via the mating variation. The most common mutation entails randomly perturbing the atomic positions or lattice vectors of a single parent organism to produce an offspring organism. Some approaches call for mutating both the lattice vectors and the atomic positions [26, 27, 33], while others affect only one type of variable. To apply a mutation to the lattice vectors, they are subjected to a randomly generated symmetric strain matrix of the form

$$S = \begin{pmatrix} 1+e_1 & e_6/2 & e_5/2 \\ e_6/2 & 1+e_2 & e_4/2 \\ e_5/2 & e_4/2 & 1+e_3 \end{pmatrix}, \tag{1}$$

where the components e_i are taken from uniform or Gaussian distributions [31–33].

Mutations of the atomic positions are achieved in a similar fashion. Each of the three spatial atomic coordinates is perturbed by a random amount, often obtained from a uniform or Gaussian distribution [26, 27]. To keep the size of these perturbations reasonable, either an allowed range or a standard deviation is set by the user. Most formulations do not mutate every atom in the cell but instead specify a probability that any given atom in the cell will be displaced. The approach of Abraham et al. combines mutation with the mating variation, perturbing atomic positions after mating has been performed [38]. However, most authors treat mutation as a separate operation.

Glass et al. claim that randomly mutating atomic positions is not necessary because enough unintentional change occurs during mating and local optimization to make it mostly redundant [32]. However, in later work, Lyakhov et al. use a "smart" mutation operation where atoms with low-order local environments (see Sect. 2.11) are shifted more [40, 41]. A further refinement to mutation has been made by shifting all the atoms along the eigenvector of the softest phonon mode [40, 41].

Permutation is another mutation-type operation for multi-component systems that swaps the positions of different types of atoms in the cell. Generally, the user specifies which types of atoms can be exchanged, and the algorithm performs a certain number of these exchanges each time the permutation variation is used on a parent organism [26, 27, 37]. The extent to which exchanging atomic positions affects the energy is strongly system-dependent. For ionic systems, exchanging an anion with a cation is likely to result in a much larger energy change than exchanges between two different types of cations or anions. In metals, on the other hand, the change in energy under permutation corresponding to anti-site defects is generally small. It is often helpful to use a permutation variation when studying these systems in order to find the minimum among several competing low energy configurations.

The number of swaps carried out can be random within a specified range, or can be pulled from a user-specified distribution. Randomly exchanging all types of atoms has the drawback that many energetically unfavorable exchanges may be performed, especially in ionic systems. If the number of atoms in the cell is small, Trimarchi et al. do an exhaustive search of all possible ways to place the atoms on the atomic sites [33].

Lonie et al. employ a "ripple" variation, in which all atoms in the cell are shifted by varying amounts [31]. First, one of the three lattice vectors is randomly chosen and then atomic displacements are made parallel to this axis. The amount by which each atom is shifted is sinusoidal with respect to the atom's fractional coordinates along the other two lattice vectors. This produces a ripple effect through the cell. Lonie et al. argue that this variation makes sense because many materials display ripple-like structural motifs. Combining the ripple variation with other variations such as the lattice vector mutation and the permutation leads to hybrid variations that can improve the performance of the EA by reducing the number of redundant structures encountered in the search [31].

Zhu et al. employ an additional mutation when searching for molecular solids. Since molecules are not usually spherically symmetric, a rotational mutation operator was introduced, in which a randomly selected molecule is rotated by a random angle [29].

2.8 System Size

The number of atoms per cell, N, is an important parameter that needs to be considered. If N is fixed to a value which is not a multiple of the size of the ground state primitive cell of the material, the search cannot identify the correct global minimum. However, N is a difficult parameter to search over. In the case of other degrees of freedom for the solution, such as interatomic distances and cell volume, the local optimization performed by the energy code helps to find the best values. No analogous operation is possible in the case of N. Furthermore, the energy hypersurface is not particularly well behaved with respect to this parameter. It is likely that values of N surrounding the optimum will lead to structures quite high in energy while values of N further from the ideal may lead to closer-to-ideal structures.

Several approaches exist to search over this parameter. The first is simply to "guess" the correct value of N [31–33]. Guessing N can make it easy to miss the global minimum, especially for systems about which little is known a priori. To increase the chances of finding the right number, searches can be performed at several different values of N, but this is inefficient. A second technique is to allow the cell size of candidate solutions to vary during the search. This can be done passively through the mating variation by not enforcing a constraint on the number of atoms in the offspring structure [27, 34, 38]. Incorporating a mutation-type variation specifically designed for varying N is an additional option.

Another way to aid the search for the correct number of atoms per cell is to use large cells. Large supercells effectively allow several possible primitive cell sizes to be searched at once because the cells can be supercells of multiple smaller cells. For example, a search with a 50-atom supercell is capable of finding ground state structures with primitive unit cells containing 1, 2, 5, 10, 25, and 50 atoms. However, because the number of local minima of the energy landscape increases exponentially with N [13], and because individual energy calculations are much more expensive for larger structures, efficiency suffers.

Lyakhov et al. describe another difficulty with the large supercell approach that arises when generating the initial population. Randomly generated large cells almost always have quite poor formation energies, and disordered glass-like structures dominate. This discovery implies that there exists an upper limit to the size of randomly generated structures that can provide a useful starting point for the search. Starting an evolutionary algorithm with a low-diversity initial population comprised of low fitness structures provides a small chance of finding the global minimum [40, 41]. To obtain reasonably good large cells for the initial population, Lyakhov et al. generate smaller random cells of 15–20 atoms, and then take supercells of these [40, 41]. In this way, the organisms in the initial population can still contain many atoms, but they possess some degree of order and therefore tend to be more successful.

An alternative approach is to start with smaller supercells and encourage them to grow through the course of the search [26, 27]. This is achieved by occasionally doubling the cell size of one of the parents prior to performing the mating variation. The speed of cell growth in the population can be controlled through the frequency of the random doubling. The advantage of this technique is that it searches over N while still gaining (eventually) the benefits of large supercells. In addition, considering smaller structures first ensures that the quicker energy calculations are performed early in the search, and the more expensive energy calculations required for larger cells are only carried out once the algorithm has already gained some knowledge about what makes good structures.

2.9 Development and Screening

After a new organism has been created by one of the variations, it is checked against the constraints described in Sect. 2.1 and tested for redundancy (see Sect. 2.10). At this stage, many EAs scale the atomic density of the new organisms using an estimate of the optimal density [26, 27, 32]. Starting from an initial guess of the optimal density ρ_0, the density estimate is updated each generation by taking a weighted average of the old best guess ρ_i and the average density of the best few structures in the most recent generation ρ_{ave}:

$$\rho_{i+1} = w\rho_{ave} + (1 - w)\rho_i,$$

where w is the density weighting factor. Then any time a new organism is made it is scaled to this atomic density before local relaxation. The primary reason for the density scaling is a practical one. Many minimization algorithms are quite time consuming if the initial solution is far from a minimum. This scaling is an easy first pass at moving solutions towards a minimum. Because the density scaling of an organism alters the interatomic distances, etc., the constraints checks are performed after the scaling of the density.

2.10 Maintaining Diversity in the Population

As the evolutionary algorithm searches the potential energy surface, equivalent structures sometimes occur in the population. If a pair of structures mates more than once, they are likely to create similar offspring. If the set of best structures does not change from generation to generation due to promotion, the set of parents, and thus the resulting set of children, can also be very similar. In addition, as the generation as a whole converges to the global minimum, all the organisms are likely to become more similar. What is worse, once a couple of low energy, often selected organisms are in the population, they can reproduce and similar structures will effectively fill up the next generations.

Duplicate structures hinder progress for several reasons. The most computationally expensive part of the algorithm is the energy calculations, and performing multiple energy calculations on the same structure is wasteful. However, this is exactly what happens if duplicate structures are not identified and removed from the population. Furthermore, low diversity in the population makes it difficult for the algorithm to escape local minima and to explore neighboring regions of the potential energy surface. This leads to premature convergence which is in practice indistinguishable from convergence to the correct global minimum. For these reasons, it is desirable to maintain the diversity of the population by identifying and removing equivalent structures. This is not a trivial task because, as discussed in Sect. 2.1, there exist infinitely many ways to represent a structure. Numerical noise adds to the difficulty of identifying equivalent structures.

Some authors directly compare atomic positions to determine whether two structures are identical. Lonie et al. developed an algorithm for this purpose, and it correctly identified duplicate structures that had been randomly rotated, reflected, or translated, and had random cell axes [42].

Tipton et al. also use a direct comparison of structures, with a slight modification [26, 27]. During the search, two lists of previously-observed structures are maintained. The first contains all the structures, relaxed and unrelaxed, that the algorithm has seen. If a new unrelaxed offspring structure matches one of the structures in the list, it is discarded. The assumption is that if it was good enough

to keep the first time, it was promoted, and if not, there is no reason to spend more effort on it. A second list contains the relaxed structures of all the organisms in the current generation. If a new relaxed offspring structure matches one of the structures in this list, it is discarded to avoid having duplicate structures in the generation. This approach both minimizes the number of redundant calculations performed and prevents the population from stagnating.

Wang et al. employ a bond characterization matrix to identify duplicate structures in the population [8]. The components of the matrix are based on bond lengths and orientations, and the types of atoms participating in the bond. Bahmann et al. identify duplicate structures by choosing a central atom in each organism and comparing the bond lengths between the central atom and the other atoms in a supercell [30]. These authors introduced an additional technique to help prevent the population from stagnating by stipulating an organism age limit. If an organism survives unchanged (via promotion) for a user-specified number of generations, it is removed from the population. This feature is meant to prevent a small number of good organisms from dominating the population and reducing its diversity.

Another method involves defining a fingerprint function which describes essential characteristics of a structure. When two organisms are found to have the same fingerprints, they are likely identical, and one is discarded. Several fingerprint functions have been used. The simplest is just the energy [22, 39]. The logic is that if two structures are in fact identical, they should have the same energies to within numerical noise. An interval is chosen to account for the noise. However, the size of the interval is fairly arbitrary and system-dependent, and this method is prone to false positives. Eliminating good unique organisms from the population can be even more detrimental to the search than not removing any organisms at all [42]. Lonie et al. expanded the fingerprint function to include three parameters: energy, space group, and the volume of the cell [31]. Again, intervals were set on the volume and energy. This is an improvement over simply using the energy as a fingerprint, but it can still occasionally fail, especially for low-symmetry structures or when atoms are displaced slightly from their ideal positions. Lonie et al. found that their direct comparison algorithm outperformed their fingerprint function at identifying duplicate structures [42].

Valle et al. employ a fingerprint function that is based on the distributions of the distances between different pairs of atom types in an extended cell [41, 43, 44]. For example, a binary system contains three interatomic distance distributions. The fingerprint function takes all three distributions into account. They used this fingerprint function to define an order parameter (see Sect. 2.11). Zhu et al. modified this fingerprint function slightly when searching for molecular solids; since distances between atoms within molecules do not change significantly, these distances are not considered when calculating structures' fingerprints [29].

Discarding duplicate structures from the population is not the only method employed to maintain diversity. Abraham et al. use a fingerprint function to determine how similar all the structures in a generation are to the lowest energy structure in the generation. Instead of simply removing similar structures, a modified fitness function is used which penalizes organisms based on their similarity to this best structure [45].

2.11 Order Parameters

Order parameters give a measure of the degree of order of an entire structure and also of the local environment surrounding individual atoms. Since energy is often correlated with local order, this can be a useful tool. Valle et al. extended their fingerprint function by using it to define an order parameter [40, 41, 44]. They used it to guide the algorithm at various points, as mentioned previously.

2.12 Frequency of Promotion and Variations

The user-specified parameters of an evolutionary algorithm affect its performance. However, running hundreds or thousands of structure searches to optimize these parameters can be prohibitively expensive, especially if an ab initio energy model is used. Furthermore, optimal values depend on the system under study. Physical and chemical intuition can be used to specify some of the parameters, such as the minimum interatomic distance constraint, but there exists no clear way to determine many of the others without performing enough searches to obtain reliable statistics.

Many authors arbitrarily choose how much each variation contributes to the next generation [32–34]. Lonie et al. performed thousands of searches for the structure of TiO_2 using empirical potentials to determine the best set of parameters for their algorithm [31]. They found that the relative frequency of the different variations did not significantly affect the success rate of the algorithm. However, the parameters associated with each variation did. For example, the lattice mutation variation was found to produce more duplicate structures when the magnitude of the mutation was small. This is likely to be due to structures relaxing back to their previous local minima when only slightly perturbed.

There is an important distinction between the relative frequency with which a given variation is called by the algorithm and the actual proportion of organisms in the next generation that are produced by that variation. The difference arises because not all the variations have the same likelihood of creating viable offspring. For example, mating is more likely to give good offspring structures than mutation of atomic positions because the latter will more frequently produce offspring that violate the minimum interatomic distance constraint. For this reason, the researcher's intention may be more clearly communicated if the proportion of offspring created by each variation is specified rather than the frequency that each variation is called by the algorithm.

Sometimes a situation arises in which it is not possible for one of the variations to produce a viable offspring organism. This could happen, for example, if the variation increases the size of the structures in the system, but all the potential parent organisms are already close to the maximum allowed cell size. In this case, the search will stop unless there is some way for the algorithm to get around the user-specified requirement that a certain percentage of the offspring come from this variation. Setting an upper limit on the number of failed attempts per variation is one way to achieve this [26, 27].

2.13 Convergence Criteria: Have We Found the Global Minimum?

When searching for an unknown structure, there is no known criterion that guarantees that the best structure encountered by the evolutionary algorithm is in fact the global minimum. One common technique to analyze the success rate of a heuristic search algorithm was given by Hartke [46]. In this method, many independent structure searches are performed on the same system, using the same set of parameters for the algorithm. For each search the energy of the best structure in each generation is recorded. These values are then used to create a plot of the energy vs generation number (or the total number of energy evaluations) that contains three curves: the energy of the highest-energy best structure, the energy of the lowest-energy best structure, and the average energy of the best structures. One shortcoming of the Hartke plot is that the lowest and highest best energies encountered are outliers, and in practice they depend strongly on the choice of the number of independent structure searches.

Tipton et al. employs a statistically more relevant approach to quantifying an EA's performance in which the median, the 10th percentile, and the 90th percentile energies of the best structures are plotted [27]. The 10th and 90th percentiles offer a better characterization of the distribution of results and are less susceptible to outliers and the number of independent structure searches performed to characterize the efficiency of the algorithm.

Figure 6 shows an example of a performance distribution plot for Zr_2Cu_2Al that was obtained by performing 100 independent runs of an EA with an embedded atom model potential [27]. These plots provide insight into the expected performance of the algorithm for the given material system and parameter settings and enable statistical comparisons of the performance of different methodologies or parameterizations of an EA. Of course, the strength of these conclusions depends on how many searches were used to construct the performance distribution plot, and the algorithm must be tested on systems with known ground state structures to be certain when the search was successful.

Lonie et al. showed that a decaying exponential fits the average-best energy curve of a Hartke plot well for a system with a known ground state structure [31]. The halflife of the exponential fit provides a measure of how fast the algorithm converges and can be used to determine a stopping criterion for the search. However, not all of the searches find the global minimum, so allowing a search to run for many halflives still does not guarantee that the global minimum will be found, but it does increase confidence in the result.

A more common approach is to stop the search after a user-specified number of generations has elapsed without improvement of the best organism [26, 27, 30, 31]. Stopping once an allocated amount of computational resources has been expended is a popular alternative. Bahmann et al. have determined convergence when population diversity falls below a certain threshold or when all the organisms have very similar energies [30].

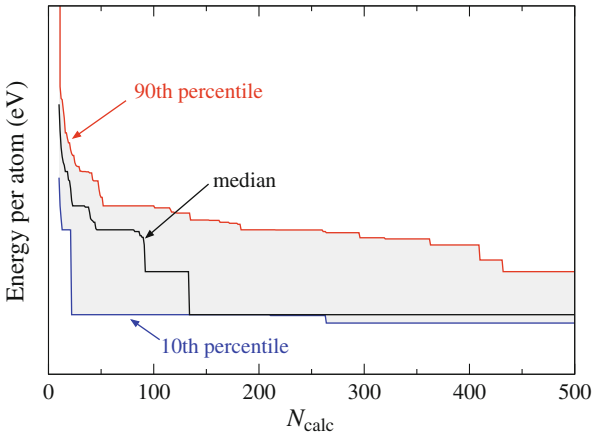


Fig. 6 Performance distribution plot for 100 structure searches at fixed composition for Zr_2Cu_2Al using an embedded atom model potential [27]. The energy of the 90th percentile best structure is shown in *red*, the tenth percentile best structure in *blue*, and the median best structure in *black*

Although most authors use one of the fairly simple convergence criteria mentioned above, a quantitative statistical approach has been proposed by Venkatesh et al. [14]. Using Bayesian analysis, they determined the distribution of local minima based on the number found by a random search. This distribution was then used to calculate how many attempts would be required to find the global minimum with a specified probability.

3 Phase Diagram Searching

Even when one can say with a reasonable degree of confidence that the evolutionary algorithm has converged to the global minimum of the potential energy landscape, the result might still not represent the lowest energy structure that would be observed in nature. Skepticism is justified for several reasons [47]. First, as discussed in Sect. 2.8, unless the number of atoms in the cell is correctly guessed or allowed to vary, the EA cannot find the global minimum. Second, the structure identified as the global minimum might not be mechanically or dynamically stable, which would be reflected in energy-lowering imaginary phonon modes for the proposed global minimum crystal structure. Third, the reported global minimum might actually represent a metastable phase that decomposes into two or more structures with different stoichiometries. To determine whether this is the case, a phase diagram search must be performed. In addition to predicting the decomposition of structures into phases of other stoichiometries, phase diagrams are of great interest for many practical applications.

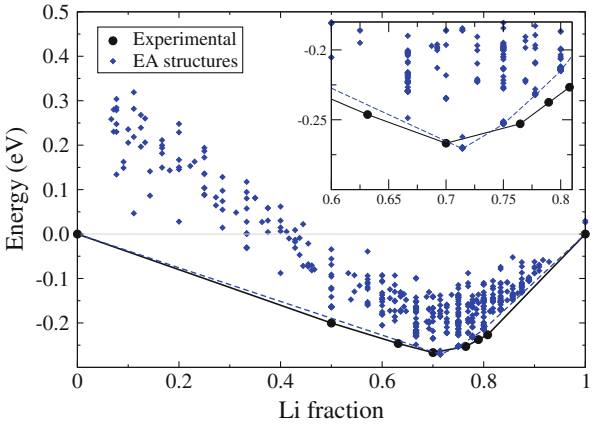


Fig. 7 A phase diagram search of the Li–Si binary system by Tipton et al. [48] using the method of Trimarchi et al. [49] showed that a search for relatively small unit cell structures could approximate the structural and energetic characteristics of the known very large experimental structures and thus be used to predict the voltage characteristics of a Li–Si battery anode. The search also identified a previously unknown member of the low-temperature phase diagram with composition Li_5Si_2

In order to perform a phase diagram search, we make use of the convex hull construction [9]. The formation energies of all structures with respect to the elemental constituents are plotted vs the composition. To determine the elemental references, one can either refer to the literature or perform preliminary searches. The smallest convex surface bounding these points is the convex hull, and the lowest energy facet for each composition is of physical interest. Thus, the convex hull is a graphical representation of the lowest energy a system can attain at each composition, and the points that lie on the convex hull correspond to stable structures. Figure 7 is an example of a convex hull for the Li–Si binary system [48].

Two approaches have been used to construct the convex hull. The first is to perform fixed-stoichiometry searches at many compositions [9, 50]. The lowest energy structure found in each search is then placed on an energy vs composition plot and the convex hull is constructed. However, this method is computationally expensive because it requires many separate searches to adequately sample the composition space [27].

The second approach entails modifying the evolutionary algorithm to search over composition space in the course of a single run. This requires two changes to the standard algorithm. The first is that the stoichiometry of structures the algorithm considers must be allowed to vary. This can be achieved simply by giving the initial population random stoichiometries and removing stoichiometry constraints on offspring structures [49]. The second modification involves the objective function. The algorithm constructs a current convex hull for each generation of structures, and a structure's objective function is defined as its distance from the current convex hull [26, 27, 49]. In this way, structures that lie on the current convex hull

have the highest fitness, and those above the convex hull have lower fitnesses. Selection then acts on this value in the standard way described in Sect. 2.4. As the search progresses, the true convex hull of the system is approached.

As discussed in Sect. 2.8, the global minimum cannot be found if the cell does not contain an integer multiple of the correct number of atoms. Since the structures lying on the convex hull often do not contain the same numbers of atoms, allowing the number of atoms to vary (Sect. 2.8) during the phase diagram search helps the algorithm find the correct convex hull. An alternative approach is to perform several composition searches with different, fixed system sizes. Each search generates a convex hull, and these hulls can be overlaid to obtain the overall lowest convex hull [49]. It should also be noted that the stoichiometries accessible to the algorithm are constrained by the number of atoms in the cell. For example, a cell containing four atoms in a binary system provides the algorithm with only five possible compositions (0, 25, 50, 75, and 100% A or B). The use of larger system sizes may be necessary for the algorithm to find the correct convex hull.

Another difficulty with phase diagram searches is inadequate sampling of the entire composition range. Mating between parents with different stoichiometries tends to produce offspring structures of intermediate composition. Because of this, over time the population as a whole may drift toward the middle region of the composition range, making it difficult to sufficiently sample more extreme compositions. Two solutions to this problem have been proposed [26, 27]. The first is to modify the selection criteria in such a way that mating between parents with similar compositions is encouraged. The second is to divide the composition range into sections and perform separate searches over each section. Agglomerating the results from all the sections gives the overall convex hull.

4 Energy Calculations and Local Relaxation

The potential energy landscape over which an evolutionary algorithm searches is defined by the code used for the energy calculations. These energy codes approximate the true potential energy landscape of the system, so the global minimum found by an EA will only represent the true global minimum of the system insofar as the approximate Hamiltonian accurately represents the physics of the system.

As discussed in Sect. 1.1, the potential energy surface is divided into basins of attraction by the local structure optimization or relaxation available in most energy codes (see Fig. 1). In order to find the global minimum, we must only sample a structure that resides in its basin of attraction, and the local optimizer will do the rest. This tremendously reduces the effective size of the space that must be searched; a relatively sparse sampling of a region can identify most of the local minima [7]. Local optimization is therefore crucial to the success of the search. Although the method depends on the energy code used, the local optimization problem is relatively well understood and its solutions are generally stable.

Glass et al. observed that the energies of a relaxed and unrelaxed structure are only weakly correlated [32]. This implies that the energy of an unrelaxed structure is not a reliable indicator of how close that structure is to a minimum in the potential energy landscape. Although omitting local optimization is computationally cheaper, an evolutionary search performed this way is unlikely to be successful. Woodley et al. compared the performance of an evolutionary algorithm with and without local relaxation and found that locally relaxing every structure greatly improved the efficiency and success rate of the algorithm [51].

Both empirical and ab initio energy codes have successfully been used in evolutionary algorithms to perform energy calculations and local relaxations. Due to their approximate nature, empirical potential energy landscapes often contain unphysical minima [52]. In addition, the cut-off distances imposed in many interatomic potentials leave discontinuities in the energy landscape, which can impede local relaxation. Although they mimic the true potential energy landscape more accurately than empirical potentials, density functional theory (DFT) calculations are also capable of misguiding the search if care is not taken. Pickard et al. found that insufficiently dense k-point sampling can lead to false minima, and for calculations at high pressures, pseudopotentials with small enough core radii must be used to give accurate results [7].

Many EA implementations are interfaced with multiple energy codes, and more than one type of energy calculation may even be used in a single search. Ji et al. employed both empirical potentials and DFT calculations when searching for structures of ice at high pressures [53]. Lennard-Jones potentials were used for most energy calculations, but ab initio calculations were performed periodically and the parameters of the Lennard-Jones potentials were fitted to the DFT results. In this way, the empirical potential improved as the search progressed, and fewer computational resources were consumed than if ab initio methods alone had been used.

5 Summary of Methods

Tables 1 and 2 list the salient details of several implementations of evolutionary algorithms for structure prediction. The codes listed in Table 1 are production codes available to other users; the codes in Table 2 are research codes. In the following we summarize some of the distinguishing features of these evolutionary algorithms.

The Genetic Algorithm for Structure Prediction (GASP) is interfaced with VASP, GULP, LAMMPS, and MOPAC and has phase diagram searching capability [26, 27]. In addition, GASP can perform searches with a variable number of atoms in the cell, and it implements a highly tunable probability distribution for selecting organisms for mutation, mating, and promotion. The Open-Source Evolutionary Algorithm for Crystal Structure Prediction (XTALOPT) is interfaced with VASP, PWSCF, and GULP. It incorporates a unique ripple mutation, as well as hybrid mutations. It does not use a generational scheme but rather allows offspring structures to act as parents

Table 1 Comparison of the methods implemented into evolutionary algorithms in various available production codes

Name	GASP	XTALOPT	USPEX	EVO	MAISE
Authors	Tipton et al. [26, 27]	Lonie et al. [31]	Glass et al. [32]	Bahmann et al. [30]	Kolmogorov et al. [54]
Selection strategy	Probability distribution	Linear distribution	Elitist, linear, or quadratic probability distribution	Elitist	
Mutation of lattice vectors	Yes	Yes	Yes	Yes	Yes
Mutation of atomic positions	Yes	No	No [32], yes [40]	No	Yes
Permutation of atomic positions	Yes	Yes	Yes	No	Yes
Promotion	Yes	No	Yes	Yes	
Volume scaling	Yes	Yes	Yes	No	
Number of atoms in the cell	Variable	Fixed	Fixed	Fixed	
Cell reduction	Yes	Yes	Yes	No	
Diversity protection	Direct comparison	Fingerprinting [31], direct comparison [42]	Fingerprinting	Fingerprinting	
Phase diagram searching	Yes	No	No	No	No
Energy codes	VASP, MOPAC, GULP, LAMMPS	VASP, PWSCF, GULP, CASTEP	VASP, SIESTA, GULP, DMACRYS, CP2K, PWSCF	PWSCF, GULP	VASP
Unique features	Flexible probability distribution for selection	Ripple and hybrid Ripple and hybrid scheme without generations	Use of order parameters	Age limit, alternative mating operation	

Table 2 Comparison of the methods implemented into evolutionary algorithms in various research codes

Authors	Ji et al. [34]	Trimarchi et al. [33]	Bush et al. [2]	Abraham et al. [38]
Selection strategy	Elitist	Not specified	Elitist	Roulette wheel, elitist
Mutation of lattice vectors	No	Yes	No	No
Mutation of atomic positions	No	Yes	Yes	Yes
Permutation of atomic positions	Yes	Yes	No	No
Promotion	Yes	Yes	Yes	Yes
Volume scaling	No	No	No	No
Number of atoms in the cell	Fixed	Fixed	Fixed	Variable
Cell reduction	No	No	No	No
Diversity protection	None	None	None	None
Phase diagram searching	No	Yes	No	No
Energy codes	VASP	VASP	GULP	CASTEP
Unique features	Cells constrained to constant volume		Surrogate objective function	Periodic slicing, mutation only after mating

as soon as they are created [31]. The Universal Structure Predictor: Evolutionary Xtallography (USPEX) code is interfaced with VASP, SIESTA, PWSCF, GULP, DMACRYS, and CP2K. It incorporates a unique order parameter, both for cells and individual atoms, that is used to help guide the search [32, 40]. The Evolutionary Algorithm for Crystal Structure Prediction (EVO) is interfaced with PWSCF and GULP. It applies an age limit to structures encountered in the search, and it also employs a mating operation that is different from the cut-and-splice technique used in most other evolutionary structure searches [30]. Finally, the Module for Ab Initio Structure Evolution (MAISE) is interfaced with VASP. Both planar and periodic slices can be used during mating, and it has the option to perform mating and mutation in a single variation [54].

Trimarchi et al. developed an evolutionary algorithm for structure prediction that is interfaced with VASP [33]. They later extended the algorithm to include phase diagram searching [49]. Abraham et al. designed an evolutionary algorithm with several unique features, including periodic slicing during the mating operation and mutation of atomic positions only after mating [38]. The algorithm also accepts offspring structures with different numbers of atoms than the parents. It is interfaced with CASTEP. The evolutionary algorithm of Ji et al. is interfaced with VASP, and it constrains the structures it considers to a constant volume

[34]. Bush et al. developed an evolutionary algorithm that incorporates a surrogate objective function [2].

6 Applications

Evolutionary algorithms have been used to solve the structures of many types of systems including molecules, clusters, surfaces, nanowires, and nanoporous materials [39, 55–57]. Here we focus on applications of EAs to bulk, 3D periodic systems. Within this constraint, we have made an effort to provide a comprehensive review of prior applications. We grouped the application into six categories based on the type of material studied: pure elements, hydrogen-containing compounds, intermetallics, minerals, molecular solids, and other inorganic compounds. Tables 3, 4, 5, 6, 7, and 8 correspond to these categories and list the applications of the method. For each study, we indicate the system studied, the number of atoms in the configuration space searched over, the energy code used, and the lead author.

6.1 Elemental Solids

Table 3 describes searches for elemental solids. Some elemental phase diagrams are still not fully characterized, especially under extreme conditions such as high pressure. Several elements have been predicted to display unusual properties at high pressure, such as superconductivity. Ma and Oganov studied several different elements under pressure. They found a new phase of boron with 28 atoms in the unit cell that is predicted to be stable in the pressure range 19–89 GPa [58]. A search of carbon under high pressures led to the prediction that the bc8 structure is more stable than diamond above 1 TPa [25]. Oganov et al. predict several new superconducting phases of calcium at pressures up to 120 GPa [61]. A study of hydrogen at pressures up to 600 GPa predicted that it remained a molecular solid throughout this pressure range [25]. The interesting case of hydrogen under pressure will be further described below in the discussion of hydrogen-containing compounds. Ma et al. predict that potassium and rubidium follow the same sequence of phase transitions under pressure (40–300 GPa) observed experimentally for cesium, but predict a new cubic phase of lithium above 300 GPa [63]. Ma et al. also studied nitrogen under pressure, predicting new polymeric insulating phases above 188 GPa, and studied sodium at pressures up to 1 TPa, predicting a new optically transparent, insulating phase above 320 GPa [65]. They have also reported a monoclinic, metallic, molecular phase of oxygen in the range of 100–250 GPa whose calculated XRD diffraction pattern is in agreement with experiment [66].

Bi et al. searched for phases of europium at pressures up to 100 GPa and predicted several nearly degenerate structures in the range 16–45 GPa, which

Table 3 Application of evolutionary algorithms to single element systems

Element	Pressure (GPa)	Number of atoms	Number of atoms	References
Al	0	4, 8, 12	PWSCF	Bahmann et al. [30]
B	0,100,300	2, 3, 4, 6, 8, 9, 16, 12, 24, 26, 28, 30, 32	VASP	Oganov et al. [58]
	0	24, 25, 26, 27, 28, 29, 30, 31, 32	VASP	Ji et al. [34]
Ba	0–300	Variable, up to 15	VASP	Taillon et al. [59]
C	10–100	2, 4, 6, 8	VASP	Li et al. [60]
	0	Not specified	CASTEP	Abraham et al. [38]
	0	8, others tried	PWSCF	Bahmann et al. [30]
	0–2,000	8	VASP	Oganov et al. [25]
Ca	20–600	3, 4, 6, 8, 9, 12, 16	VASP	Oganov et al. [61]
Cl	100	8	VASP	Oganov et al. [25]
Eu	0–90	Variable, up to 30	VASP	Bi et al. [62]
F	50, 100	8	VASP	Oganov et al. [25]
Fe	350	8	VASP	Oganov et al. [25]
Ga	0	8	PWSCF	Bahmann et al. [30]
H	Up to 600	2, 3, 4, 6, 8, 12, 16	VASP	Oganov et al. [25]
In	0	4	PWSCF	Bahmann et al. [30]
K	40–300	4, 6, 8, 12, 16	VASP	Ma et al. [63]
Li	30–1,000	4, 6, 8, 12, 16, 24	VASP	Ma et al. [63]
Mo	0	Variable, up to 40	LAMMPS	Park et al. [52]
N	100–350	8, 12, 16, 32	VASP	Ma et al. [64]
	100	6, 8, 12, 16	VASP	Oganov et al. [25]
Na	0–1000	Not specified	VASP	Ma et al. [65]
O	100–250	4, 8	VASP	Ma et al. [66]
	25, 130, 250	4, 6, 8, 12, 16	VASP	Oganov et al. [25]
Rb	40–200	4, 6, 8, 12	VASP	Ma et al. [63]
S	12	3, 4, 6, 8, 9, 12	VASP	Oganov et al. [25]
Si	0	8	VASP	Trimarchi et al. [33]
	10, 14, 20	8	VASP	Oganov et al. [25]
Tl	0	4	PWSCF	Bahmann et al. [30]
Xe	200, 1,000	8	VASP	Oganov et al. [25]

may help explain the mixed phase structure observed experimentally in this pressure regime [62] and the occurrence of superconductivity and magnetism in these phases [109]. In a novel application of evolutionary algorithms, Park et al. used an EA to verify that a new modified embedded atom potential for molybdenum accurately reproduced the energy landscape of molybdenum [52].

Table 4 Application of evolutionary algorithms to hydrogen-containing compounds

Compound	Pressure (GPa)	Formula units per cell	Energy code	References
Li_mH, $m = 2 - 9$	60, 80, 100	2, 4	VASP	Hooper et al. [67]
LiH_n, $n = 2 - 8$	0–300	0–300	VASP	Zurek et al. [50]
NaH_n, $n = 6 - 12$	100, 300	2, 3, 4	VASP	Baettig et al. [68]
KH_n, $n = 2$–12	10, 100, 250	2, 4	VASP	Hooper et al. [69]
RbH_n, $n = 2$–14	2–250	2	VASP	Hooper et al. [70]
CsH_n, $n = 2$–9;16	30–200	2, 3, 4	VASP	Shamp et al. [71]
BeH_n, $n = 2$–5	50, 150, 200	2, 4	VASP	Hooper et al. [72]
MgH_n, $n = 2$–16	0–250	2, 3, 4, 5, 6, 8	VASP	Lonie et al. [73]
BaH_n, $n = 2$–13	50–200	2, 4	VASP	Hooper et al. [72]
WH_n, $n = 1$–6;8	25, 150	1, 2, 3, 4	VASP	Labet et al. [74]
CH_4	20–800	2, 3, 4	VASP	Gao et al. [75]
	11	Not specified	VASP	Zhu et al. [29]
SiH_4	40–300	1, 2, 3, 4, 6	VASP	Martinez et al. [76]
GeH_4	50–250	1, 2, 3, 4	VASP	Gao et al. [77]
SnH_4	30–250	1, 2, 3, 4	VASP	Gao et al. [78]
PbH_4	0–500	2, 4	VASP	Zaleski-Ejgierd et al. [79]
NaH	0, 29.3	4	PWSCF, VASP	Lonie et al. [31]
CH	0–300	Fixed, up to 8	VASP	Wen et al. [80]
PtH_x, $x = \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, 1, 2, 3, 4$	120	6, 10, 12	VASP	Zhou et al. [81]
FeH_x $x = \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, 1, 2, 3, 4$	300, 400	2, 3, 4, 6, 8	VASP	Bazhanova et al. [82]
$LiBeH_3$	50–550	2, 4	VASP	Hu et al. [83]
$LiNH_2$	0–360	2, 4	VASP	Prasad et al. [84]
$NaAlH_4$	0–20	1, 2	VASP	Zhou et al. [85]
$NaPtH_2$	0	Not specified	VASP	Wen et al. [86]
$NaNH_2$	0, 10, 20	1, 2, 3, 4	VASP	Zhong et al. [87]
$SrFeH_4$	0	Not specified	VASP	Wen et al. [86]
$Mg(BH_4)_2$	0–20	2, 4	VASP	Zhou et al. [88]

6.2 *Hydrogen-Containing Compounds*

Table 4 summarizes EA structure searches performed on hydrogen-containing compounds. Ashcroft suggested in 1968 that hydrogen could become a high-temperature superconductor under pressure [110] and in 2006 that doping hydrogen to form chemically precompressed hydrogen-rich materials could be a potential route to reduce the pressure required for superconductivity [111].

Hooper, Lonie, and Zurek have performed several studies on polyhydrides of alkali and alkaline earth metals under pressure. A metallic phase of LiH_6

Table 5 Application of evolutionary algorithms to intermetallic compounds

Compound	Pressure (GPa)	Formula units per cell	Energy code	References
Al–Sc system	0	8 atoms	VASP	Trimarchi et al. [89]
	0	6, 8 atoms	VASP	Trimarchi et al. [49]
	0	8 atoms	VASP	Ji et al. [34]
$Al_{13}K$	0	1	VASP	Oganov et al. [37]
Au_2Pd	0	4	VASP	Trimarchi et al. [33]
	0	4	VASP	Trimarchi et al. [89]
$CaLi_2$	10–250	1, 2, 3, 4, 6, 8	VASP	Xie et al. [90]
$CdPt_3$	0	2	VASP	Trimarchi et al. [89]
CuPd	0	4	VASP	Trimarchi et al. [89]
Na–Ca system	50	Not specified	VASP	Taillon et al. [59]
$PdTi_3$	0	2	VASP	Trimarchi et al. [89]

Table 6 Application of evolutionary algorithms to minerals

Mineral	Pressure (GPa)	Formula units per cell	Energy code	References
Al_2O_3	300	4	VASP	Oganov et al. [25]
Al_2SiO_5	10	4	GULP	Zhu et al. [91]
$CaCO_3$	50, 80, 150	1, 2, 4	VASP	Oganov et al. [25]
FeC_x, $x = \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, 1, 2, \frac{7}{3}, 3, 4$	300, 400	2, 3, 4, 6, 8	VASP	Bazhanova et al. [82]
FeS	56, 120, 400	2, 3, 4, 6, 8	VASP	Ono et al. [92]
$FeSi_x$, $x = \frac{1}{3}, \frac{1}{2}, 1, \frac{5}{3}, 2, 3$	300, 400	8, 9, 12, 16, 18, 24 atoms	VASP	Zhang et al. [93]
$MgCO_3$	110, 150	1, 2, 4, 6	VASP	Oganov et al. [25, 37]
Mg-O system	Up to 850	Up to 20 atoms	VASP	Zhu et al. [94]
$MgSiO_3$	250	32	GULP	Zhu et al. [91]
	80, 120, 1,000	4, 8	VASP	Oganov et al. [25]
Na-Cl system	0–250	Up to 16 atoms	VASP	Zhang et al. [95]
SiO_2	500, 2,000	1–8	PWSCF	Wu et al. [96]
	0	3	VASP	Oganov et al. [25]
	10	24	GULP	Zhu et al. [91]
TiO_2	0	2, 4, 8	GULP	Woodley et al. [51]
	0	16	GULP	Lonie et al. [31]
	Not specified	4	VASP	Lonie et al. [31]

was predicted to be stable above 110 GPa [50], and a stable phase of NaH_9 was predicted to metallize at 250 GPa [68]. Stable rubidium polyhydride phases were predicted to metallize at pressures above 200 GPa [70]. A stable, superconducting phase of MgH_{12} was identified under pressure, with a predicted

Table 7 Application of evolutionary algorithms to molecular solids

Compound	Pressure (GPa)	Formula units per cell	Energy code	References
Benzene C_6H_6	0–300	Not specified	VASP	Wen et al. [80]
	0, 5, 10, 25	4	VASP	Zhu et al. [29]
Ice H_2O	1,000–4,000	8	LAMMPS, PWSCF	Ji et al. [53]
	0	4	VASP	Zhu et al. [29]
	0	4	VASP	Oganov et al. [25]
CO_2	50, 100, 150	1, 2, 3, 4, 6, 8	VASP	Oganov et al. [37]
	12–20	2	VASP	Zhu et al. [29]
	50	1, 2, 3, 4, 6, 8	VASP	Oganov et al. [25]
NH_3	5, 10, 25, 50	4	VASP	Zhu et al. [29]
Glycine	1, 2	2, 3, 4	VASP	Zhu et al. [29]
Butane-1,4- diammonium dibromide	0	2	VASP	Zhu et al. [29]
Urea $CO(NH_2)_2$	0	2	VASP	Oganov et al. [25]

T_c of 47–60 K at 140 GPa [73]. Lonie et al. also identified a new phase of BeH_2 above 150 GPa, as well as a stable superconducting phase of BaH_6, with a T_c of 30–38 K at 100 GPa [72].

Several studies have been performed on the group IV hydrides ranging from methane to plumbane. Gao et al. searched for methane structures under pressure and predicted that it dissociates into ethane and hydrogen at 95 GPa, butane and hydrogen at 158 GPa, and finally carbon and hydrogen at 287 GPa [75]. Martinez et al. looked at silane under pressure and predicted two new phases, one stable from 25 to 50 GPa, and the other from 220 to 250 GPa. The latter was predicted to be superconducting, with a T_c of 16 K at 220 GPa [76]. Gao et al. searched for germane and stannane under pressure. Germane was predicted to be stable with respect to decomposition into pure germanium and hydrogen above 196 GPa, and it was predicted to be superconducting, with a T_c of 64 K at 220 GPa [77]. Two stannane isomers were predicted – one stable from 96 to 180 GPa, and the other occurring above 180 GPa. Both phases were calculated to be superconductors [78]. Zaleski et al. performed evolutionary structure searches for plumbane (PbH_4) under pressure and predicted that it forms a stable non-molecular solid at pressures greater than 132 GPa [79]. Wen et al. predicted five low energy three-dimensional structures of graphane in the pressure range 0–300 GPa, and each was either semiconducting or insulating [80].

Zhou et al. investigated platinum hydrides under pressure and predicted a superconducting hexagonal phase of PtH to be stable above 113 GPa [81]. Hu et al. searched for $LiBeH_3$ at pressures up to 530 GPa and predicted two new insulating phases [83]. Zhou et al. found two new tetragonal structures of $Mg(BH_4)_2$ under pressure whose densities, bulk moduli, and XRD patterns match experimentally measured values [88].

Table 8 Application of evolutionary algorithms to various inorganic compounds

Compound	Pressure (GPa)	Formula units per cell	Energy code	References
$Al_{12}C$	0	1	VASP	Oganov et al. [37]
BeB_2, BeB_3, BeB_4, $BeB_{2.75}$	0, 160	4, 8	VASP	Hermann et al. [97]
CsI	5–300	2, 3, 4, 8	VASP	Xu et al. [98]
Fe-B system	0	Up to 15 atoms	VASP	Kolmogorov [54]
GaAs	0	4	VASP	Trimarchi et al. [33]
Li-B system	0–320	1, 2	VASP	Hermann et al. [99, 100]
Li-Si system	0	Variable, up to 20 atoms	VASP	Tipton et al. [48]
MgB_2	5–300	1, 2, 3, 4, 6	VASP	Ma et al. [101]
SiC	0	4	VASP	Trimarchi et al. [33]
WN_2	0, 60	1, 2, 3, 4	VASP	Wang et al. [102]
XeF_2	0–200	Up to 4	VASP	Kurzydlowski et al. [103]
Xe-O System	5–220	Up to 36 atoms	VASP	Zhu et al. [104]
$ATiO_2$, A=Ba, Be, Ca, Mg, Sr	0	Not specified	VASP	Wen et al. [105]
$NaPtF_2$	0	Not specified	VASP	Wen et al. [86, 105]
$SrFeC_4$	0	Not specified	VASP	Wen et al. [86, 105]
$CaRhO_3$	Not specified	6	VASP	Shirako et al. [106]
Li_3RuO_4	0	Not specified	GULP	Bush et al. [2]
$SrTiO_3$	0	10	GULP	Lonie et al. [31]
BC_2N	30, 100	1, 2, 4	VASP	Li et al. [107]
LiBeB	30, 100	Not specified	VASP	Hermann et al. [108]
Si_2N_2O	0	2	VASP	Oganov et al. [25]
$SrSiN_2$	0	4	VASP	Oganov et al. [25]

6.3 *Intermetallic Compounds*

Table 5 summarizes searches for intermetallic compounds. Trimarchi et al. have studied many intermetallics. Their algorithm identified the correct lattice of Au_2Pd, which is known to be fcc, but it failed to find the lowest energy atomic configuration [33] because the system exhibits several nearly degenerate structures. In another study by these authors, a new phase of IrN_2 was discovered [89]. They performed a phase diagram search on the Al–Sc system, which exhibit several crystal structures across the composition range, and successfully identified the experimentally known ground state phases [49]. Xie et al. discovered two new superconducting phases of $CaLi_2$, one stable from 35 to 54 GPa and the other stable from 54 to 105 GPa. Furthermore, they predict that $CaLi_2$ is unstable with respect to dissociation into the constituents at pressures greater than 105 GPa [90].

Sometimes the underlying lattices for these systems are known empirically, and the search reduces to finding the lowest energy arrangement of atoms on the lattice. For these cases, an efficient method to search over permutations of atomic positions is crucial. D'Avezac et al. employed a virtual atom technique, in which the species type of an atom is "relaxed" to determine if exchanging atom types at that site would likely lead to a lower energy configuration [112]. In contrast to real space mating operations, these authors also employed a reciprocal space mating scheme [112]. With this technique, the structure factors of two parent organisms are combined to form the offspring organism's structure factor, which is then transformed to real space, giving the offspring organism. A cluster expansion fitted to DFT calculations was leveraged to perform energy calculations.

6.4 Minerals

Table 6 contains a summary of EA searches for the structures of several minerals. Many of these studies were carried out at high pressure in order to simulate the conditions in planetary interiors. Oganov et al. found two new phases of $MgCO_3$ under pressure. In addition, a new phase of $CaCO_3$ was reported to be stable above 137 GPa [37], and the structure of the post-aragonite phase of $CaCO_3$, stable from 42 to 137 GPa, was solved [16]. Iron carbides of various compositions were explored under pressure, and it was predicted that the cementite structure of Fe_3C is unstable at pressures above 310 GPa, indicating that this phase does not exist in the Earth's inner core [82]. Zhang et al. searched for iron silicide structures under pressure and predicted that only FeSi with a cesium chloride structure is stable at pressures greater than 20 GPa. Ono et al. investigated the structure of FeS under pressure and reported several new phases up to 135 GPa [92]. Wu et al. predicted a new low temperature post-perovskite phase of SiO_2 with the Fe_2P-type structure [96].

6.5 Molecular Crystals

Table 7 summarizes searches for molecular crystals. Applications for molecular solids include high-energy materials, pharmaceuticals, pigments, and metal-organic frameworks [113–115]. Molecular solids are not always in their thermodynamic ground states. Instead, the system is kinetically trapped and the molecular units are maintained. Zhu et al. made several changes to the standard EA to facilitate searching for molecular crystals, and they applied their algorithm to search for structures of ice, methane, ammonia, carbon dioxide, benzene, glycine, and butane-1,4-diammonium dibromide. Experimentally known structures were recovered by the algorithm [29]. Ji et al. searched for ice at terapascal pressures and predicted three new phases [53]. Lennard-Jones potentials were used to model the system, but

ab initio calculations were periodically performed and the results used to fit the empirical potentials. Hermann et al. performed evolutionary searches for high-pressure phases of ice and predicted that ice becomes metallic at 4.8 TPa [116]. Oganov et al. predicted that the β-cristobalite structure is the most stable for CO_2 between 19 and 150 GPa [37].

6.6 *Inorganic Compounds*

Table 8 summarizes searches for inorganic compounds. Many of these studies aimed to clarify regions of various phase diagrams. Others sought to identify phases with desirable properties, such as superconductivity.

Tipton et al. applied an evolutionary algorithm to investigate Li-Si anode battery materials and carried out a phase diagram search on the Li–Si system. They discovered a new stable phase with composition Li_5Si_2 [48]. Hermann et al. searched for structures in the Li–B system under pressure and found several stable structures. LiB was found to become increasingly stable as the pressure was increased beyond 300 GPa [99, 100]. Hermann et al. also predicted a new stable phase of LiBeB at ambient pressure and several additional phases under pressures up to 320 GPa [108]. Kolmogorov et al. searched for Fe–B structures at several different compositions and reported new phases with compositions FeB_4 and FeB_2.

Xu et al. discovered a new orthorhombic phase of CsI that is predicted to be stable from 42 GPa up to at least 300 GPa [98]. This material is predicted to metallize at 100 GPa and to become superconducting at 180 GPa. Zhu et al. searched for xenon oxides under pressure and found three stable compounds: XeO above 83 GPa, XeO_2 above 102 GPa, and XeO_3 above 114 GPa [104]. Bush et al. solved the structure of Li_3RuO_4 at zero pressure. They used an alternative fitness function in their EA and only calculated energies at the end of the search [2]. Li et al. resolved the structure of superhard BC_2N and found it to have a rhombohedral lattice [107].

7 Conclusions

Crystal structure prediction is a long-standing challenge in the physical sciences. If we frame the structure prediction problem as one of global optimization, robust and accurate free energy methods such as those described in Sect. 4 can be used as objective functions, which can be minimized to find the thermodynamically stable structure. This minimization problem has been effectively addressed in recent years using evolutionary algorithms. However, the EA is less of a particular algorithm and more of a general problem solving strategy. Thus, many methodological and design choices are possible when creating an EA for structure prediction, and innovation in the field is ongoing.

The method generally begins with a broad sampling of the solution phase space. Information gained from early calculations is used to try to guess new low energy candidate structures. The ability to make such inferences relies on some characterization of or knowledge about the structure of the energy landscape, as described in Sect. 1.1. In the evolutionary approach, we leverage the power of biological evolution to search for low energy structures. Parent structures that are good solutions to the problem are varied and combined in such a way so as to pass down their traits to children. In this way, favorable properties are propagated in the population while unfavorable ones tend to die out. The parent selection and variation operators are very important to the success of this approach, and the most common and successful variation operators take advantage of the partial spatial separability of the energy minimization problem. These and other methodological issues were discussed in Sect. 2.

In Sect. 6 we reviewed many applications of the method to systems of fundamental scientific as well as technological interest. As the field has begun to mature, researchers have had many successes in practical applications. Several publicly available software packages for performing these calculations exist and are described in Sect. 5. However, evolutionary algorithms are not yet a commodity method, and an understanding of the methodology remains helpful for obtaining best results.

A number of challenges remain. Since ab initio energy calculations are the most expensive part of the method, reducing the number of these necessary to obtain high quality predictions is the focus of methodological developments. The energy and relaxed structure are a small subset of the information provided by ab initio calculations, so the opportunity exists to improve results by making better use of the full set of data. Increasing the efficiency of the search through solution representation and variation operators is also a promising avenue of improvement. Searching over structures' compositional degrees of freedom remains inefficient since there is no local relaxation of these parameters, and the energy landscape has little structure with respect to them. This makes prediction of the number of atoms in the unit cell and phase diagram prediction challenging, as discussed in Sect. 3.

Although they lie beyond the scope of this review and volume, uses of evolutionary algorithms to predict the atomic structures of other systems, such as surfaces, 1D and 2D materials, atomic clusters, or molecules, etc., are also important and active areas of research. A primary goal of computational materials science is to find materials with desirable properties, and structure prediction is a necessary early step in first principles prediction of materials' properties. Evolutionary algorithms have turned out to be a useful global optimization method for addressing the structure prediction problem.

Acknowledgments This work was supported by the National Science Foundation under award number CAREER DMR-1056587 and by the Energy Materials Center at Cornell (EMC2) funded by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under award number DE-SC0001086. W.W.T. was supported by the NSF IGERT Fellowship Program "A Graduate Traineeship in Materials for a Sustainable Future" under award number DGE-0903653 and the NSF GK12 Program "Grass Roots: Advancing education in renewable energy and cleaner fuels through collaborative graduate fellow/teacher/grade-school student

interactions" under award number DGE-1045513. This research used computational resources of the Texas Advanced Computing Center under Contract Number TG-DMR050028N and of the Computation Center for Nanotechnology Innovation at Rensselaer Polytechnic Institute.

References

1. van Dover RB, Schneemeyer L, Fleming R (1998) Discovery of a useful thin-film dielectric using a composition-spread approach. Nature 392(6672):162–164
2. Bush TS, Catlow CRA, Battle PD (1995) Evolutionary programming techniques for predicting inorganic crystal structures. J Mater Chem 5:1269–1272
3. Ceder G, Morgan D, Fischer C, Tibbetts K, Curtarolo S (2006) Data-mining-driven quantum mechanics for the prediction of structure. MRS Bull 31(12):981–985
4. Goedecker S (2004) Minima hopping: an efficient search method for the global minimum of the potential energy surface of complex molecular systems. J Chem Phys 120:9911
5. Martoňnák R, Laio A, Parrinello M (2003) Predicting crystal structures: the Parrinello–Rahman method revisited. Phys Rev Lett 90(7):075503
6. Pannetier J, Bassas-Alsina J, Rodriguez-Carvajal J, Caignaert V (1990) Prediction of crystal structures from crystal chemistry rules by simulated annealing. Nature 346(6282):343–345
7. Pickard CJ, Needs RJ (2011) Ab initio random structure searching. J Phys Condens Matter 23(5):053201
8. Wang Y, Lv J, Zhu L, Ma Y (2012) CALYPSO: a method for crystal structure prediction. Comput Phys Commun 183(10):2063–2070
9. Feng J, Hennig RG, Ashcroft NW, Hoffmann R (2008) Emergent reduction of electronic state dimensionality in dense ordered Li–Be alloys. Nature 451(7177):445–448
10. Rudin SP, Jones MD, Albers RC (2004) Thermal stabilization of the HCP phase in titanium. Phys Rev B 69(9):094117
11. Souvatzis P, Eriksson O, Katsnelson MI, Rudin SP (2008) Entropy driven stabilization of energetically unstable crystal structures explained from first principles theory. Phys Rev Lett 100:095901
12. Woodley MS, Battle DP, Gale DJ, Catlow RAC (1999) The prediction of inorganic crystal structures using a genetic algorithm and energy minimisation. Phys Chem Chem Phys 1:2535–2542
13. Stillinger FH (1999) Exponential multiplicity of inherent structures. Phys Rev E 59(1):48
14. Venkatesh PK, Cohen MH, Carr RW, Dean AM (1997) Bayesian method for global optimization. Phys Rev E 55:6219–6232
15. Massen CP, Doye JP (2007) Power-law distributions for the areas of the basins of attraction on a potential energy landscape. Phys Rev E 75(3):037101
16. Oganov AR, Glass CW, Ono S (2006) High-pressure phases of $CaCO_3$: crystal structure prediction and experiment. Earth Planet Sci Lett 241(1):95–103
17. Brodmeier T, Pretsch E (1994) Application of genetic algorithms in molecular modeling. J Comput Chem 15(6):588–595
18. Dandekar T, Argos P (1994) Folding the main chain of small proteins with the genetic algorithm. J Mol Biol 236(3):844
19. Lucasius CB, Werten S, van Aert A, Kateman G, Blommers MJ (1991) Conformational analysis of DNA using genetic algorithms. In: Schwefel HP, Maenner R (eds) Parallel problem solving from nature. Springer, Berlin, pp 90–97
20. McGarrah D, Judson RS (1993) Analysis of the genetic algorithm method of molecular conformation determination. J Comput Chem 14(11):1385–1395
21. Sun S (1993) Reduced representation model of protein structure prediction: statistical potential and genetic algorithms. Protein Sci 2(5):762–785

22. Deaven DM, Ho KM (1995) Molecular geometry optimization with a genetic algorithm. Phys Rev Lett 75:288–291
23. Xiao Y, Williams DE (1993) Genetic algorithm: a new approach to the prediction of the structure of molecular clusters. Chem Phys Lett 215(1):17–24
24. Smith RW (1992) Energy minimization in binary alloy models via genetic algorithms. Comput Phys Commun 71(1):134–146
25. Oganov AR, Glass CW (2006) Crystal structure prediction using ab initio evolutionary techniques: principles and applications. J Chem Phys 124(24):244704
26. Tipton WW, Hennig RG (2013) Genetic algorithm for structure and phase prediction. Cornell University. http://gasp.mse.cornell.edu/. Accessed 22 Oct 2013
27. Tipton WW, Hennig RG (2013) A grand canonical genetic algorithm for the prediction of multicomponent phase diagrams and testing empirical potentials. J Phys Cond Matter 25:495401
28. Pickard CJ, Needs RJ (2006) High-pressure phases of silane. Phys Rev Lett 97:045504
29. Zhu Q, Oganov AR, Glass CW, Stokes HT (2012) Constrained evolutionary algorithm for structure prediction of molecular crystals: methodology and applications. Acta Crystallogr Sect B Struct Sci 68(3):215–226
30. Bahmann S, Kortus J (2013) EVO – evolutionary algorithm for crystal structure prediction. Comput Phys Commun 184(6):1618–1625
31. Lonie DC, Zurek E (2011) Xtalopt: an open-source evolutionary algorithm for crystal structure prediction. Comput Phys Commun 182(2):372–387
32. Glass CW, Oganov AR, Hansen N (2006) USPEX – evolutionary crystal structure prediction. Comput Phys Commun 175(1112):713–720
33. Trimarchi G, Zunger A (2007) Global space-group optimization problem: finding the stablest crystal structure without constraints. Phys Rev B 75:104113
34. Ji M, Wang CZ, Ho KM (2010) Comparing efficiencies of genetic and minima hopping algorithms for crystal structure prediction. Phys Chem Chem Phys 12(37):11617–11623
35. Woodley SM (2004) Prediction of inorganic crystal framework structures part 2: using a genetic algorithm and a direct approach to exclusion zones. Phys Chem Chem Phys 6:1823–1829
36. Kŕivý I, Gruber B (1976) A unified algorithm for determining the reduced (Niggli) cell. Acta Crystallogr A 32(2):297–298
37. Oganov AR, Glass CW (2008) Evolutionary crystal structure prediction as a tool in materials design. J Phys Condens Matter 20(6):064210
38. Abraham NL, Probert MIJ (2006) A periodic genetic algorithm with real-space representation for crystal structure and polymorph prediction. Phys Rev B 73:224104
39. Johnston RL (2003) Evolving better nanoparticles: genetic algorithms for optimising cluster geometries. Dalton Trans 4193–4207
40. Lyakhov AO, Oganov AR, Valle M (2010) How to predict very large and complex crystal structures. Comput Phys Commun 181(9):1623–1632
41. Oganov AR (2013) Evolutionary crystal structure prediction and computational materials discovery. doi: 10.1007/128_2013_508
42. Lonie DC, Zurek E (2012) Identifying duplicate crystal structures: Xtalcomp, an open-source solution. Comput Phys Commun 183(3):690–697
43. Valle M, Oganov AR (2008) Crystal structures classifier for an evolutionary algorithm structure predictor. In: IEEE symposium on visual analytics science and technology VAST'08, 2008. Columbus, Ohio, pp 11–18
44. Valle M, Oganov AR (2010) Crystal fingerprint space – a novel paradigm for studying crystalstructure sets. Acta Crystallogr A 66(5):507–517
45. Abraham NL, Probert MIJ (2008) Improved real-space genetic algorithm for crystal structure and polymorph prediction. Phys Rev B 77:134117
46. Hartke B (1993) Global geometry optimization of clusters using genetic algorithms. J Phys Chem 97(39):9973–9976

47. Zhang X, Trimarchi G, Zunger A (2009) Possible pitfalls in theoretical determination of groundstate crystal structures: the case of platinum nitride. Phys Rev B 79:092102
48. Tipton WW, Bealing CR, Mathew K, Hennig RG (2013) Structures, phase stabilities, and electrical potentials of Li-Si battery anode materials. Phys Rev B 87:184114
49. Trimarchi G, Freeman AJ, Zunger A (2009) Predicting stable stoichiometries of compounds via evolutionary global space-group optimization. Phys Rev B 80:092101
50. Zurek E, Hoffmann R, Ashcroft NW, Oganov AR, Lyakhov AO (2009) A little bit of lithium does a lot for hydrogen. Proc Natl Acad Sci U S A 106(42):17640–17643
51. Woodley SM, Catlow CRA (2009) Structure prediction of titania phases: implementation of Darwinian versus Lamarckian concepts in an evolutionary algorithm. Comput Mater Sci 45 (1):84–95
52. Park H, Fellinger MR, Lenosky TJ, Tipton WW, Trinkle DR, Rudin SP, Woodward C, Wilkins JW, Hennig RG (2012) Ab initio based empirical potential used to study the mechanical properties of molybdenum. Phys Rev B 85:214121
53. Ji M, Umemoto K, Wang CZ, Ho KM, Wentzcovitch RM (2011) Ultrahigh-pressure phases of H_2O ice predicted using an adaptive genetic algorithm. Phys Rev B 84:220105
54. Kolmogorov AN, Shah S, Margine ER, Bialon AF, Hammerschmidt T, Drautz R (2010) New superconducting and semiconducting Fe–B compounds predicted with an ab initio evolutionary search. Phys Rev Lett 105(21):217003
55. Chuang F, Ciobanu CV, Shenoy V, Wang CZ, Ho KM (2004) Finding the reconstructions of semiconductor surfaces via a genetic algorithm. Surf Sci 573(2):L375–L381
56. Lu N, Ciobanu CV, Chan TL, Chuang FC, Wang CZ, Ho KM (2007) The structure of ultrathin H-passivated [112] silicon nanowires. J Phys Chem C 111(22):7933–7937
57. Woodley SM (2007) Engineering microporous architectures: combining evolutionary algorithms with predefined exclusion zones. Phys Chem Chem Phys 9(9):1070–1077
58. Oganov AR, Chen J, Gatti C, Ma Y, Ma Y, Glass CW, Liu Z, Yu T, Kurakevych OO, Solozhenko VL (2009) Ionic high-pressure form of elemental boron. Nature 457(7231):863–867
59. Taillon JA, Tipton WW, Hennig RG (2012) Ab initio discovery of novel crystal structure stability in barium and sodium-calcium compounds under pressure using DFT. arXiv preprint arXiv:1207.3320
60. Li Q, Ma Y, Oganov AR, Wang H, Wang H, Xu Y, Cui T, Mao HK, Zou G (2009) Superhard monoclinic polymorph of carbon. Phys Rev Lett 102(17):175,506
61. Oganov AR, Lyakhov AO (2010) Towards the theory of hardness of materials. J Superhard Mater 32(3):143–147
62. Bi W, Meng Y, Kumar RS, Cornelius AL, Tipton WW, Hennig RG, Zhang Y, Chen C, Schilling JS (2011) Pressure-induced structural transitions in europium to 92 GPa. Phys Rev B 83(10):104106
63. Ma Y, Oganov AR, Xie Y (2008) High-pressure structures of lithium, potassium, and rubidium predicted by an ab initio evolutionary algorithm. Phys Rev B 78(1):014102
64. Ma Y, Oganov AR, Li Z, Xie Y, Kotakoski J (2009) Novel high pressure structures of polymeric nitrogen. Phys Rev Lett 102(6):065501
65. Ma Y, Eremets M, Oganov AR, Xie Y, Trojan I, Medvedev S, Lyakhov AO, Valle M, Prakapenka V (2009) Transparent dense sodium. Nature 458(7235):182–185
66. Ma Y, Oganov AR, Glass CW (2007) Structure of the metallic ζ-phase of oxygen and isosymmetric nature of the ε-ζ phase transition: ab initio simulations. Phys Rev B 76 (6):064101
67. Hooper J, Zurek E (2012) Lithium subhydrides under pressure and their superatom-like building blocks. ChemPlusChem 77(11):969–972
68. Baettig P, Zurek E (2011) Pressure-stabilized sodium polyhydrides: NaH_n ($n>1$). Phys Rev Lett 106(23):237002
69. Hooper J, Zurek E (2012) High pressure potassium polyhydrides: a chemical perspective. J Phys Chem C 116(24):13322–13328

70. Hooper J, Zurek E (2012) Rubidium polyhydrides under pressure: emergence of the linear H_3^- species. Chemistry 18(16):5013–5021
71. Shamp A, Hooper J, Zurek E (2012) Compressed cesium polyhydrides: Cs^+ sublattices and H_3^- three-connected nets. Inorg Chem 51(17):9333–9342
72. Hooper J, Altintas B, Shamp A, Zurek E (2013) Polyhydrides of the alkaline earth metals: a look at the extremes under pressure. J Phys Chem C 117(6):2982–2992
73. Lonie DC, Hooper J, Altintas B, Zurek E (2013) Metallization of magnesium polyhydrides under pressure. Phys Rev B 87(5):054107
74. Labet V, Hoffmann R, Ashcroft NW (2011) Molecular models for WH6 under pressure. New J Chem 35(10):2349–2355
75. Gao G, Oganov AR, Ma Y, Wang H, Li P, Li Y, Iitaka T, Zou G (2010) Dissociation of methane under high pressure. J Chem Phys 133:144508
76. Martinez-Canales M, Oganov AR, Ma Y, Yan Y, Lyakhov AO, Bergara A (2009) Novel structures and superconductivity of silane under pressure. Phys Rev Lett 102(8):087005
77. Gao G, Oganov AR, Bergara A, Martinez-Canales M, Cui T, Iitaka T, Ma Y, Zou G (2008) Superconducting high pressure phase of germane. Phys Rev Lett 101(10):107002
78. Gao G, Oganov AR, Li P, Li Z, Wang H, Cui T, Ma Y, Bergara A, Lyakhov AO, Iitaka T et al (2010) High-pressure crystal structures and superconductivity of stannane (SnH_4). Proc Natl Acad Sci U S A 107(4):1317–1320
79. Zaleski-Ejgierd P, Hoffmann R, Ashcroft N (2011) High pressure stabilization and emergent forms of PbH_4. Phys Rev Lett 107(3):037002
80. Wen XD, Hand L, Labet V, Yang T, Hoffmann R, Ashcroft N, Oganov AR, Lyakhov AO (2011) Graphane sheets and crystals under pressure. Proc Natl Acad Sci U S A 108(17):6833–6837
81. Zhou XF, Oganov AR, Dong X, Zhang L, Tian Y, Wang HT (2011) Superconducting highpressure phase of platinum hydride from first principles. Phys Rev B 84(5):054543
82. Bazhanova ZG, Oganov AR, Gianola O (2012) Fe–C and Fe–H systems at pressures of the Earth's inner core. Phys Uspekhi 55(5):489
83. Hu CH, Oganov AR, Lyakhov AO, Zhou HY, Hafner J (2009) Insulating states of $LiBeH_3$ under extreme compression. Phys Rev B 79(13):134116
84. Prasad DL, Ashcroft N, Hoffmann R (2012) Lithium amide (LiNH2) under pressure. J Phys Chem A 116(40):10027–10036
85. Zhou XF, Dong X, Zhao Z, Oganov AR, Tian Y, Wang HT (2012) High-pressure phases of $NaAlH_4$ from first principles. Appl Phys Lett 100(6):061905–061905
86. Wen XD, Cahill TJ, Gerovac NM, Bucknum MJ, Hoffmann R (2009) Playing the quantum chemical slot machine: an exploration of ABX_2 compounds. Inorg Chem 49(1):249–260
87. Zhong Y, Zhou HY, Hu CH, Wang DH, Oganov AR (2012) Theoretical studies of highpressure phases, electronic structure, and vibrational properties of $NaNH_2$. J Phys Chem C 116(15):8387–8393
88. Zhou XF, Oganov AR, Qian GR, Zhu Q (2012) First-principles determination of the structure of magnesium borohydride. Phys Rev Lett 109(24):245503
89. Trimarchi G, Zunger A (2008) Finding the lowest-energy crystal structure starting from randomly selected lattice vectors and atomic positions: first-principles evolutionary study of the Au–Pd, Cd–Pt, Al–Sc, Cu–Pd, Pd–Ti, and Ir–N binary systems. J Phys Condens Matter 20 (29):295212
90. Xie Y, Oganov AR, Ma Y (2010) Novel high pressure structures and superconductivity of $CaLi_2$. Phys Rev Lett 104(17):177005
91. Zhu Q, Oganov AR, Lyakhov AO (2012) Evolutionary metadynamics: a novel method to predict crystal structures. CrystEngComm 14(10):3596–3601
92. Ono S, Oganov AR, Brodholt JP, Vočadlo L, Wood IG, Lyakhov A, Glass CW, Côté AS, Price GD (2008) High-pressure phase transformations of FeS: novel phases at conditions of planetary cores. Earth Planet Sci Lett 272(1):481–487

93. Zhang F, Oganov AR (2010) Iron silicides at pressures of the earth's inner core. Geophys Res Lett 37(2), L02305
94. Zhu Q, Oganov AR, Lyakhov AO (2012) Unexpected stoichiometries in Mg-O system under high pressure. arXiv preprint arXiv:1211.6521
95. Zhang W, Oganov AR, Goncharov AF, Zhu Q, Boulfelfel SE, Lyakhov AO, Somayazulu M, Prakapenka VB (2012) Unexpected stable stoichiometries of sodium chlorides. arXiv preprint arXiv:1211.3644
96. Wu S, Umemoto K, Ji M, Wang CZ, Ho KM, Wentzcovitch RM (2011) Identification of post-pyrite phase transitions in SiO_2 by a genetic algorithm. Phys Rev B 83:184102
97. Hermann A, Ashcroft N, Hoffmann R (2012) Making sense of boron-rich binary Be–B phases. Inorg Chem 51(16):9066–9075
98. Xu Y, John ST, Oganov AR, Cui T, Wang H, Ma Y, Zou G (2009) Superconducting high-pressure phase of cesium iodide. Phys Rev B 79(14):144110
99. Hermann A, McSorley A, Ashcroft NW, Hoffmann R (2012) From Wade–Mingos to Zintl–Klemm at 100 GPa: binary compounds of boron and lithium. J Am Chem Soc 134 (45):18606–18618
100. Hermann A, Suarez-Alcubilla A, Gurtubay IG, Yang LM, Bergara A, Ashcroft NW, Hoffmann R (2012) LiB and its boron-deficient variants under pressure. Phys Rev B 86(14):144110
101. Ma Y, Wang Y, Oganov AR (2009) Absence of superconductivity in the high-pressure polymorph of MgB_2. Phys Rev B 79(5):054101
102. Wang H, Li Q, Li Y, Xu Y, Cui T, Oganov AR, Ma Y (2009) Ultra-incompressible phases of tungsten dinitride predicted from first principles. Phys Rev B 79(13):132109
103. Kurzydłowski D, Zaleski-Ejgierd P, Grochala W, Hoffmann R (2011) Freezing in resonance structures for better packing: XeF_2 becomes $(XeF_+)(F_-)$ at large compression. Inorg Chem 50(8):3832–3840
104. Zhu Q, Jung DY, Oganov AR, Glass CW, Gatti C, Lyakhov AO (2012) Stability of xenon oxides at high pressures. Nat Chem 5(1):61–65
105. Wen XD, Cahill TJ, Hoffmann R, Miura A (2009) Tuning of metal–metal bonding by counterion size in hypothetical $AeTiO_2$ compounds. J Am Chem Soc 131(41):14632–14633
106. Shirako Y, Kojitani H, Oganov A, Fujino K, Miura H, Mori D, Inaguma Y, Yamaura K, Akaogi M (2012) Crystal structure of CaRhO3 polymorph: high-pressure intermediate phase between perovskite and post-perovskite. Am Mineral 97(1):159–163
107. Li Q, Wang M, Oganov AR, Cui T, Ma Y, Zou G (2009) Rhombohedral superhard structure of BC_2N. J Appl Phys 105(5):053514–053514
108. Hermann A, Ivanov B, Ashcroft NW, Hoffmann R (2012) LiBeB: a predicted phase with structural and electronic peculiarities. Phys Rev B 86(1):014104
109. Bi W, Souza-Neto NM, Haskel D, Fabbris G, Alp EE, Zhao J, Hennig RG, Abd- Elmeguid MM, Meng Y, McCallum RW, Dennis K, Schilling JS (2012) Synchrotron X-ray spectroscopy studies of valence and magnetic state in europium metal to extreme pressures. Phys Rev B 85:205134
110. Ashcroft NW (1968) Metallic hydrogen: a high-temperature superconductor? Phys Rev Lett 21:1748–1749
111. Feng J, Grochala W, Jaro´n T, Hoffmann R, Bergara A, Ashcroft N (2006) Structures and potential superconductivity in SiH4 at high pressure: en route to metallic hydrogen. Phys Rev Lett 96(1):017006
112. dAvezac M, Zunger A (2008) Identifying the minimum-energy atomic configuration on a lattice: Lamarckian twist on Darwinian evolution. Phys Rev B 78(6):064102
113. Agrawal J (1998) Recent trends in high-energy materials. Prog Energ Combust 24(1):1–30
114. Baburin I, Leoni S, Seifert G (2008) Enumeration of not-yet-synthesized zeolitic zinc imidazolate MOF networks: a topological and DFT approach. J Phys Chem B 112(31):9437–9443

115. Price SL (2004) The computational prediction of pharmaceutical crystal structures and polymorphism. Adv Drug Deliv Rev 56(3):301–319
116. Hermann A, Ashcroft NW, Hoffmann R (2012) High pressure ices. Proc Natl Acad Sci U S A 109(3):745–750
117. Johannesson GH, Bligaard T, Ruban AV, Skriver HL, Jacobsen KW, Nørskov JK (2002) Combined electronic structure and evolutionary search approach to materials design. Phys Rev Lett 88(25):255506

Top Curr Chem (2014) 345: 223–256
DOI: 10.1007/128_2013_508
© Springer-Verlag Berlin Heidelberg 2014
Published online: 15 February 2014

Crystal Structure Prediction and Its Application in Earth and Materials Sciences

Qiang Zhu, Artem R. Oganov, and Xiang-Feng Zhou

Abstract Evolutionary algorithms, based on physically motivated forms of variation operators and local optimization, proved to be a powerful approach in determining the crystal structure of materials. This review summarized the recent progress of the USPEX method as a tool for crystal structure prediction. In particular, we highlight the methodology in (1) prediction of molecular crystal structures and (2) variable-composition structure predictions, and their applications to a series of systems, including $Mg(BH_4)_2$, Xe-O, Mg-O compounds, etc. We demonstrate that this method has a wide field of applications in both computational materials design and studies of matter at extreme conditions.

Keywords Crystal structure prediction · Molecular crystals · Variable composition · High pressure · Novel compounds · Ab initio simulations · Density functional theory

Q. Zhu
Department of Geosciences, Center for Materials by Design, Institute for Advanced Computational Science, SUNY Stony Brook, New York, NY 11794-2100, USA

A.R. Oganov (✉)
Department of Geosciences, Center for Materials by Design, Institute for Advanced Computational Science, SUNY Stony Brook, New York, NY 11794-2100, USA

Moscow Institute of Physics and Technology, 9 Institutskiy Lane, Dolgoprudny City, Moscow Region 141700, Russia

School of Materials Science, Northwestern Polytechnical University, Xi'an 710072, China
e-mail: artem.oganov@sunysb.edu

X.-F. Zhou
Department of Geosciences, Center for Materials by Design, Institute for Advanced Computational Science, SUNY Stony Brook, New York, NY 11794-2100, USA

School of Physics, Nankai University, Tianjin 300071, China

Contents

In thermodynamic equilibrium, and at temperatures below melting, materials tend to form crystalline states, which possess long-range order and translational symmetry. Understanding the structure of materials is crucial for understanding their properties. However, the prediction of crystal structure has been a long-standing challenge in physical science. Back in 1988, Maddox summarized this problem with the following words [1]:

> One of the continuing scandals in the physical sciences is that it remains in general impossible to predict the structure of even the simplest crystalline solids from knowledge of their chemical composition. . . Solids such as crystalline water (ice) are still thought to lie beyond mortals' ken.

Over the next few years, programs started appearing that attempted to do just this and, in 1994, Gavezzotti [2] addressed the fundamental questions "Are crystal structures predictable?" The answer was again asserted as "No."

Crystal structure prediction (CSP) is particularly necessary when crystal structure information is not readily available. At normal conditions, the crystal structure of most materials can be trivially determined by modern experimental techniques such as X-ray diffraction. However, the same treatment becomes extremely problematic when it comes to extreme conditions, and computer simulation becomes essential for obtaining structural information. Not only at extreme but also at normal conditions crystal structure prediction is of enormous value – this is one of the most fundamental problems in materials science and a necessary key step in computational materials discovery.

What do we mean precisely by CSP problem? The simplest and most important case is to find, at given pressure (and temperature) conditions, the stable crystal

structure knowing only the chemical formula [3].[1] Many types of advanced techniques have been proposed to address this problem [4–13] and these are described in a recent book [3]. Among these methods, the USPEX method [13–18], based on evolutionary algorithm, is the leading one, and has been viewed as a revolution in crystallography [19]. It has led to many exciting discoveries, early examples of which, confirmed by experiment, include the superhard phase of boron with partially ionic bonding [20], transparent insulating phase of sodium [21], etc. In this chapter we will give an overview of the modern crystal structure prediction field, and particularly the methodology and a few recent applications based on evolutionary algorithms. Discussions here follow closely those in [13, 15, 18].

1 Methodology

1.1 Energy Landscape

Before talking about the prediction of the crystal structure, let us first consider the energy landscape that needs to be explored. The number of distinct points on the landscape can be estimated as:

$$C = \binom{V/\delta^3}{N} \prod \binom{N}{n_i},$$

where N is the number of atoms in the unit cell of volume V, δ is a relevant discretization parameter (for instance, 1 Å), and n_i is the number of atoms of ith type in the unit cell. Even for small systems ($N \approx 10$), C is astronomically large (roughly 10^N if one uses $\delta = 1$ Å and a typical atomic volume of 10 Å^3). Such an enormous number of structures cannot possibly be sampled, even on the most advanced supercomputer, making direct solution of the CSP impossible.

The dimensionality of the energy landscape is

$$d = 3N + 3,$$

where $3N - 3$ degrees of freedom are from N atoms, and the remaining six dimensions are defined by the lattice. CSP is an NP-hard problem, and the difficulty increases exponentially with the dimensionality. Yet great simplification can be achieved if structures are relaxed, i.e., brought to the nearest local energy minima. Relaxation introduces some intrinsic chemical constraints (bond lengths, bond angles, avoidance of unfavorable contacts). Therefore, the intrinsic dimensionality can be reduced:

[1] Two extended formulations of this problem include simultaneous searches for stable chemical compositions and structures in multicomponent systems, and finding the structures (and compositions) that possess required physical properties.

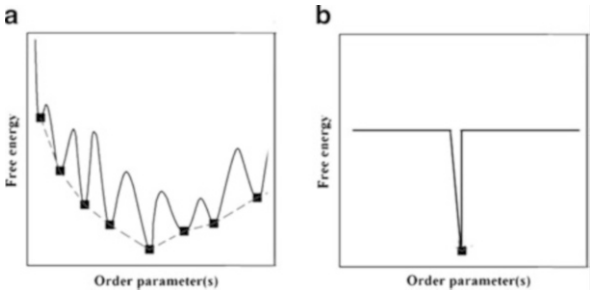


Fig. 1 Simplified illustration of energy landscape: (**a**) the general landscape; (**b**) golf-course like landscape. The landscape (**a**) could be transformed to a *bowl-shaped* one without noise by interpolating local minima points as shown by the *dashed line*, but (**b**) does not have such a helpful transformation

$$d^* = 3N + 3 - \kappa,$$

where κ is the number of correlated dimensions, which could vary greatly according to the intrinsic chemistry in the system. For example, the dimensionality drops a lot from 99 to 11.6 for $Mg_{16}O_{16}$, while only a little, from 39 to 32.5, for $Mg_4N_4H_4$. Thereby, the reduced complexity for the energy landscape of local minima is

$$C^* = \exp(\beta d^*).$$

This implies that any efficient search method must include structure relaxation (*local optimization*). We also note that all global optimization methods rely on the assumption that the reduced energy landscape should have an overall shape (Fig. 1a). An extreme (and, fortunately, unrealistic) case of a golf-course landscape (Fig. 1b) gives an opposite example, where total lack of structure of the landscape will lead any global optimization method to fail.

1.2 Global Optimization Methods

As the stable structure corresponds to the global minimum of the free energy surface, crystal structure prediction is mathematically a global optimization problem. Several global optimization algorithms have been devised and used with some success in CSP – for instance, simulated annealing [4, 5], metadynamics [6, 7], genetic algorithms [8], evolutionary algorithms [13], random sampling [9], basin hopping [10], minima hopping [11], and data mining [12].

One either has to start already in a good region of configuration space (so that no effort is wasted on sampling poor regions) or has to use a "self-improving" method that locates, step by step, the best structures. The first group of methods includes metadynamics, simulated annealing, basin hopping, and minima hopping approaches.

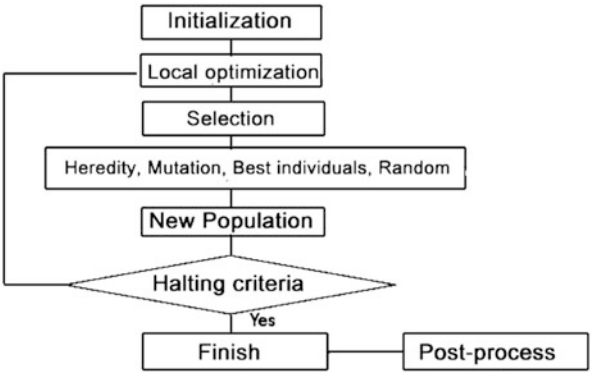


Fig. 2 The EA implemented in the USPEX code for crystal structure prediction. Several versions of this algorithm, as well as other algorithms, such as evolutionary metadynamics [7] and variable-cell NEB method [28], are implemented in USPEX as well

The second group essentially includes only evolutionary algorithms. Alternatively, data mining approaches use advanced machine learning concepts and predict the structures based on a large database of known crystal structures [12]. Among all these groups of methods, evolutionary algorithms present a particularly attractive approach for solving CSP. The strength of evolutionary simulations is that they do not require any system-specific knowledge except chemical composition, and are self-improving, i.e., in subsequent generations increasingly good structures are found and used to generate new structures. Its power has been evidenced by many recent discoveries in the field of CSP [20–27].

1.3 Evolutionary Algorithm

The evolutionary algorithm (EA) mimics Darwinian evolution and employs natural selection of the fittest and such variation operators as genetic heredity and mutations. It can perform well for different types of free energy landscapes. Unlike in genetic algorithms, we represent the coordinates of atoms in the unit cell and lattice vectors by real numbers (rather than binary "0/1" strings) – and therefore our algorithm is not genetic but evolutionary. The search space here is continuous and not discrete as with binary string representation.

The procedure is as shown in Fig. 2:

1. Initialization of the first generation, that is, a set of structures satisfying the hard constraints are randomly generated.
2. Determination of the quality for each member of the population using the so-called fitness function.
3. Selection of the best members from the current generation as parents, from which the new generation is created by applying specially designed variation operators.
4. Evaluation of the quality of all new trial solutions (i.e., structures).
5. Repeat steps 3 and 4 until pre-specified halting criteria are achieved.

The above algorithm has been implemented in the USPEX (Universal Structure Predictor: Evolutionary Xtallography) code [13–17]. Fitness function mathematically describes the target direction of the global search, which can be either a thermodynamic fitness (to find stable states) or a physical property (to find materials with desired properties).

1.4 Variation Operators

An essential step in an EA is to deliver the good gene to the next population. In USPEX, such delivery is done via variation operators. In general, the choice of variation operators follows naturally from the representation and the nature of the fitness landscape, and may or may not be inspired by physical processes representing transformations between likely good solutions.

Heredity is a core part of the EA approach, as it allows communication between different trial solutions or classes of solutions by combining parts from different parents. In USPEX, to generate a child from two parents, the algorithm first chooses a plane which is parallel to one lattice plane, and then cuts a slice with a random thickness and random position along the other lattice vector; such slices from two parent structures are then matched to form a child structure. In this process, the number of atoms of each type is adjusted to ensure conservation of chemical composition.

Mutation operators use a single parent to produce a child. *Lattice mutation* applies a stain matrix with zero-mean Gaussian random strains to the lattice vectors; *soft-mode mutation* (which we call *softmutation* for brevity) displaces atoms along the softest mode eigenvectors, or a random linear combination of softest eigenvectors; the *permutation* operator swaps chemical identities of atoms in randomly selected pairs of unlike atoms.

1.5 Fingerprints: A Tool to Identify Similar Crystal Structures and to Prevent Premature Convergence

A general challenge for global optimization methods is to avoid getting stuck in a local minimum and thus skip the global minimum. In the context of EA, this is due to the fact that good structures tend to produce children that bear resemblance to them, and it is possible for a good low-energy (but still not the global minimum) structure to come to dominate the population. Such behavior is especially common for energy landscapes with many good local minima, and a successful algorithm should address this problem. To prevent this, the key is to control the diversity of the population. Thus one question comes up – how can we detect similar structures and measure the similarity quantitatively?

Direct comparison of atomic coordinates will not work due to translational invariance (i.e., adding a constant vector to coordinates of all atoms will not change the structure) and because they are represented in lattice vectors units and there are many equivalent ways to choose a unit cell. Free energy difference is not a good parameter either: two completely different structures can have very close energies.

An ideal function characterizing a structure should be (1) derived from the structure itself, rather than its properties, (2) invariant with respect to shifts, rotations, and reflections in the coordinate system; (3) sensitive to different orderings of the atoms; (4) formally related to experiment; (5) robust against numerical errors, and (6) capable of incorporating short-range and long-range order. In USPEX, we use the so-called fingerprint function [29] to describe a crystal structure. It has the formulation very similar to pair distribution function (PDF), which for an elemental solid is

$$\mathrm{PDF}(R) = \sum_i \sum_{j \neq i} \frac{1}{4\pi R_{ij}^2 \frac{N}{V} \Delta} \delta\left(R - R_{ij}\right),$$

where R_{ij} is the distance between atoms i and j, V is the unit cell volume, N is the number of atoms in the unit cell, and Δ is a bin width (in Å). The index i goes over all atoms in the unit cell and index j goes over all atoms within the cutoff distance from the atom i. The PDF at long distances oscillates around the value +1, which is not convenient for our purposes, and we subtract this "background" value for convenience. Generalizing to systems containing more than one atomic type, we introduce fingerprint as a matrix, the components of which are fingerprint functions for A–B type distances:

$$F_{\mathrm{AB}}(R) = \sum_{\mathrm{A}_i, \mathrm{cell}} \sum_{\mathrm{B}_j} \frac{\delta\left(R - R_{ij}\right)}{4\pi R_{ij}^2 \frac{N_{\mathrm{A}} N_{\mathrm{B}}}{V} \Delta} - 1.$$

One can measure the similarity between two structures by calculating the cosine distance between two fingerprint functions:

$$d_{ij} = 0.5 \cdot \left(1 - \frac{f_i f_j}{|f_i||f_j|}\right).$$

Using this new crystallographic descriptor, we can improve the selection rules and variation operators above. During the selection process, only one copy of each distinct structure is used, and all its copies are killed. Fingerprint theory brings many other benefits (quantification and visualization of energy landscapes, use of ordered fragments of crystal structures, etc.); see [14, 17, 29].

2 New Developments

USPEX has been widely successfully in applications to very different kinds of systems, enjoying high success rate and efficiency. To predict very large and complex crystal structures, this method has been improved in many ways (generation of random symmetric structures, smart variation operators learning about preferable local environments and directed mutations, ageing technique, etc. [17]). Below we give examples of two major subjects, prediction of structures from molecular building blocks, and simultaneous optimization of both configurational and compositional space to find novel compounds.

2.1 Predicting Structures from Building Blocks

Molecular crystals are extremely interesting because of their applications as pharmaceuticals, pigments, explosives, and metal-organic frameworks [30]. The periodically conducted blind tests of organic crystal structure prediction, organized by Cambridge Crystallographic Data Centre (CCDC), have been the focal point for this community and they reflect steady progress in the field [31–36]. The tests show that it is now possible to predict the packing of a small number of rigid molecules, provided there are cheap force fields accurately describing the intermolecular interactions. In these cases, efficiency of search for the global minimum on the energy landscape is not crucial. However, if one has to use expensive ab initio total energy calculations or study systems with a large number of degrees of freedom (many molecules, especially if they have conformational flexibility, lead to astronomically large numbers of possible structures), efficient search techniques become critically important.

Compared to the prediction of atomic structures, there are several features to be taken into account for molecular crystals:

1. A typical unit cell contains many more atoms than a normal inorganic structure, which means an explosion of computing costs if all these atoms are treated independently.
2. Molecules interact with each other by weak forces, such as the van der Waals (vdW) interactions, and the inter-molecular distances are typically larger than those in atomic crystals, which leads to the availability of large empty space.
3. Most of the molecular compounds are thermodynamically less stable than simpler molecular compounds from which they can be obtained (such as H_2O, CO_2, CH_4, NH_3, H_2). This means that a fully unconstrained global optimization approach in many cases will produce a mixture of these simple molecules, which are of little interest to organic chemists. To study the packing of the actual molecules of interest it is necessary to fix the intra-molecular connectivity.
4. Crystal structures tend to be symmetric, and the distribution of structures over symmetry groups is very uneven [37, 38]. For example, 35% of inorganic and

45% of organic materials have the point group 2/*m*. Compared to inorganic crystals, there is a strong preference of organic crystals to a small number of space groups. Most organic crystals are found to possess space groups: $P2_1/c$ (36.59%), *P*-1 (16.92%), $P2_12_12_1$ (11.00%), *C2/c* (6.95%), $P2_1$ and *Pbca* (4.24%).

If we start to search for the global minimum with randomly generated structures, it is very likely that most of the time will be spent on exploring those uninteresting disordered structures far away from the target. Fortunately, the prediction of stable complex molecular structures can be achieved under the constraint of fixed molecules (or partially flexible molecules) as building blocks. The truly interesting problem for most organic chemists can be solved by *constrained global optimization*, finding the most stable packing of molecules with fixed bond connectivity. This will not only make the global optimization process meaningful, but at the same time will simplify it, leading to a drastic reduction of the number of degrees of freedom and of the search space. In order to apply constraints on the EA, we mainly need to modify the initialization of structures and variation operators.

2.1.1 Initialization: Generation of Molecular Structures

It is essential that all newly generated structures consist of molecules with desired bond connectivity. The efficiency can be greatly enhanced by using symmetry (so that different molecules in the unit cell are symmetrically related to each other) in the random generation of new structures – a population of symmetric structures is usually more diverse than a set of fully random (often disordered) structures. Diversity of the population of structures is essential for the success and efficiency of evolutionary simulations.

The initial structures are usually generated randomly, with randomly selected space groups. First, we randomly pick 1 of 230 space groups, and set up a Bravis cell according to the prespecified initial volume with random cell parameters consistent with the space group. Then one molecule is randomly placed on a general Wyckoff position and is multiplied by space group operations. If two or more symmetry-related molecules are found close to each other, we merge them in one molecule that sits on a special Wyckoff position and has averaged coordinates of the molecular center and averaged orientational vectors (or random, when the average value is zero). Adding new molecular sites one by one, until the correct number of molecules is reached, we get what we call a random symmetric structure (Fig. 3). During this process we also make sure that no molecules overlap or sit too close to each other.

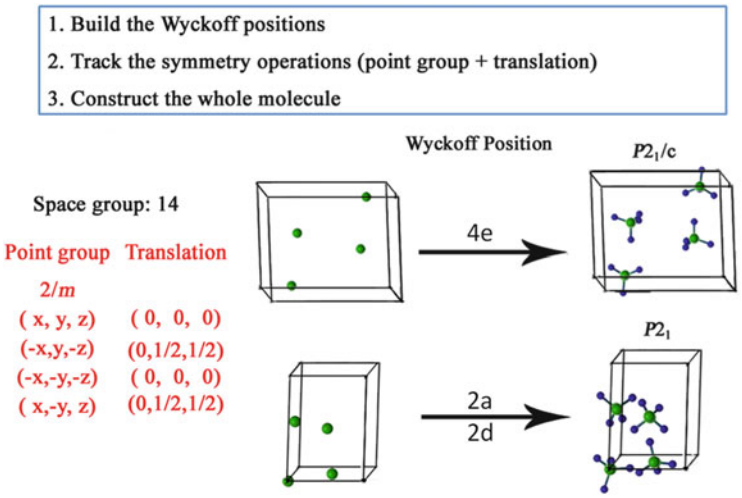


Fig. 3 Illustration of generating a random symmetric structure with four molecules per cell. For a given space group randomly assigned by the program (in this case, $P2_1/c$), the Bravis cell is generated, and molecular center is placed onto a random position (in this case, the general position 4*e* or 2a + 2d). Molecules are then built at the Wyckoff sites preserving their intramolecular connectivity and with their orientations obeying space group symmetry operations. Molecular geometry often breaks space group symmetry, leading to a subgroup, and we allow this. For clarity of the figure, molecules occupying positions at the corners and faces of the unit cell are shown only once

2.1.2 Variation Operators

Child structures (new generation) are produced from parent structures (old generation) using one of the following variation operators:

1. Heredity.
2. Permutation.
3. Coordinate mutation.
4. Lattice mutation (seldom used for molecular crystals).

   These are the same as in atomic crystal structures, with the only difference that variation operators act on the geometric centers of the molecules and their orientations, i.e., whole molecules, rather than single atoms, are considered as the minimum building blocks. Since molecules cannot be considered as spherically symmetric point particles, additional variation operators must be introduced.
5. Rotational mutation of the whole molecules.
6. Modified softmutation, which must retain molecular connectivity and is thus a hybrid operator of coordinate and rotational mutation. Figure 4 shows how variation operators work in our algorithm. Below we describe how these variation operators were used in our test cases.

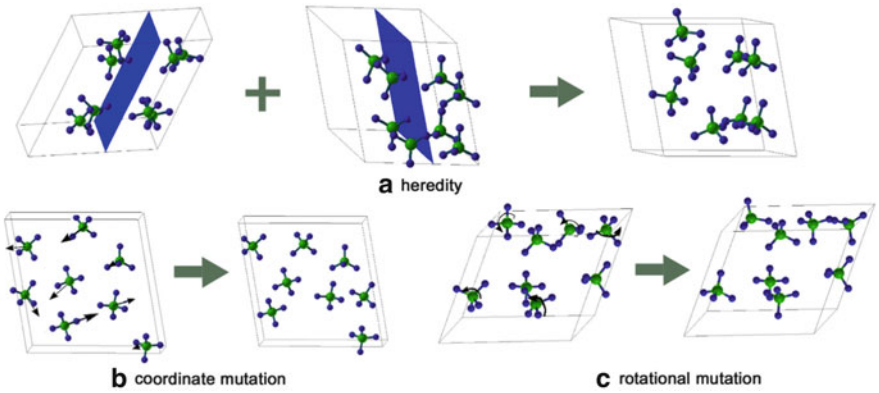


Fig. 4 Variation operators: (**a**) heredity; (**b**) coordinate mutation; (**c**) rotational mutation

Heredity

This operator cuts planar slices from each individual and combines these to produce a child structure. In heredity, each molecule is represented by its geometric center (Fig. 4a) and orientation. From each parent, we cut (parallel to a randomly selected coordinate plane of the unit cell) a slab of random thickness (within the bounds of 0.25–0.75 of the cut lattice vector) from a random height in the cell. If the total number of molecules of each type obtained from combining the slabs does not match the desired number of molecules, a corrector step is performed: molecules in excess are removed while molecules in shortage are added; molecules with a higher local degree of order have higher probability to be added and lower probability to be removed. This is equivalent to our original implementation of heredity for atomic crystals.

Rotational Mutation

A certain number of randomly selected molecules are rotated by random angles (Fig. 4c). For rigid molecules there are only three variables to define the orientation of the molecules. For flexible molecules, we also allow the mutation of torsional angles of the flexible groups. A large rotation can have a marked effect on global optimization, helping the system to jump out of the current local minimum and find optimal orientational ordering and optimal molecular conformation.

Softmutation

This powerful operator, first introduced for atomic crystals [14], involves atomic displacements along the softest mode eigenvectors, or a random linear combination of the softest eigenvectors. In the context of molecular crystals it becomes a hybrid

operator, combining rotational and coordinate mutations. In this case, the eigenvectors are calculated first and then projected onto translational and rotational degrees of freedom of each molecule and the resulting changes of molecular positions and orientations are applied, preserving rigidity of the fixed intra-molecular degrees of freedom. To calculate efficiently the normal modes, we construct the dynamical matrix from bond hardness coefficients [14]. The same structures can be softmutated many times, each time along the eigenvector of a new mode.

2.2 *Method for Variable-Composition Searches: Prediction of New Compounds*

This is a function to enable simultaneous prediction of all stable stoichiometries and structures. A pioneering study was done by Johanesson et al. [39], who succeeded in predicting stable stoichiometries of alloys within a given structure type. However, a simultaneous search for stable structures and compositions is much more challenging. This means that we are dealing with a complex landscape consisting of compositional and structural coordinates which require a series of modification of the standard EA approaches. This was done in 2008 in the USPEX code (see [40, 41]).

In order to involve the variation of chemical composition, we need to consider the following issues:

1. The sampling should cover the whole range of compositions of interest.
2. Proper fitness should be devised to evaluate the quality of structures that have different compositions.
3. Smart selection rules are needed, based on the fitness function.
4. Variation operators should allow the variation of stoichiometries.

2.2.1 Fitness: Representation as a Convex Hull

For a system with a given chemical formula, the optimizing target only involves energy per formula unit. If one wants to study a system of compounds with different stoichiometries, the stability can be evaluated by the formation energy towards the decomposition into mixtures of other compounds. Let us take a simple binary system AB as an example. The energy of formation of A_xB_{1-x} can be expressed as

$$E_{\text{formation}} = E_{\text{AB}} - xE_{\text{A}} - (1-x)E_{\text{B}},$$

where E_{A} and E_{B} correspond to the energy of the elemental A and B forms. Cleary $E_{\text{formation}}$ is a function of the compositional ratio x, and its calculation requires the knowledge of E_{A}, E_{B}, and E_{AB}. Stable compounds have negative energy of formation. If we draw the plot of $E_{\text{formation}}(x)$ for a series of structures/compositions in the A–B system as shown in Fig. 5, any structure with negative $E_{\text{formation}}$ can be stable

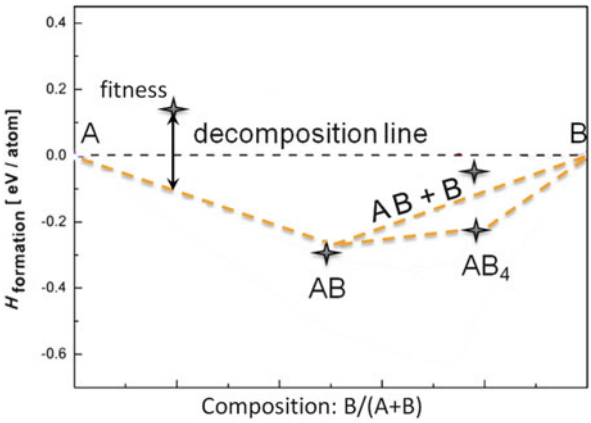


Fig. 5 Energy of formation as a function of composition. The stable structures need to be below all the possible "decomposition lines," and form a convex hull. The fitness can be defined as the minimum vertical distance from the convex hull

towards decomposition into the elements A and B – this is visually easy to detect, as structure AB, stable against decomposition into A and B, is below the line drawn from A to B. However, for a compound A_xB_{1-x} to be thermodynamically stable, this is necessary but not sufficient – a sufficient condition is that this compound is stable to decomposition into any other compounds (not only elements A and B), i.e., is below all the possible "decomposition lines." All thermodynamically stable compounds form a convex hull. The fitness of a structure/composition can be defined as the minimum vertical distance from the convex hull (see Fig. 5).

2.2.2 Selection

With the fitness available, we can proceed to the selection process. In a standard EA approach we select low-energy structures from the current generation. That is, the current population is considered as the *selection pool*. For variable-composition calculations, a modified selection rule can be beneficial: we are facing a much more complex search space, and the population size is usually insufficient to represent the diversity of the whole system. Thus we need to build the *selection pool* from the whole history. At the end of each generation, we update the convex hull and then calculate the fitness for the structures from all previous generations, and rank them after discarding identical structures identified by fingerprints. One common behavior of this data set is that the distribution of "high fitness structures" is very uneven in the compositional space. There might exist many low-energy structures for some particular compositions while only a few structures for other compositions. This indicates that the energy window varies a lot with stoichiometries, and thus a direct selection from the ranking list might bias the search considerably. To revise this, we use a simple rule to set the maximum number of structures for each composition when building the selection pool (Figs. 6 and 7).

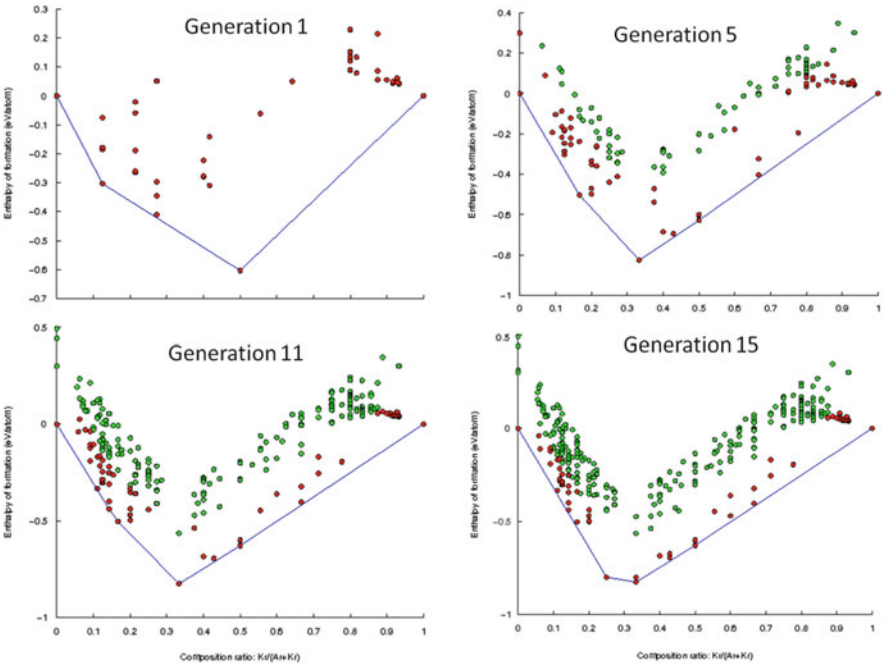


Fig. 6 The evolution of selection pool in USPEX for variable-composition structure prediction of a binary Lenard–Jones system (see Fig. 10 for details)

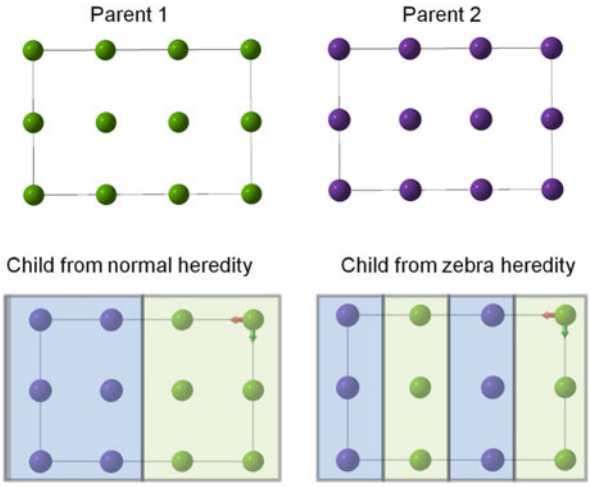


Fig. 7 Illustration of zebra heredity operator. It is quite obvious that in the case of variable composition the child structure obtained from many slices would be much more reasonable than the one obtained from traditional two-slice heredity

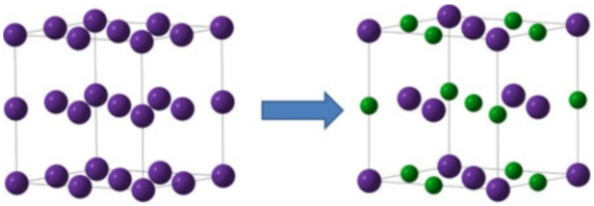


Fig. 8 Illustration of transmutation: one can obtain the NaCl-type structure from the simple cubic structure by transmuting half of the atoms

2.2.3 Variation Operators

Some of the variation operators, like softmuation and permutation, have the same formulation as used in standard EA. Heredity, however, is defined in a slightly different way. First, the chemistry-preserving constraints in the heredity operator should be removed. Second, if we consider two parent structures with quite different stoichiometries, their child structures obtained by normal heredity will very likely have two distinct chemical blocks as shown in Fig. 8, and such structures will be closer to the idea of a two-phase assemblage (a result of decomposition) than a single phase with a definite chemical composition. To remedy this we cut many slices from both parents (the thickness determined stochastically according to the approximate atomic radii) in a "zebra" pattern – the modified heredity operator is called "zebra heredity."

To allow further change of chemical composition, we introduce a "chemical transmutation" operator. This operator turns out to be quite efficient for driving the system from a known minimum to another good minimum in a different area of compositional space.

2.2.4 Implementation and Tests

After considering all the above ideas developments, the EA for variable-composition searches can be designed, as shown in Fig. 9.

An example of a (very difficult) system is given in Fig. 10. Consider a simple binary Lennard–Jones A–B system; the potential for each atomic ij-pair is given by

$$U_{ij} = \varepsilon_{ij}\left[\left(\frac{R_{\min}}{R}\right)^{12} - 2\left(\frac{R_{\min}}{R}\right)^{6}\right],$$

where $R_{\min}$ is the distance at which the potential reaches minimum, and ε is the depth of the minimum. In these simulations we use additive atomic dimensions $R_{\min}(\mathrm{BB})=1.5R_{\min}(\mathrm{AB}) = 2R_{\min}(\mathrm{AA})$ and non-additive energies (to favor compound formation) $\varepsilon_{\mathrm{AB}} = 1.25$; $\varepsilon_{\mathrm{AA}} = 1.25\varepsilon_{\mathrm{BB}}$. Odd as it may seem, a binary Lennard–Jones system with a 1:2 ratio of radii exhibits a large number of ground states – including the exotic $A_{14}B$ compound and the well-known AlB_2-type structure, and several marginally unstable compositions (such as A_8B_7, $A_{12}B_{11}$,

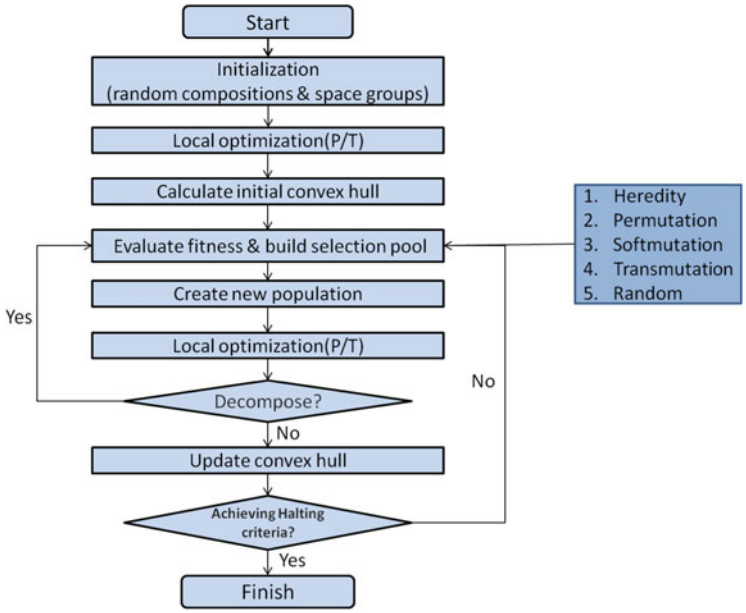


Fig. 9 The flowchart of variable-composition prediction in the USPEX code

A_6B_7, A_3B_4, AB_2). The correctness of these predictions is illustrated by the fact that a fixed-composition test simulation at AB_2 stoichiometry produced results perfectly consistent with the variable-composition runs.

Figure 11 shows a practically interesting example of variable-composition simulations – B–N system at ambient and high (50 GPa) pressure. At ambient pressure, hexagonal BN is thermodynamically stable, and $B_{13}N$ is right at the border between stability and metastability. On increasing pressure, $B_{13}N$ becomes metastable and only BN (in the cubic, diamond-like, form) is stable. One can also notice a strong increase of stability of BN – its enthalpy of formation increases from ~ -1.5 eV/atom at 1 atm to ~ -2.2 eV/atom at 50 GPa. Variable-composition calculations are a very powerful tool to explore chemical reactivity of the elements and its dependence on external conditions, such as pressure.

3 Applications

As an illustration of constrained global optimization for molecular crystals, we consider a promising material for hydrogen storage, $Mg(BH_4)_2$. To illustrate how pressure leads to the formation of new chemical compounds (which are most efficiently predicted by variable-composition searches), we show recent results on the Xe–O and Mg–O systems. In all the calculations, global optimizations were carried out by the USPEX code, and the VASP code [42] was employed for local

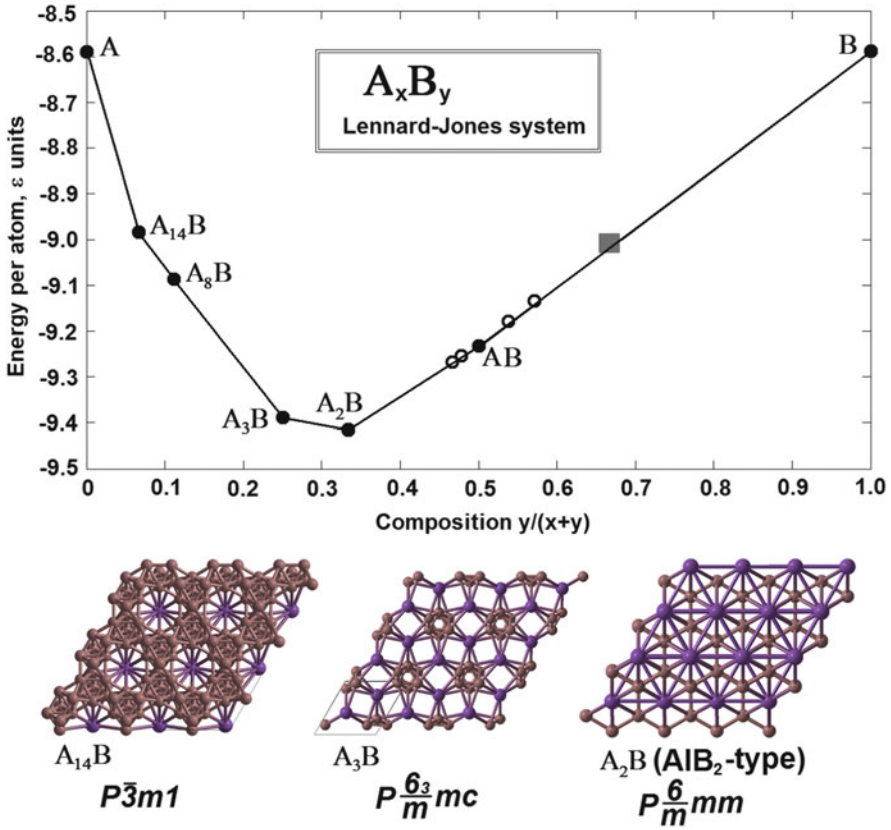


Fig. 10 Variable-composition USPEX simulation of the A_xB_y binary Lennard–Jones system. In the *upper panel*: stable compositions ($A_{14}B$, A_8B, A_3B, A_2B, AB). The *lower panel* shows some of the stable structures

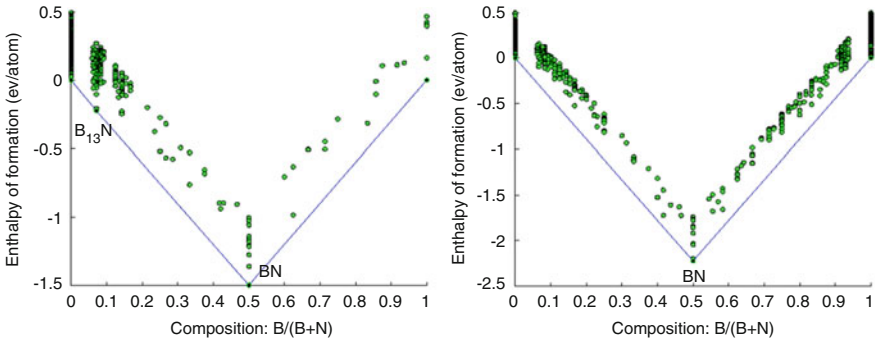


Fig. 11 Variable-composition USPEX simulation of B–N system at 1 atm (*left*) and 50 GPa (*right*)

optimization (i.e., structural relaxation), using the PBE exchange-correlation functional [43] and the PAW method [44].

3.1 $Mg(BH_4)_2$ [16, 27]

Lightweight metal borohydrides have recently received much attention owing to their high gravimetric and volumetric hydrogen densities compared to other complex hydrides [45]. Of these, magnesium borohydride, $Mg(BH_4)_2$, as a prominent lightweight solid-state hydrogen storage material with a theoretical hydrogen capacity of 14.8 wt%, has been extensively studied at both ambient and high pressure conditions.

3.1.1 $Mg(BH_4)_2$ at Ambient Condition

As a test, we first explore the energy landscape of $Mg(BH_4)_2$ at ambient condition. $Mg(BH_4)_2$ at ambient condition has been extensively studied as a template for developing novel hydrogen-storage solutions. Based on the experimental data, the ground-state α and β phases have been assigned space groups $P6_122$ (330 atoms per unit cell) and *Fddd* (704 atoms/cell), and turned out to have unexpectedly complex crystal structures [46–49]. There had been disputes between experimentalists and theoreticians regarding the nature of these ground-state structures [50–52]. Recent theoretical work then predicted a new body-centered tetragonal phase (with $I4m2$ symmetry), which has slightly lower energy than the $P6_122$ phase, by using the prototype electrostatic ground-state approach (PEGS) [50]. Later, based on the prototype structure of $Zr(BH_4)_4$, another orthorhombic phase with $F222$ symmetry was found to have even lower energy than all previously proposed structures [52].

In general, the previous theoretical discoveries of novel $Mg(BH_4)_2$ phases were conducted either by ad hoc extensive searching or by chemical intuition. However, USPEX does not rely on any prior knowledge except chemical composition, and could be particularly useful for predicting stable crystal structures for these complex metal hydride systems. If we consider the BH_4^- ion as a molecular group, the search space would be dramatically reduced. Within 10 generations (or just 400 structure relaxations), USPEX found the $F222$ phase (Fig. 12a) as the most stable structure at ambient pressure. Moreover, the $I\text{-}4\ m2$ structure (Fig. 12b) was also found by USPEX in the same calculation, with enthalpy less than 1.2 meV/atom above that of the $F222$ phase. Compared to the previous work, our method is clearly more universal, systematic, and robust, enables efficient structure prediction for complex molecular systems, both organic and inorganic.

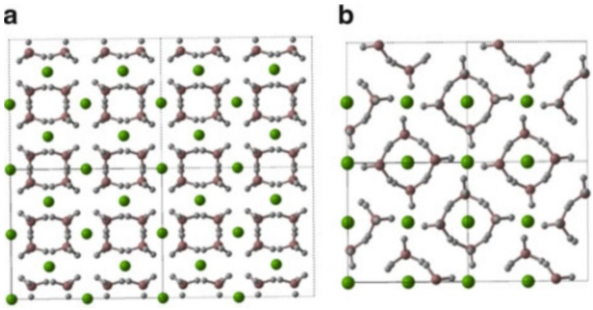


Fig. 12 $Mg(BH_4)_2$ polymorphs at ambient conditions found by USPEX. (**a**) *F*222 phase; (**b**) *I*-4*m*2

3.1.2 $Mg(BH_4)_2$ Under High Pressure

To improve the reversible hydrogen absorption or desorption kinetics or get new metastable polymorphs, recent studies focused on the stabilization of the high-pressure phases of $Mg(BH_4)_2$ at ambient pressure. Most recently, new δ, δ′, and ε phases of $Mg(BH_4)_2$ were successfully synthesized under pressure [53]. Many of them turned out to retain their structure upon decompression to ambient conditions. Crystal structures of γ and δ phases were, apparently convincingly, resolved using powder synchrotron X-ray diffraction [53]. Unexpectedly, theoretical phonon calculations showed the $P4_2nm$ structure (proposed by Filinchuk for the δ phase [53]) to be dynamically unstable at ambient pressure, which means that the exact crystal structure of the δ phase is still unresolved, even for such a simple structure with only 22 atoms per cell, and still less for the poorly characterized δ′ and ε phases. Therefore, the polymorphism and phase diagram of this important compound required further investigation.

According to our prediction, the tetragonal $I4_1/acd$ and trigonal P-3m1 phases are found to be the most stable ones in structure searches at 2–5 GPa and 10–20 GPa, respectively. Interestingly, within the whole pressure range (up to 20 GPa), we did not find the $P4_2nm$ structure proposed by Filinchuk et al. [53], but instead found the $I4_1/acd$ phase with 4 formula units (44 atoms) per cell and P-4 phase with 2 formula units per cell at pressures below 5 GPa (see Fig. 13). Given that the $P4_2nm$ structure is dynamically unstable at ambient pressure, and based on our enthalpy calculations, we hypothesized that the $I4_1/acd$ and P-4 structures might correspond to the experimentally observed δ and δ′ phases. Further investigation confirmed this suggestion, as we will show below.

$I4_1/acd$-$Mg(BH_4)_2$ becomes more stable than the γ phase at pressures above 0.7 GPa (Fig. 13). In the room-temperature experiment, a pressure-induced structural transformation is observed for the porous γ phase, and occurs in two steps: the γ phase turns into a diffraction-amorphous phase at 0.4–0.9 GPa and then, at approximately 2.1 GPa, into the δ phase [53]. We note a tiny enthalpy difference between $I4_1/acd$ and P-4 structures at pressures around 1 GPa. As pressure increases to 9.8 GPa, the P-3 m1 structure becomes the most stable structure, in agreement with earlier predictions [46, 54]. Bil et al. [55] indicated that it is important to treat long-range dispersion interactions to get the ground state

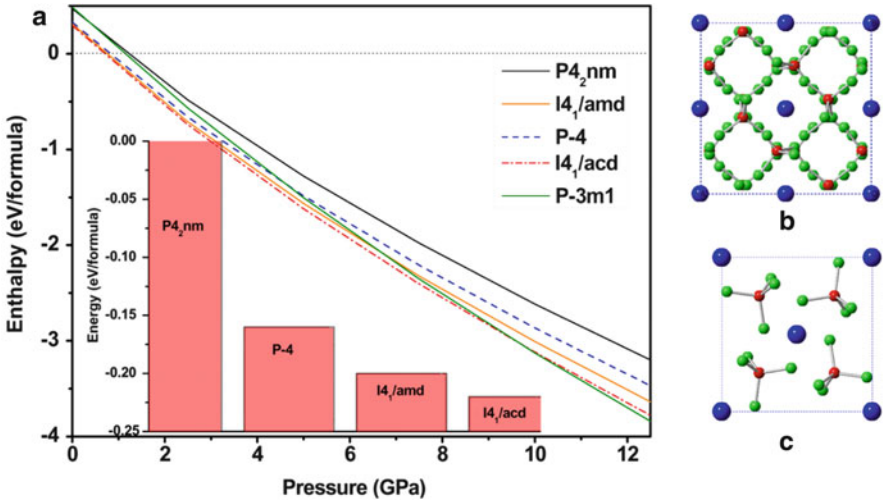


Fig. 13 (**a**) Enthalpy curves (relative to the γ phase) of various structures of $Mg(BH_4)_2$ as a function of pressure; (**b**) the $I4_1/acd$ structure; (**c**) the P-4 structure. Enthalpies are given per formula unit. The *inset* in (**a**) shows the energy per formula unit of $I4_1/acd$, P-4, and $P4_2nm$ structures (relative to the $P4_2nm$ structure) at zero pressure, including vdW interactions

structures of magnesium borohydrides correctly. We have examined the energetic stability of the considered structures through a semi-empirical Grimme correction to DFT energies, stresses and forces [56]. When this correction is included, the $I4_1/acd$ and P-4 structures once again come out as more stable than the $P4_2nm$ structure, by 21.2 kJ/mol and 15.4 kJ/mol, respectively. Energetic stability seems to correlate with the degree of disparity of bond lengths and atomic Bader charges. The $P4_2nm$ structure has two inequivalent Mg–H distances, 2.26 and 2.07 Å, compared to 2.11 and 2.07 Å in the $I4_1/acd$ structure, and 2.12 and 2.06 Å in the P-4 structure. As we can see, the more homogeneous bond lengths, the greater stability. Bader charges, computed using the code [57], show the same picture: for H atoms we find them to be $-0.63e$ and $-0.59e$ in the $P4_2nm$ structure, $-0.63e$ and $0.62e$ in the P-4 structure, and $-0.63e$ and $-0.61e$ in the $I4_1/acd$ structure. More homogeneous Bader charges and bond lengths in the $I4_1/acd$ and P-4 structures correlate with their greater thermodynamic stability at ambient pressure, in agreement with proposed correlations between local bonding configurations and energetic stability [52].

Our calculations suggest that the $P4_2nm$ structure, proposed by experiment for the δ phase, is unstable. This implies that either density functional theory calculations are inaccurate for this system, or experimental structure determination was incorrect. To assess these possibilities, we simulated the XRD patterns of the $I4_1/acd$ and P-4 structures, and compared them with the experimental XRD pattern of the δ phase at ambient pressure (see Fig. 14a). One observes excellent agreement, both for the positions and the intensities of the peaks (including both strong and weak peaks), of

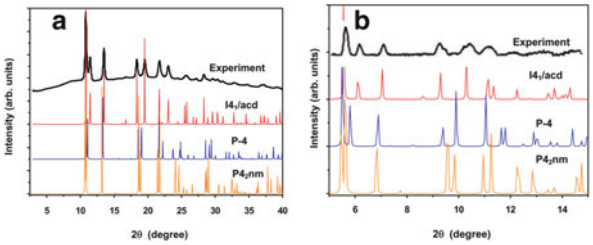


Fig. 14 Simulated XRD patterns of the $I4_1/acd$, $P\text{-}4$ and $P4_2nm$ structures of $Mg(BH_4)_2$ with the X-ray wavelength of 0.770518 Å at ambient pressure (**a**) and 0.36814 Å at 10 GPa (**b**) in comparison with the corresponding experimental results [46, 53]

the $I4_1/acd$ structure with experiment [53]. The situation is very peculiar: two structures, $I4_1/acd$ and $P4_2nm$, have nearly identical XRD patterns, both compatible with experiment – but one, $I4_1/acd$, is the true thermodynamic ground state (global minimum of the enthalpy), whereas the other, $P4_2nm$, is not even a local minimum of the enthalpy (dynamically unstable structure, incapable of sustaining its own phonons). In this situation, the true structure is clearly $I4_1/acd$. This case gives a clear real-life example of the fact that very different structures can have very similar powder XRD patterns, making structure determination from powder data dangerous, and in such cases input from theory is invaluable. The $P\text{-}4$ structure also has a rather similar XRD pattern, but the peak positions are slightly shifted. Comparison with an independent experimental XRD pattern collected at 10 GPa (Fig. 14b) shows that the peak positions and intensities of the $I4_1/acd$ structure are once again in excellent agreement with the experimental data [46], while the strong peaks of the $P\text{-}4$ structure at 9.9°, 11.6°, and 11.8° obviously deviate from the observed ones. This reinforces our conclusion that the $I4_1/acd$ structure is the best candidate for the high pressure δ phase. At pressures below 10 GPa, a mixture of $I4_1/acd$ and $P\text{-}4$ phases is possible, as the XRD peaks of these two structures are quite similar. We must remember that in the experiment, the δ and δ′ phases are nearly indistinguishable [53]. This example highlights the importance of theoretical simulations in establishing crystal structures, when only powder XRD data are available: purely experimental solutions may be dangerous even for simple structures, such as the structure of the δ phase with only six non-hydrogen atoms in the unit cell.

3.2 Xe–O system [25]

Xenon is a noble gas, chemically inert at ambient conditions. A few xenon fluorides have been found [58–61], with Xe atoms in the oxidation states +2, +4, or +6. Upon application of high pressure, insulating molecular structure of XeF_2 was found to transform into two- and three-dimensional extended solids and to become metallic [61]. Clathrate Xe–H solids were also observed [62]. Two xenon oxides (XeO_3,

XeO_4) [63] are known at atmospheric pressure, but are unstable and decompose explosively above 25°C (XeO_3) and −40°C (XeO_4) [64]. A crystalline XeO_2 phase with local square-planar XeO_4 geometry has recently been synthesized at ambient conditions [65].

Growing evidence shows that noble gases, especially Xe, may become much more reactive under pressure [66]. The formation of stable xenon oxides and silicates could explain the missing xenon paradox, i.e., the observation that the amount of Xe in the Earth's atmosphere is an order of magnitude less than what it would be if all Xe were degassed from the mantle into the atmosphere [67]. One possibility to explain this deficiency is to assume that Xe is largely retained in the Earth's mantle. In fact, a recent experiment discovered that xenon reacts with SiO_2 at high pressures and temperatures [68, 69]. At the same time, recent theoretical investigation showed that no xenon carbides are stable, at least up to the pressure of 200 GPa [70], and experimental and theoretical high pressure work [71] found no tendency for xenon to form alloys with iron or platinum.

Here we address possible stability of xenon oxides using quantum-mechanical calculations of their energetics. We have performed structure prediction simulations for the Xe–O system for the compositions of XeO, XeO_2, XeO_3, XeO_4 at 5, 50, 100, 120, 150, 180, 200, and 220 GPa. Our calculation at 5 GPa yielded lowest-enthalpy structures that always contained the O_2 molecules, indicating the tendency for segregation of the elements, and indeed at 5 GPa decomposition was found to be energetically favorable. This suggests that the reaction observed by Sanloup et al. [68, 69] at 0.7–10 GPa was an entropically driven incorporation of Xe impurities into the structure of SiO_2, rather than enthalpically-driven formation of a stoichiometric xenon silicate or oxide. Indeed, solid solutions and point defects are stabilized by entropy (rather than enthalpy) [72].

3.2.1 Stable Xe–O Compounds Under High Pressure [25]

Figure 15 shows the enthalpy of formation of all the Xe oxides as a function of pressure. Below 83 GPa all xenon oxides are unstable. At 83 GPa, XeO-*Pbcm* becomes stable, followed by XeO_2-$P2_1/c$ above 102 GPa and XeO_3-$P4_2/mnm$ above 114 GPa. There is a clear trend of increasing the oxidation number of Xe on increasing pressure.

A simple and clear analysis of chemical bonding can be carried out using the electron localization function (ELF) [73]. The ELF gives information about the valence electron configuration of an atom in a compound. States with closed-shell electronic configurations (Xe^0, $5s^25p6$, and Xe^{6+}, $5s^2$) will exhibit a spherical ELF distribution, whereas open-shell states (Xe^{2+}, Xe^{4+}) will not. For Xe^{2+} one *p*-orbital is empty and the ELF will have a toroidal shape; likewise, Xe^{4+} can be formed by the removal of two *p*-orbitals and the ELF will show a two-lobe maximum corresponding to the shape of the lone *p*-electron pair.

The most stable structure of XeO at 100 GPa has space group *Pbcm* and eight atoms in the unit cell. As shown in Fig. 15c, Xe atoms are in a twofold (linear) coordination and Xe–O bonds form chains, with O–Xe–O angles of 175.6° and

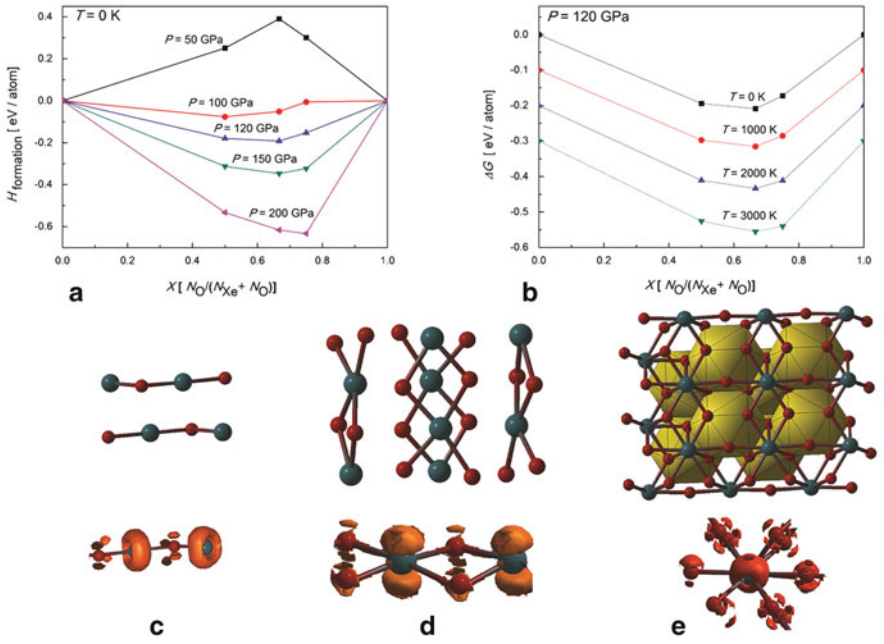


Fig. 15 (**a**) Predicted enthalpies of formation of Xe–O compounds at high P and $T = 0$ K; (**b**) predicted Gibbs free energy of formation of Xe–O compounds at different temperatures (shifted for clarity by −0.1 eV/atom at each successive temperature) and $P = 120$ GPa; (**c**) crystal structure of XeO (*Pbcm*) at 100 GPa, and its ELF isosurface (ELF = 0.85) on the Xe–O chain; (**d**) crystal structure of XeO_2 ($P2_1/c$) at 120 GPa, and its ELF isosurface (ELF = 0.85) on the XeO_4 square; (**e**) crystal structure of XeO_3 (*Pmmn*) at 200 GPa, and its ELF isosurface (ELF = 0.82) on XeO_{12} anticuboctahedra

Xe–O–Xe angles of 112.6°. The alternating Xe–O bond lengths are 2.0 and 2.1 Å. The ELF picture shows a toroidal maximum of ELF around each Xe atom, exactly what one should expect for Xe^{2+} state.

For XeO_2, the stable structure above 102 GPa has space group $P2_1/c$ and 24 atoms in the unit cell. Xenon atoms have a slightly non-planar square coordination and the structure consists of 1D-ribbons of edge-sharing XeO_4-squares (Xe–O distances are 2.0 and 2.1 Å), with four Xe–O bonds and two lone pair maxima forming an octahedron, consistent with the geometry proposed by recent experiments [65]. Just as in XeO, there are no peaks visible in the ELF isosurface along the Xe–O bonds (Fig. 15d). Above 198 GPa it transforms into the XeO_2-*Cmcm* structure.

XeO_3 becomes stable at 114 GPa. Its structure has space group $P4_2/mnm$ and 16 atoms in the unit cell. It is stable against decomposition into Xe and O_2 as well as into XeO or XeO_2 and O_2. $P4_2/mnm$ phase is composed of two sublattices: square XeO_2 chains, again suggesting the Xe^{4+} state, and linear chains made of O_2 dumbbells. Above 145 GPa, the molecules in the linear $–O_2–O_2–$ chains are partly dissociated and we observe the $–O_2–O–$ chains in the $C2/c$ phase that has 48 atoms

Table 1 Representative chemical reactions involving xenon oxides and silicates in Earth's lower mantle at 100 GPa

Reaction	ΔH (eV)	ΔV (Å^3)
$FeO + Xe \rightarrow XeO + Fe$	2.170	−1.35
$FeO + Xe + ½O_2 \rightarrow XeO_2 + Fe$	2.203	−3.09
$FeO + 2Xe + SiO_2 \rightarrow FeSi + 2XeO$	8.540	−1.88
$Fe + Xe + SiO_2 \rightarrow FeSi + XeO_2$	8.687	−2.03
$4XeO + 2SiO_2 \rightarrow 2XeSiO_4$	0.910	−0.61
$4XeO + 2MgSiO_3 \rightarrow 2XeSiO_4 + 2MgO$	1.490	0.27
$4XeO + 2CaSiO_3 \rightarrow 2XeSiO_4 + 2CaO$	4.409	−0.33
$Xe + SiO_2 \rightarrow Si + XeO_2$	12.205	0.33
$2Xe + SiO_2 \rightarrow Si + 2XeO$	12.057	0.48

per unit cell. Above 198 GPa, the structure transforms to a *Pmmn* phase with eight atoms per unit cell. In this remarkable structure, the oxygen atoms form anticuboctahedra in which the Xe atoms sit in the center (Fig. 15e). The ELF distribution around Xe atoms in the *Pmmn* phase is spherical around the xenon, which points at the Xe^{6+} valence state with a spherically symmetric 5 s^2 valence shell. Again, we observe the tendency of increasing oxidation states under pressure.

Xenon fluorides are stable at ambient conditions, whereas xenon oxides become stable above 83 GPa. Xenon carbides are unstable up to 200 GPa at least [69]. It appears that xenon forms compounds most readily with the most electronegative atoms, and that in turn suggests that ionicity is essential. This is somewhat counter-intuitive, given that the xenon atom has a very stable closed valence shell and its electronegativity is rather high. The electronegativity difference (1.4 for Xe–F, 0.8 for Xe–O, and 0.56 for Xe–C) determines the degree of ionicity at ambient conditions. However, ionicity often seems to be enhanced under pressure. Spontaneous ionization under pressure was recently found even in a pure element, boron [20].

3.2.2 Xe–Si–O System in the Earth's Mantle [25]

Table 1 shows the representative chemical reactions involving xenon oxides and silicates in the Earth's lower mantle at 100 GPa. Xe oxides are only stable above 83 GPa, i.e., at pressures corresponding to the lower mantle. Since in the Earth's mantle metallic Fe should be present [74, 75], stability of Xe oxides needs to be explored in the presence of metallic Fe. In our calculations of phase equilibria, we took into account that at lower mantle conditions Fe has the hcp structure and FeO has the antiferromagnetic inverse NiAs structure [76, 77]. Calculations show that all the predicted xenon oxides are very strong oxidants and will oxidize Fe, producing iron oxide and free xenon (FeO + Xe). Therefore, Xe oxides cannot be present in the lower mantle, where free Fe should exist.

Since xenon oxides are not stable in coexistence with metallic Fe, we investigated the formation of stable xenon silicates under pressure, focusing on $XeSiO_3$ and Xe_2SiO_4, which contain the least oxidized divalent xenon. All of the investigated compositions were unstable towards decomposition into XeO, XeO_2, SiO_2,

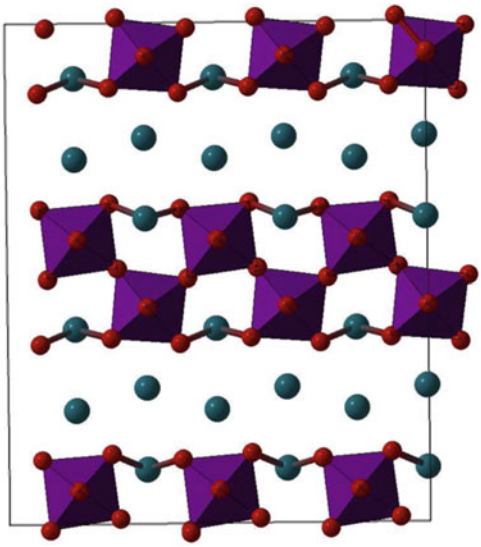


Fig. 16 Crystal structure of the least unstable Xe_2SiO_4 obtained from USPEX

and elemental Xe; Xe_2SiO_4 (Fig. 16) proved to be one of the least unstable silicates, but is still unstable. In this structure, Xe atoms terminate the silicate perovskite layers, suggesting that xenon could also be stored in perovskite/post-perovskite stacking faults [78] or at grain boundaries or dislocations.

3.3 Mg–O system [26]

Magnesium oxide (MgO) is one of the most abundant phases in planetary mantles, and understanding its high-pressure behavior is essential for constructing models of the Earth's and planetary interiors. For a long time, MgO was believed to be among the least polymorphic solids – only the NaCl-type structure has been observed in experiments at pressures up to 227 GPa [79]. Static theoretical calculations have proposed that the NaCl-type (B1) MgO would transform into CsCl-type (B2) and the transition pressure is approximately 490 GPa at 0 K (474 GPa with the inclusion of zero-point vibrations) [80–82]. Calculations also predicted that MgO remains non-metallic up to extremely high pressure (20.7 TPa) [81], making it to our knowledge the most difficult mineral to metalize. Thermodynamic equilibria in the Mg–O system at 0.1 MPa have been summarized in previous studies [83–85], concluding that only MgO is a stable composition, though metastable compounds (MgO_2, MgO_4) can be prepared at very high oxygen fugacities.

Using ab initio variable-composition evolutionary simulations, we explored the entire range of possible stoichiometries for the Mg–O system at pressures up to 850 GPa. In addition to MgO, our calculations find that two extraordinary compounds (MgO_2 and Mg_3O_2) become thermodynamically stable in the regions of

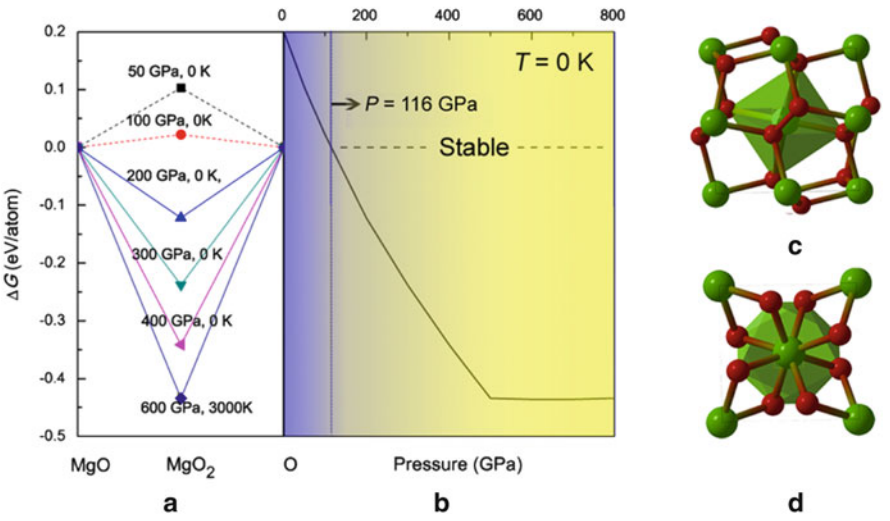


Fig. 17 (**a**) Convex hull for the MgO–O system at high pressures; (**b**) the enthalpy of formation of MgO_2 as a function of pressure; (**c**) *Pa*3 structure (*c*-MgO_2); (**d**) *I*4/*mcm* structure (*t*-MgO_2)

high and low oxygen chemical potential at 116 GPa and 500 GPa, respectively. To confirm this and to obtain the most detailed picture, we then focused our search on two separate regions of chemical space: Mg–MgO and MgO–O, respectively. Since the structures in the two regions exhibit different properties, we describe them separately.

3.3.1 MgO_2

It is well known that monovalent (H–Cs) and divalent (Be–Ba and Zn–Hg) elements are able to form not only normal oxides but also peroxides and even superoxides [86] (for instance, BaO_2 has been well studied at both ambient and high pressure [87, 88]). Our structure prediction calculations identified the existence of magnesium peroxide with *Pa*3 symmetry and 12 atoms in the unit cell at ambient pressure, which is in good agreement with experimental results [89]. In this cubic phase, Mg is octahedrally coordinated by oxygen atoms (which form O_2 dumbbells); see Fig. 17c. However, *Pa*3 MgO_2 (*c*-MgO_2 from now on) is calculated to have a positive enthalpy of formation from MgO and O_2, and is therefore metastable. The calculation shows that, on increasing pressure, *c*-MgO_2 transforms into a tetragonal form with space group *I*4/*mcm*. In the *t*-MgO_2 phase (Fig. 17d), Mg is eight-coordinate. Here we see the same trend of change from six- to eightfold coordination as in the predicted B1–B2 transition in MgO. However, in MgO_2 it happens at a mere 53 GPa, compared to 490 GPa for MgO. Most remarkably, above 116 GPa the *t*-MgO_2 structure has a negative enthalpy of formation from MgO and

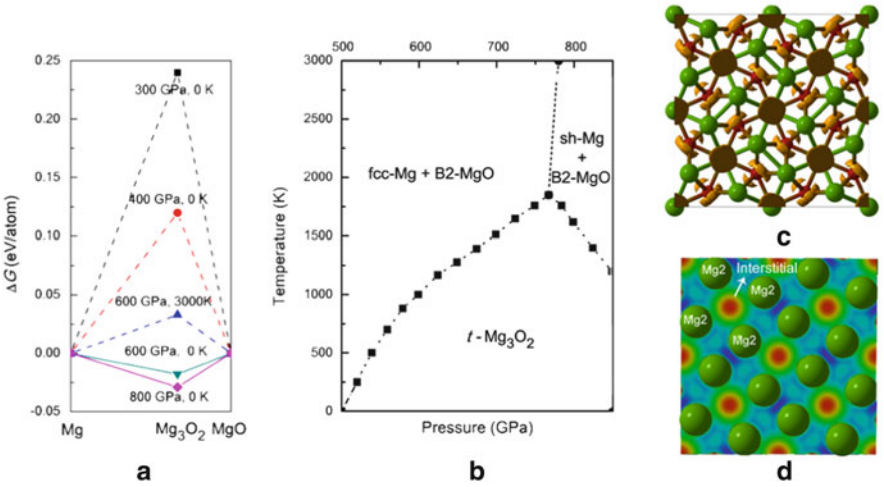


Fig. 18 (**a**) Convex hull for the Mg–MgO system at high pressures; (**b**) the corresponding P–T stability diagram of Mg_3O_2; (**c**) ELF isosurfaces of *t*-Mg_3O_2 (ELF = 0.83); (**d**) charge density distribution of *t*-Mg_3O_2 viewed along the *c*-axis showing interstitial charge density maxima

O_2, indicating that *t*-MgO_2 becomes thermodynamically stable. Furthermore, its stability is greatly enhanced by pressure and its enthalpy of formation becomes impressively negative, −0.43 eV/atom, at 500 GPa!

We also examined the effect of temperature on its stability by performing quasiharmonic free energy calculations using the PHONOPY code [90]. Thermal effects tend to decrease the relative stability of MgO_2 by 0.008 meV/(atom*K), which is clearly insufficient to change the sign of the formation free energy (G), and MgO_2 remains stable even at extremely high temperatures.

3.3.2 Mg_3O_2

For the Mg-rich part of the Mg-O phase diagram (Fig. 18), USPEX shows completely unexpected results. First of all, elemental Mg is predicted to undergo several phase transitions induced by pressure: hcp–bcc–fcc–sh. At ambient conditions, Mg adopts the hcp structure, while bcc-Mg is stable from 50 to 456 GPa, followed by the transition to fcc and simple hexagonal phase at 456 GPa and 756 GPa, respectively. These results are in excellent agreement with previous studies [91–93]. Unexpectedly, Mg-rich oxides, such as Mg_2O and Mg_3O_2, begin to show very competitive enthalpy of formation at pressures above 100 GPa. However, they are still not stable against decomposition into Mg and MgO, and their crystal structures could be thought of as a combination of blocks of Mg and B1–MgO. This situation qualitatively changes at 500 GPa, where we find that Mg_3O_2 becomes thermodynamically stable. This new stable (*t*-Mg_3O_2) phase has

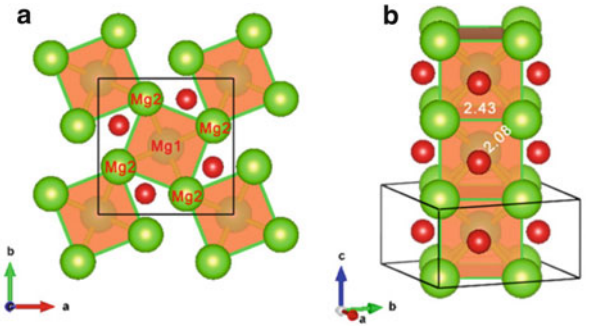


Fig. 19 (**a**) Crystal structures of t-Mg_3O_2 at 500 GPa, space group $P4/mbm$, $a = 4.508$ Å, $c = 2.367$ Å, Mg1(0.3494, 0.1506, 0.5); Mg2(0, 0, 0); O(0.8468, 0.6532, 0); (**b**) 1D-column of body-centered Mg-cubes

a very unusual tetragonal structure with the space group $P4/mbm$. This crystal structure can be viewed as a packing of O atoms and 1D-columns of almost perfect body-centered Mg-cubes. As shown in Fig. 19, there are two types of Mg atoms in the unit cell, Mg1 and Mg2. Here, Mg2 atoms form the cubes, joined into vertical columns and filled by Mg1 atoms.

Within the cubic columns, one can notice empty $(Mg1)_2(Mg2)_4$ clusters with the shape of flattened octahedra, with Mg–Mg distances ranging from 2.08 Å (Mg1–Mg2) to 2.43 Å (Mg2–Mg2). The coordination environments are quite different: each Mg1 is bonded to two Mg1 atoms and eight Mg2 atoms, while each Mg2 atoms is bonded to six O atoms (trigonal prismatic coordination) and two O atoms. Oxygen atoms in t-Mg_3O_2 are coordinated by eight Mg_2 atoms.

The ELF distribution in t-Mg_3O_2 (Fig. 18c) also shows strong charge transfer from Mg to O. However, we surprisingly found a very strong interstitial ELF maximum (ELF = 0.97) located in the center of the Mg-octahedron (Fig. 18d). To obtain more insight we performed Bader analysis. The resulting charges are $+1.592e$ for Mg1, $+1.687e$ for Mg2, $-1.817e$ for O, and $-1.311e$ for the interstitial electron density maximum. Such a strong interstitial electronic accumulation requires an explanation. At high pressure, strong interstitial electron localization was found in some alkali and alkaline-earth elements; for instance, sodium becomes a transparent insulator due to strong core–core orbital overlap [21]. As a measure of size of the core region we use the Mg^{2+} ionic radius (0.72 $Å^3$ [94]), while the size of the valence electronic cloud is represented by the $3s$ orbital radius (1.28 Å [95]). In Mg_3O_2, Mg–Mg contacts at 500 GPa (2.08 Å for Mg1–Mg2, 2.37 Å for Mg1–Mg1, and 2.43 Å for Mg2–Mg2) are only slightly shorter than the sum of valence orbital radii, but longer than the distance at which strong core-valence overlap occurs between neighboring Mg atoms (0.72 + 1.28 = 2.00 Å). Thus, the main reason for strong interstitial electronic localization is the formation of strong multicenter covalent bonds between Mg atoms; the core-valence expulsion (which begins at distances slightly longer than the sum of core and valence radii and increases as the distance decreases) could also play some role for valence electron localization.

Strong Mg–Mg covalent bonding is not normally observed; the valence shell of the Mg atom only has a filled $3s^2$ configuration, unsuitable for strong bonding.

Under pressure, the electronic structure of the Mg atom changes (*p*- and *d*-levels become significantly populated), and strong covalent bonding can appear as a result of *p*–*d* hybridization. There is another way to describe chemical bonding in this unusual compound. We must remember that Mg_3O_2 is anion-deficient compared with MgO; the extra localized electrons in Mg octahedron interstitial play the role of anions, screening Mg atoms from each other. These two descriptions are complementary.

3.3.3 Geophysical Implications

What are the implications of these two Mg–O compounds for planetary sciences? High pressures, required for their stability, are within the range corresponding to deep planetary interiors. In the interiors of terrestrial planets, reducing conditions dominate, due to the excess of metallic iron. This makes the presence of MgO_2 unlikely. However, given the diversity of planetary bodies it is not impossible to imagine that on some planets strongly oxidized environments can be present at depths corresponding to the pressure of 116 GPa and greater (in the Earth this corresponds to depths below ~2,600 km), which would favor the existence of MgO_2. At the more usual reducing conditions of planetary interiors, Mg_3O_2 could exist at pressures above 500 GPa in deep interiors of giant planets. There it can coexist in equilibrium with Fe (but probably not with FeO, according to our DFT and DFT + U calculations of the reaction of Fe + 3MgO = FeO + Mg_3O_2). According to our calculations (Fig. 17), Mg_3O_2 can only be stable at temperatures below 1,800 K, which is too cold for deep interiors of giant planets; however, impurities and entropy effects stemming from defects and disorder might extend its stability field into planetary temperatures. Exotic compounds MgO_2 and Mg_3O_2, in addition to their general chemical interest, might be important planet-forming minerals in deep interiors of some planets.

4 Outlook

Evolutionary algorithms, based on physically motivated forms of variation operators and local optimization, are a powerful tool enabling reliable and efficient prediction of stable crystal structures. This method has a wide field of applications in computational materials design (where experiments are time-consuming and expensive) and in studies of matter at extreme conditions (where experiments are very difficult or sometimes beyond the limits of feasibility).

One of the current limitations is the accuracy of today's ab initio simulations; this is particularly critical for strongly correlated and for systems where van der Waals interactions are essential [96] – although for the case of van der Waals bonding good progress has been achieved recently [97, 98]. Note, however, that our method itself does not make any assumptions about the way energies are calculated

and can be used in conjunction with any method that is able to provide total energies. Most practical calculations are done at $T = 0$ K, but temperature can be included as long as the free energy can be calculated efficiently. Difficult cases are aperiodic and disordered systems (for which only the lowest-energy periodic approximants and ordered structures can be predicted at this moment).

We are suggesting USPEX as the method of choice for crystal structure prediction of systems with up to ~100 atoms/cell, where no information (or just the lattice parameters) is available. Above ~100 atoms/cell runs become expensive due to the "curse of dimensionality" (although still feasible), eventually necessitating the use of other ideas within USPEX or another approach. There is, however, hope of enabling structure prediction for very large (>200 atoms/cell) systems. USPEX has been applied to many important problems. Here we highlighted the methodology and some applications in (1) prediction of molecular crystal structures and (2) variable-composition structure predictions. Due to lack of space, we did not describe here the following important advances:

- Methods to predict structures of nanoparticles [17] and surfaces [99], including variable-cell and variable-composition surface reconstructions.
- Hybrid optimization approach to optimize physical properties [23, 24] – this technique can be used for practically any physical property, and its variable-composition extension is available in USPEX.
- Evolutionary metadynamics [7], a powerful hybrid of the evolutionary algorithm USPEX and metadynamics.

One can expect many more applications to follow, both in high-pressure research and in materials design.

Acknowledgments Calculations were performed at the supercomputer of the Center for Functional Nanomaterials, Brookhaven National Laboratory. We gratefully acknowledge funding from DARPA (Grants No. W31P4Q1210008 and No. W31P4Q1310005), NSF (No. EAR-1114313 and No. DMR-1231586), the AFOSR (No. FA9550-13-C-0037), CRDF Global (No. UKE2-7034-KV-11), and Government of the Russian Federation (No. 14.A12.31.0003). X.F.Z thanks National Science Foundation of China (Grant No. 11174152).

References

1. Maddox J (1988) Crystals from first principles. Nature 335:201
2. Gavezzotti A (1994) Are crystal structures predictable? Acc Chem Res 27:309–314
3. Oganov AR (ed) (2010) Modern methods of crystal structure prediction. Wiley, Weinheim
4. Pannetier J, Bassas-Alsina J, Rodriguez-Carvajal J et al (1990) Prediction of crystal structures from crystal chemistry rules by simulated annealing. Nature 346:343–345
5. Schon JC, Jansen M (1996) First step towards planning of syntheses in solid-state chemistry: determination of promising structure candidates by global optimization. Angew Chem Int Ed Engl 35:1286–1304
6. Martonak R, Laio A, Parrinello M (2003) Predicting crystal structures: the Parrinello–Rahman method revisited. Phys Rev Lett 90:075503

7. Zhu Q, Oganov AR, Lyakhov AO (2012) Evolutionary metadynamics: a novel method to predict crystal structures. CrystEngComm 14:3596–3601
8. Woodley MS, Battle DP, Gale DJ et al (1999) The prediction of inorganic crystal structures using a genetic algorithm and energy minimisation. Phys Chem Chem Phys 1:2535–2542
9. Freeman CM, Newsam JM, Levine SM et al (1993) Inorganic crystal structure prediction using simplified potentials and experimental unit cells: application to the polymorphs of titanium dioxide. J Mater Chem 3:531–535
10. Wales DJ, Doye JPK (1997) Global optimization by basin-hopping and the lowest energy structures of Lennard–Jones clusters containing up to 110 atoms. J Phys Chem A 101:5111–5116
11. Goedecker S (2004) Minima hopping: searching for the global minimum of the potential energy surface of complex molecular systems without invoking thermodynamics. J Chem Phys 120:9911–9917
12. Curtarolo S, Morgan D, Persson K et al (2003) Crystal structures with data mining of quantum calculations. Phys Rev Lett 91:135503
13. Oganov AR, Glass CW (2006) Crystal structure prediction using evolutionary algorithms: principles and applications. J Chem Phys 124:244704
14. Lyakhov AL, Oganov AR, Valle M (2010) How to predict very large and complex crystal structures. Comput Phys Comm 181:1623–1632
15. Oganov AR, Lyakhov AO, Valle M (2011) How evolutionary crystal structure prediction works – and why. Acc Chem Res 44:227–237
16. Zhu Q, Oganov AR, Glass CW et al (2012) Constrained evolutionary algorithm for structure prediction of molecular crystals: methodology and applications. Acta Crystallogr B68:215–226
17. Lyakhov AO, Oganov AR, Stokes HT et al (2013) New developments in evolutionary structure prediction algorithm USPEX. Comput Phys Comm 184:1172–1182
18. Zhu Q (2013) Crystal structue prediction and its applications to Earth and materials sciences. Dissertation, Stony Brook University
19. Chaplot SL, Rao KR (2006) Crystal structure prediction – evolutionary or revolutionary crystallography? Curr Sci 91:1448–1450
20. Oganov AR, Chen J, Gatti C et al (2009) Ionic high-pressure form of elemental boron. Nature 457:863–867
21. Ma Y, Eremets MI, Oganov AR et al (2009) Transparent dense sodium. Nature 458:182–185
22. Li Q, Ma Y, Oganov AR et al (2009) Superhard monoclinic polymorph of carbon. Phys Rev Lett 102:175506
23. Zhu Q, Oganov AR, Salvado MA et al (2011) Denser than diamond: ab initio search for superdense carbon allotropes. Phys Rev B 83:193410
24. Lyakhov AO, Oganov AR (2011) Evolutionary search for superhard materials: methodology and applications to forms of carbon and TiO_2. Phys Rev B 84:092103
25. Zhu Q, Jung DY, Oganov AR et al (2013) Stability of xenon oxides at high pressures. Nat Chem 5:61–65
26. Zhu Q, Oganov AR, Lyahkov AO (2013) Novel stable compounds in the Mg–O system under high pressure. Phys Chem Chem Phys 15:7696–7700
27. Zhou XF, Oganov AR, Qian GR et al (2012) First-principles determination of the structure of magnesium borohydride. Phys Rev Lett 109:245503
28. Qian GR, Dong X, Zhou XF et al (2013) Variable cell nudged elastic band method for studying solid–solid structural phase transitions. Comput Phys Comm 184:2111–2118
29. Oganov AR, Valle M (2009) How to quantify energy landscapes of solids. J Chem Phys 130:104504
30. Price SL (2004) The computational prediction of pharmaceutical crystal structures and polymorphism. Adv Drug Deliv Rev 56:301–319
31. Lommerse JPM, Motherwell WDS, Ammon HL et al (2000) A test of crystal structure prediction of small organic molecules. Acta Crystallogr B56:697–714

32. Motherwell WDS, Ammon HL, Dunitz JD et al (2002) Crystal structure prediction of small organic molecules: a second blind test. Acta Crystallogr B58:647–661
33. Day GM, Motherwell WDS, Ammon HL et al (2005) A third blind test of crystal structure prediction. Acta Crystallogr B61:511–527
34. Day GM, Cooper TG, Cruz-Cabeza AJ et al (2009) Significant progress in predicting the crystal structures of small organic molecules: a report on the fourth blind test. Acta Crystallogr B65:107–125
35. Bardwell DA, Adjiman CS, Arnautova YA et al (2011) Towards crystal structure prediction of complex organic compounds: a report on the fifth blind test. Acta Crystallogr B67:535–551
36. Day GM (2011) Current approaches to predicting molecular organic crystal structures. Crystallogr Rev 17:3–52
37. Brock CP, Dunitz JD (1994) Towards a grammar of crystal packing. Chem Mater 6:1118–1127
38. Baur WH, Kassner D (1992) The perils of CC: comparing the frequencies of falsely assigned space groups with their general population. Acta Crystallogr B48:356–369
39. Johannesson GH, Bligaard T, Ruban AV et al (2002) Combined electronic structure and evolutionary search approach to materials design. Phys Rev Lett 88(255506):2002
40. Oganov AR, Ma Y, Lyakhov AO et al (2010) Evolutionary crystal structure prediction as a method for the discovery of minerals and materials. Rev Mineral Geochem 71:271–298
41. Lyakhov AO, Oganov AR (2010) Crystal structure prediction using evolutionary approach. In: Oganov AR (ed) Modern methods of crystal structure prediction. Wiley, Weinheim, pp 147–180
42. Kresse G, Furthmuller J (1996) Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. Phys Rev B 54:11169–11186
43. Perdew JP, Burke K, Ernzerhof M (1996) Generalize gradient approximation made simple. Phys Rev Lett 77:3865–3868
44. Blochl PE (1994) Projector augmented-wave method. Phys Rev B 50:17953–17979
45. Tekin A, Caputo R, Zuttel A (2010) First-principles determination of the ground-state structure of $LiBH_4$. Phys Rev Lett 104:215501
46. George L, Drozd V, Saxena SK et al (2009) Structural phase transitions of $Mg(BH_4)_2$ under pressure. J Phys Chem C 113:15087–15090
47. Cerny R, Filinchuk Y, Hagemann H et al (2007) Magnesium borohydride: synthesis and crystal structure. Angew Chem Int Ed 46:5765–5767
48. Her J, Stephens PW, Gao Y et al (2007) Structure of unsolvated magnesium borohydride $Mg(BH_4)_2$. Acta Crystallogr B63:561–568
49. Cerny R, Ravnsbak DB, Schouwink P et al (2012) Potassium zinc borohydrides containing triangular $[Zn(BH_4)_3]^-$ and tetrahedral $[Mg(BH_4)_xCl_{4-x}]^{2-}$ anions. J Phys Chem C 116:1563–1571
50. Ozolins V, Majzoub EH, Wolverton C (2008) First-principles prediction of a ground state crystal structure of magnesium borohydride. Phys Rev Lett 100:135501
51. Voss J, Hummelshj JS, Odziana Z et al (2009) Structural stability and decomposition of $Mg(BH_4)_2$ isomorphsan ab initio free energy study. J Phys Condens Matter 21:012203
52. Zhou XF, Qian GR, Zhou J et al (2009) Crystal structure and stability of magnesium borohydride from first principles. Phys Rev B 79:212102
53. Filinchuk Y, Richter B, Jensen TR (2011) Porous and dense magnesium borohydride frameworks: synthesis, stability, and reversible absorption of guest species. Angew Chem Int Ed 50:11162–11166
54. Fan J, Bao K, Duan DF et al (2012) High volumetric hydrogen density phases of magnesium borohydride at highpressure: a first-principles study. Chin Phys B 21:086104
55. Bil A, Kolb B, Atkinson R et al (2011) Van der Waals interactions in the ground state of $Mg(BH_4)_2$ from density functional theory. Phys Rev B 83:224103
56. Grimme S (2006) Semiempirical GGA-type density functional constructed with a long-range dispersion correction. J Comput Chem 27:1787–1799

57. Henkelman G, Arnaldsson A, Jonsson H (2006) A fast and robust algorithm for Bader decomposition of charge density. Comput Mater Sci 36:354–360
58. Levy HA, Agron PA (1963) The crystal and molecular structure of xenon difluoride by neutron diffraction. J Am Chem Soc 85:241–242
59. Templeton DH, Zalkin A, Forrester JD et al (1963) Crystal and molecular structure of xenon trioxide. J Am Chem Soc 85:817
60. Hoyer S, Emmler T, Seppelt K (2006) The structure of xenon hexafluoride in the solid state. J Fluorine Chem 127:1415–1422
61. Kim M, Debessai M, Yoo CS (2010) Two- and three-dimensional extended solids and metallization of compressed XeF_2. Nat Chem 2:784–788
62. Somayazulu M, Dera P, Goncharov AF et al (2010) Pressure-induced bonding and compound formation in xenon-hydrogen solids. Nat Chem 2:50–53
63. Smith DF (1963) Xenon trioxide. J Am Chem Soc 85:816–817
64. Selig H, Claassen HH, Chernick CL et al (1964) Xenon tetroxide: preparation and some properties. Science 143:1322–1323
65. Brock DS, Schrobilgen GJ (2011) Synthesis of the missing oxide of xenon, XeO_2, and its implications for Earth missing xenon. J Am Chem Soc 133:6265–626
66. Grochala W (2007) Atypical compounds of gases, which have been called 'noble'. Chem Soc Rev 36:1632–1655
67. Anders E, Owen T (1977) Mars and Earth: origin and abundance of volatiles. Science 198:453–465
68. Sanloup C, Hemley RJ, Mao HK (2002) Evidence for xenon silicates at high pressure and temperature. Geophys Res Lett 29:1883–1886
69. Sanloup C, Schmidt BC, Perez EMC (2005) Retention of xenon in quartz and Earth's missing xenon. Science 310:1174–1177
70. Oganov AR, Ma Y, Glass CW et al (2007) Evolutionary crystal structure prediction: overview of the USPEX method and some of its applications. Psi-K Newsletter 84:142–171
71. Caldwell WA, Nguyen JH, Pfrommer BG et al (1997) Structure, bonding, and geochemistry of xenon at high pressures. Science 277:930–933
72. Urusov VS (1977) Theory of isomorphous miscibility. Nauka, Moscow
73. Becke AD, Edgecombe KE (1990) A simple measure of electron localization in atomic and molecular systems. J Chem Phys 92:5397–5403
74. Frost DJ, Liebske C, Langenhorst F et al (2004) Experimental evidence for the existence of iron-rich metal in the Earth's lower mantle. Nature 428:409–412
75. Zhang FW, Oganov AR (2006) Valence state and spin transitions of iron in Earth's mantle silicates. Earth Planet Sci Lett 249:436–443
76. Lin J, Heinz DL, Mao HK et al (2003) Stability of magnesiowstite in Earth's lower mantle. Proc Natl Acad Sci U S A 100:4405–4408
77. Mazin II, Fei Y, Downs R et al (1998) Possible polytypism in FeO at high pressures. Am Mineral 83:451–457
78. Oganov AR, Martonak R, Laio A et al (2005) Anisotropy of Earth's D" layer and stacking faults in the $MgSiO_3$ postperovskite phase. Nature 438:1142–1144
79. Duffy TS, Hemley RJ, Mao HK (1995) Equation of state and shear strength at multimegabar pressures: magnesium oxide to 227 GPa. Phys Rev Lett 74:1371–1374
80. Mehl MJ, Cohen RE, Krakauer H (1988) Linearized augmented plane wave electronic structure calculations for MgO and CaO. J Geophys Res 118:8009–8022
81. Oganov AR, Gillan MJ, Price GD (2003) Ab initio lattice dynamics and structural stability of MgO. J Chem Phys 118:10174–10182
82. Belonoshko AB, Arapan S, Martonak R et al (2010) MgO phase diagram from first principles in a wide pressure–temperature range. Phys Rev B 81:054110
83. Wriedt H (1987) The MgO (magnesium-oxygen) system. J Phase Equil 8:227–233
84. Recio JM, Pandey R (1993) Ab initio study of neutral and ionized microclusters of MgO. Phys Rev A 47:2075–2082

85. Wang ZL, Bentley J, Kenik EA (1992) In-situ formation of MgO_2 thin films on MgO single-crystal surfaces at high temperatures. Surf Sci 273:88–108
86. Vannerberg N (2007) Peroxides, superoxides, and ozonides of the metals of groups Ia, IIa, and IIb. In: Cotton FA (ed) Progress in inorganic chemistry. Wiley, New York, p 2007. ISBN 9780470166055
87. Abrahams SC, Kalnajs J (1954) The formation and structure of magnesium peroxide. Acta Crystallogr 7:838–842
88. Efthimiopoulos I, Kunc K, Karmakar S et al (2010) Structural transformation and vibrational properties of BaO_2 at high pressures. Phys Rev B 82(134125):2010
89. Vannerberg NG (1959) The formation and structure of magnesium peroxide. Ark Kemi 14:99–105
90. Togo A, Oba F, Tanaka I (2008) First-principles calculations of the ferroelastic transition between rutile-type and $CaCl_2$-type SiO_2 at high pressures. Phys Rev B 78:134106
91. Olijnyk H, Holzapfel WB (1985) High-pressure structural phase transition in Mg. Phys Rev B 31:8412–4683
92. Wentzcovitch RM, Cohen ML (1988) Theoretical model for the hcp-bcc transition in Mg. Phys Rev B 37:5571–5576
93. Li P, Gao G, Wang Y et al (2010) Crystal structures and exotic behavior of magnesium under pressure. J Phys Chem C 114:21745–21749
94. Shannon RD, Prewitt CT (1969) Effective ionic radii in oxides and fluorides. Acta Crystallogr B25:925–946
95. Waber JT, Cromer DT (1965) Orbital radii of atoms and ions. J Chem Phys 42:4116–4123
96. Neumann MA, Perrin M (2005) The computational prediction of pharmaceutical crystal structures and polymorphism. J Phys Chem B 109:15531
97. Dion M, Rydberg H, Schroder E et al (2004) Van der Waals density functional for general geometries. Phys Rev Lett 92:246401
98. Román-Pérez G, Soler JM (2009) Efficient implementation of a van der Waals density functional: application to double-wall carbon nanotube. Phys Rev Lett 103:096102
99. Zhu Q, Li L, Oganov AR et al (2013) Evolutionary method for predicting surface reconstructions with variable stoichiometry. Phys Rev B 87:195317

Top Curr Chem (2014) 345: 257–290
DOI: 10.1007/128_2013_490
© Springer-Verlag Berlin Heidelberg 2013
Published online: 7 November 2013

Large-Scale Generation and Screening of Hypothetical Metal-Organic Frameworks for Applications in Gas Storage and Separations

Christopher E. Wilmer and Randall Q. Snurr

Abstract Metal-organic frameworks (MOFs) are porous crystals that are synthesized in a building-block approach that greatly facilitates rational design. MOFs are promising materials for gas storage and separation applications, but they are also intriguing for their potential use as catalysts, electrodes, and drug delivery vehicles. For these reasons, MOFs have spurred a renewed interest in the concept of "crystal engineering," where the crystal structure of a material is designed to meet application-specific criteria. This chapter reviews recent work in the computational design of MOFs, with an emphasis on high-throughput methods that generate and screen many thousands of candidates automatically.

Keywords Adsorption · Molecular modeling · Porous coordination polymers · Porous crystals

Contents

C.E. Wilmer (✉)
Department of Chemical and Petroleum Engineering, University of Pittsburgh, 1249 Benedum Hall, 3700 O'Hara Street, Pittsburgh 15261, PA
e-mail: wilmer@pitt.edu

R.Q. Snurr
Department of Chemical and Biological Engineering, Northwestern University, 2145 Sheridan Road, Evanston, IL 60208, USA

1 Unique Properties of Metal-Organic Frameworks

Metal-organic frameworks (MOFs) are crystalline materials that share much in common with (non-crystalline) highly cross-linked polymers [1, 2]. Like conventional polymers, MOFs are synthesized by the self-assembly of molecular "building blocks" that form into an extended structure. Unlike most polymers, however, the extended structure is a rigid, three-dimensional *porous* crystal whose order can be maintained over millimeter scales. MOFs are so named because the monomers are divided into two distinct groups: metal ions (derived from dissolved metal salts) and the organic ligands that coordinate to them (see Fig. 1). Before the term MOF came into widespread use [3, 4] these materials were (and still are) referred to as coordination polymers [5–7] and, more recently, porous coordination polymers (PCPs) [1, 8, 9]. Note, however, that amorphous materials have been referred to as coordination polymers [7], but MOFs are exclusively crystalline.

A significant milestone in the history of MOFs was reached when it was discovered that the solvent (used in the synthesis procedure) could be removed, leaving behind a freestanding "permanently porous" structure with a very high internal surface area [3, 4, 10, 11].

Two of the earliest permanently porous MOF structures are HKUST-1, reported by Chui et al. [10], and MOF-5 (later named IRMOF-1) by Li et al. [4]. The former is formed from the self-assembly of benzene-1,3,5-tricarboxylic acid and a copper salt and the latter from benzene-1,4-dicarboxylic acid and a zinc salt. In both cases the metal salt and organic ligand were dissolved in dimethylformamide (DMF) solvent at elevated temperatures for 12–24 h. In each case, the synthesis resulted in micrometer scale crystals that could be analyzed by X-ray diffraction to obtain the detailed crystal structures (see Fig. 2). The discovery of these and other permanently porous MOFs [8, 11] over a decade ago catapulted MOFs from being materials of purely scientific interest to materials of potential industrial importance.

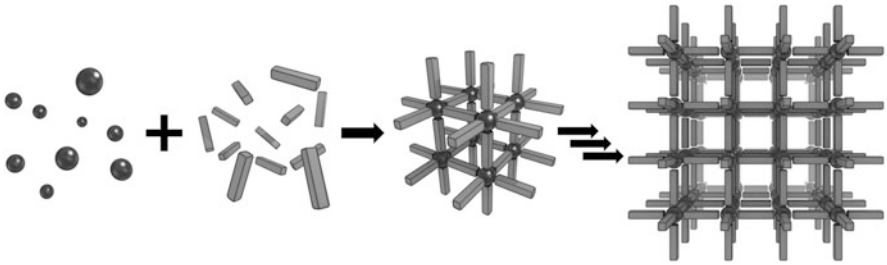


Fig. 1 Schematic of MOF self-assembly. MOFs are synthesized by the self-assembly of organic and metal-containing (inorganic) building blocks to form extended crystalline frameworks. Note that MOFs can have a wide variety of framework topologies beyond the cubic framework depicted here

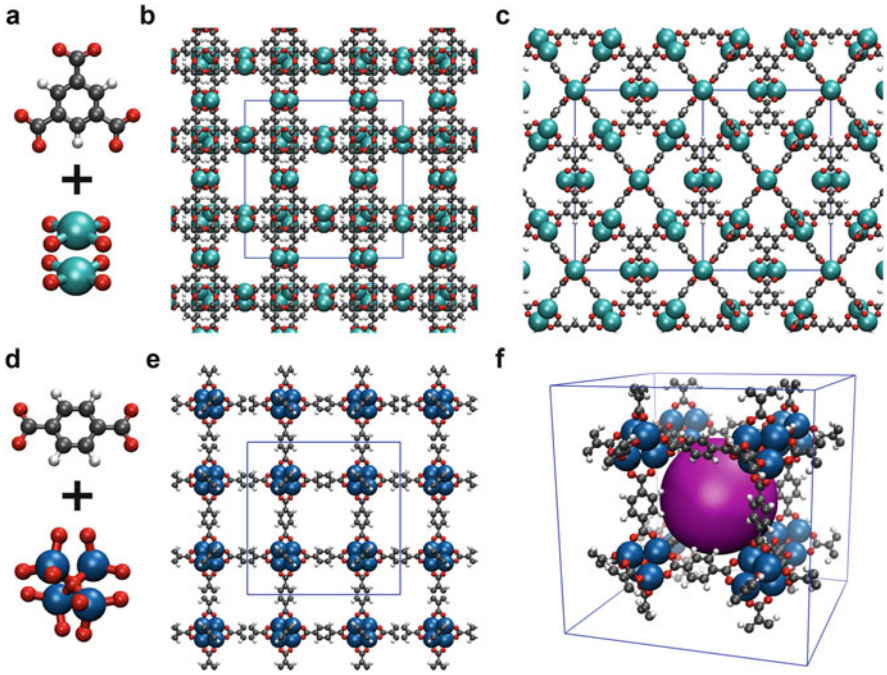


Fig. 2 Building blocks and crystal structures of (**a–c**) HKUST-1 and (**d–f**) IRMOF-1. (**a**) Benzene-1,3,5-tricarboxylic acid and copper ions, which arrange into octahedral "paddle wheel" clusters in solution, form (**b**, **c**) the HKUST-1 structure [10]. (**d**) Benzene-1,4-dicarboxylic acid and zinc ions, which form Zn_4O tetrahedral clusters in solution, form **(e, f)** the IRMOF-1 structure [4]. *Grey*, *white*, *red*, *cyan*, and *blue* spheres represent carbon, hydrogen, oxygen, copper, and zinc atoms, respectively. The *purple sphere* **(f)** indicates the size of the pore that is available to guest molecules

1.1 Utility in Gas Storage and Separation Applications

In the past decade a significant impetus for studying MOFs has been their potential use as adsorbents for industrial gas storage and separations applications [12]. MOFs can have surface areas as high as 7,200 m^2/g, significantly higher than the best activated carbon or zeolite materials [13]. Because of their high internal surface areas, MOFs have a high density of adsorption sites that can bind gases of interest, which collectively concentrate the gas without increasing the pressure. This phenomenon can have a dramatic effect on gas storage under certain conditions. For example, at 35 bar and 298 K, a vessel filled with **MOF-177** would store as much CO_2 as nine equally sized vessels without a sorbent material [14]. There has also been marked interest in developing MOFs to store gaseous fuels, such as methane [15, 16] and hydrogen [17, 18], compactly in vehicles.

In addition to storage, MOFs can selectively adsorb certain gases over others, which makes them promising for separating and purifying mixtures of gases [19]. Adsorption-based gas separation is an alternative to distillation, which is a pervasive and energetically costly process in the chemicals industry [20]. In the recent literature, MOFs have been reported as promising adsorbents for separating mixtures of H_2/CO_2 [21], H_2/NH_3 [21], CH_4/CO_2 [22], olefin/paraffin mixtures [23], and *p*-, *o*-, *m*-xylene mixtures [24, 25], among many others [26–30].

It would be particularly beneficial to industry to *design* a MOF that is optimal for a particular industrial process, as opposed to relying on serendipitous discovery. This is not an unrealistic goal, since both the crystal structure and gas adsorption behavior of a MOF can potentially be computationally predicted a priori.

1.2 Predictable Self-Assembly

Since the molecular building blocks used in their self-assembly only coordinate in very specific orientations and stoichiometries, MOF structures can be relatively straightforward to predict. The creation of new MOFs based on the known shape and connectivity of the building blocks has come to be called "reticular design" [31] and reports of thousands of new structures over the past decade are a testament to the reliability of this approach [32].

However, knowledge of the final structure does not equal knowledge of the detailed synthesis pathway. It is still a significant challenge in the field of MOFs to find the synthesis conditions that lead to the formation of a desired crystal structure. To this end, high throughput robotics have been developed to test rapidly thousands of synthesis conditions given a particular choice of building blocks [33, 34].

Even with predictable self-assembly, the development of MOFs for industrial applications must rely on chemical intuition and trial-and-error testing unless we can predict how the crystal structure determines the gas adsorption behavior. Fortunately, significant advances have been made in accurately modeling gas adsorption in MOFs using molecular simulations.

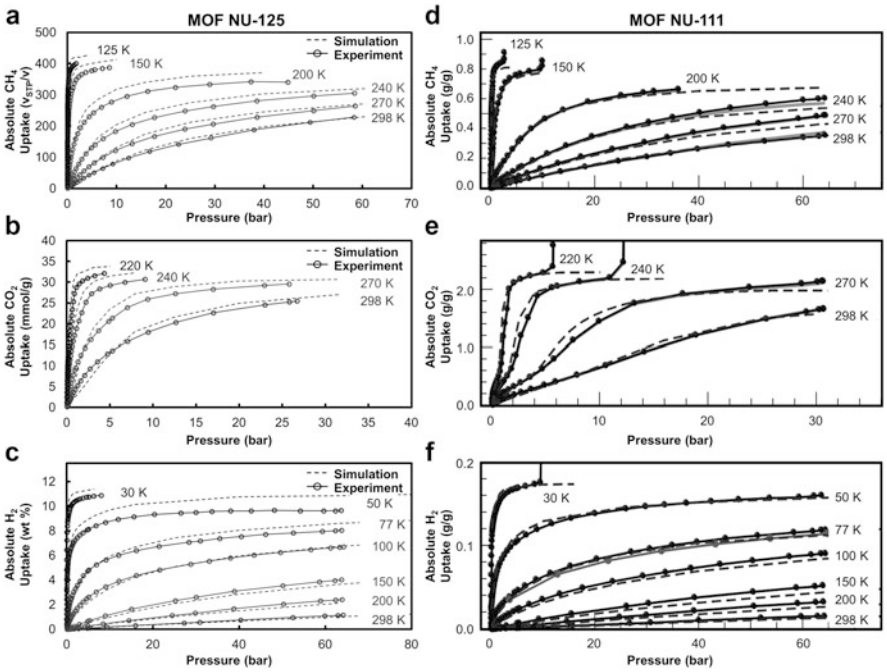


Fig. 3 Accuracy of molecular simulations for predicting gas adsorption in two different MOFs. **(a–c)** Simulated (*dashed lines*) and experimentally measured (*solid lines*) adsorption isotherms for three different gases over a wide range of temperatures for the MOF **NU-125** [16]. **(d–f)**, Simulated (*dashed lines*) and experimentally measured (*solid lines*) adsorption isotherms for three different gases over a wide range of temperatures for the MOF **NU-111**. Note that no fitting was used. Figure parts **(a–c)** adapted from [16] with permission from The Royal Society of Chemistry. Figure parts **(d–f)** adapted from [38] with permission from The Royal Society of Chemistry

1.3 Predictable Gas Adsorption Behavior

For gases that are relatively inert, molecules interact weakly with the walls of a MOF and do not appreciably change their electronic structure (i.e., no new bonds are formed or broken). Thus, under these conditions (i.e., physisorption rather than chemisorption) the details of the electronic structure for both the gas molecules and framework structure can be ignored and all atoms can be modeled as classical particles where inter-atomic interactions are governed by Lennard Jones and Coulombic potentials (for more details see Sect. 5.2) [35, 36]. While such models may seem overly simple, they have often predicted gas adsorption behavior in remarkable agreement with experimental measurements for a variety of gases over a wide range of temperatures and pressures (e.g., see Fig. 3) [15, 16, 35, 37, 38]. Such close agreement between experimental measurements and simulation data would not be possible if not for the crystalline nature of MOFs.

Certain gas-temperature combinations are more challenging to model accurately than others. Cryogenic hydrogen adsorption requires taking quantum diffraction effects into account [35, 39–41]. Notably, at low pressures and temperatures, where adsorption is dominated by host–guest interactions (rather than guest–guest) [42], the accuracy of adsorption predictions will depend strongly on how the strongest interaction sites are modeled. For example, simulations of water adsorption [43–45] and CO_2 adsorption at sub-atmospheric pressures are often challenging [46, 47] due to strong interactions that occur at the open metal sites [48, 49], which are not well described by Lennard Jones potentials [35, 50].

However, even when molecular simulations are unable to predict gas adsorption accurately, computational analysis of MOFs can precisely calculate pore sizes, surface areas, and other properties of indirect importance to the application of interest. It is also sometimes possible to use molecular simulations to rank correctly MOFs from best to worst, even when the predictions are not quantitatively accurate [46, 47, 51].

1.4 Structure Tunability

Perhaps the most attractive feature of MOFs is the enormous diversity of possible structures, due to the practically unlimited variety of organic ligands that can be used. A small subset of the many organic ligands that have been incorporated in MOF structures is shown in Fig. 4. Therefore, MOFs can have a very wide range of possible pore geometries and surface chemistries [31, 52]. By choosing the building blocks appropriately, one can tune the properties of the resulting MOF to behave in an optimal way for a given context [53]. For example, in the context of membranes for gas separations, one might create a material with pores tuned to be just large enough to let through one gas species (perhaps the desired product), but too small for the other species in the mixture. Recent reports of "post-synthesis" functionalization [54] and "solvent assisted ligand exchange" [55, 56] even allow for nuanced modifications to MOF structures and continue to expand what it is possible to synthesize.

This tunability presents a challenge however: given the ability to create almost any kind of MOF structure, *which* one should we make? This chapter mainly addresses this challenge.

2 Challenges in MOF Design

2.1 Large Space of Possible MOFs

As alluded to above, the space of possible MOFs is vast. Even when considering only hundreds of molecular building blocks (i.e., organic ligands and metal salts),

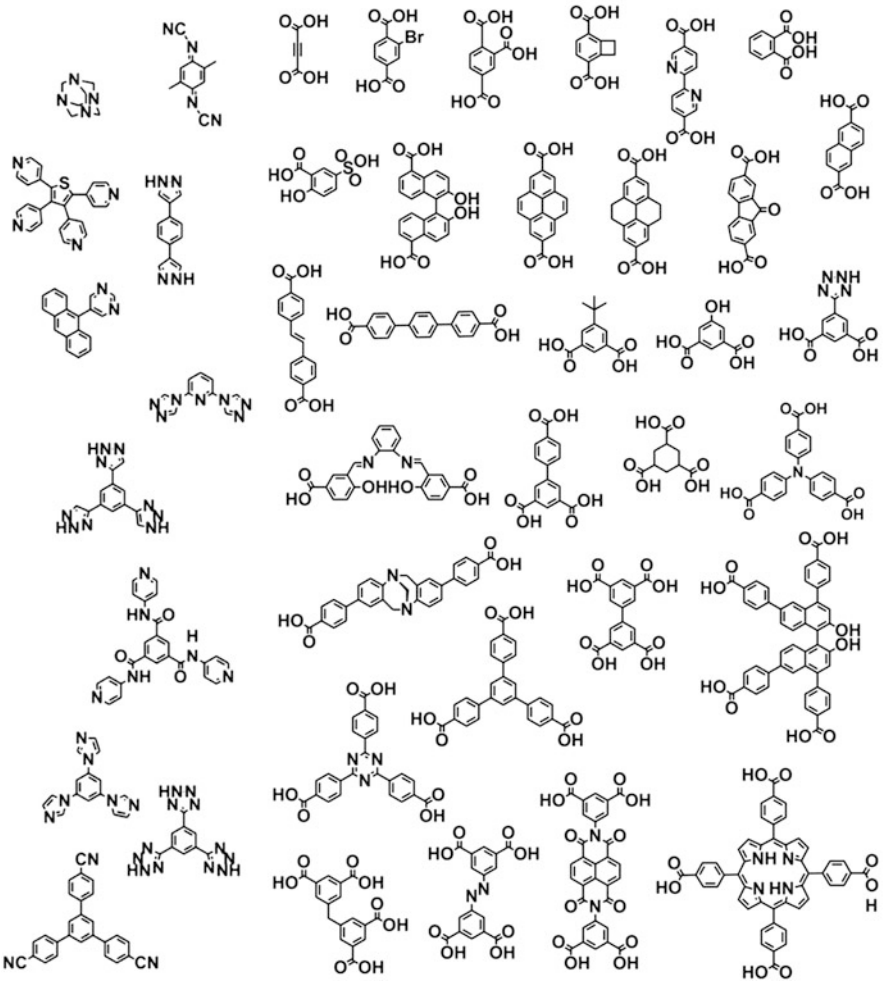


Fig. 4 A selection of organic ligand building blocks for MOFs. Figure adapted and reprinted with permission from [58]. Copyright 2013 American Chemical Society

the combinatorial possibilities allow for millions of hypothetical MOF structures [57]. Given that the universe of organic chemistry from which the ligands are chosen is itself vast (and growing), exploring the space of possible MOFs appears daunting. Although design by intuition continues to yield materials with impressive and successively better properties, maximizing the potential of MOFs will be greatly accelerated by efficient methods to search this enormous space for optimal candidates.

2.2 *Difficult to Predict Structural Details*

While from the geometry and connectivity of the building blocks it is possible to get an approximate sense of the self-assembled MOF structure, certain structural details can be difficult to estimate without appealing to more sophisticated calculations. For example, even after organic ligands coordinate to the metal ions in the framework, they may retain rotational or other non-translational degrees of freedom [59, 60]. The detailed geometry of the pores of a MOF in such cases may be temperature dependent, independent of expected thermal expansion effects. If the organic ligands exhibit conformational degrees of freedom then even the resulting framework topology can depend on the synthesis conditions, resulting in supramolecular isomerism [53, 61].

MOFs are also able to self-assemble such that one (or more) framework is interpenetrated (also called "catenated") within another [62, 63] (see Fig. 5). A common rule of thumb is that if there is enough space for another framework, then the MOF will interpenetrate. However, it is possible to control interpenetration via the synthesis procedure [63]. For example, an intermediate MOF can be constructed from an organic ligand with a bulky leaving group that does not leave room for interpenetration. By subsequently removing this bulky leaving group, a non-interpenetrated MOF with "extra" space can be synthesized.

It is not only challenging to predict whether or not MOFs will (or have the ability to) interpenetrate; it is also difficult to know how the interpenetrated frameworks will pack relative to each other. Whether or not the interpenetrated frameworks are packed tightly (see Fig. 6) can have a significant impact on the gas adsorption properties [64].

The inorganic building blocks, referred to sometimes as secondary building units (SBUs) [65, 66], are themselves self-assembled from dissolved metal salts. The structure of these precursor assemblies would be very challenging to predict from ab initio calculations, but MOF design is broadly based on the assumption that, under similar conditions, dissolved metal salts will always form the same SBUs. Thus, the SBUs can be thought of as rigid building blocks in the same way as the rigid organic ligands. Nevertheless, many distinct SBUs are derived from the same metal salts but under different conditions, and so one should be cautious when predicting MOF structures on the assumption that a particular SBU will form.

Finally, it is perhaps fundamentally impossible to predict the location and arrangement of defects (e.g., missing organic ligands) or of interchangeable ligands in so-called multivariate MOFs (MTV-MOFs) [67]. In the latter, ligands that are identical except for having different chemical functional groups are allowed to self-assemble simultaneously, resulting in crystalline materials with a random spatial distribution of functional groups. Creating MOFs with multiple chemical functional groups in a single crystal is attractive for catalytic applications and for gas masks, where each functional group can respond to a different toxic molecule. Unfortunately, designing such an MTV-MOF cannot currently rely on any particular arrangement of functional groups.

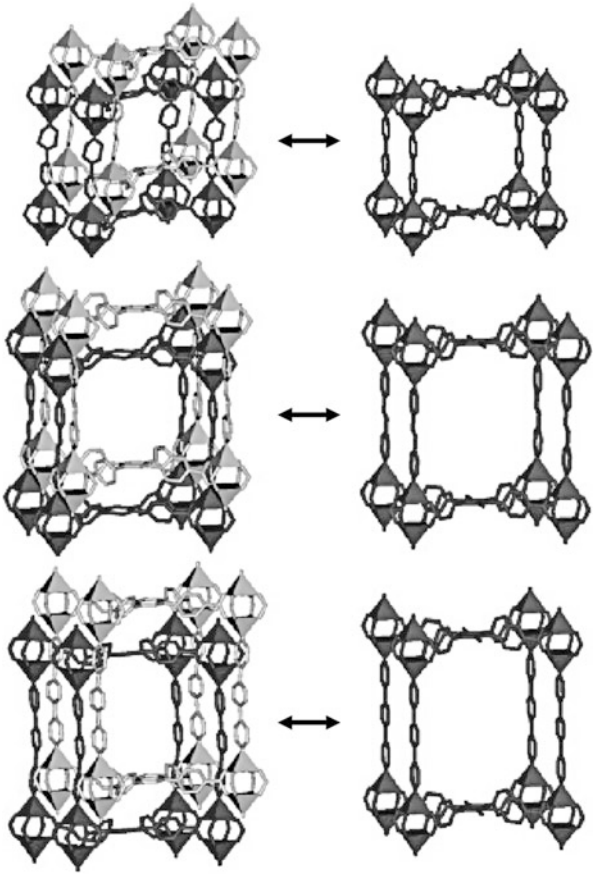


Fig. 5 Framework interpenetration can occur in MOFs, where one framework grows inside another. Whether or not this occurs can depend on the synthesis path taken to produce the final structure. Figure adapted and reprinted with permission from [63]. Copyright 2010 American Chemical Society

2.3 Unclear Structure–Property Relationships

Even if it were possible to know the detailed MOF structure based on the choice of building blocks, that would still not answer the question of which MOF should be designed for a particular engineering problem. This is because, in general, we do not know the relationship of the crystal structure to the gas adsorption property of interest (without performing either experiments or detailed molecular simulations).

For a single MOF, or a small set, it is possible to predict the gas adsorption properties of each structure using molecular simulations. However, as described above, the space of possible MOFs is so vast that even computational trial-and-error

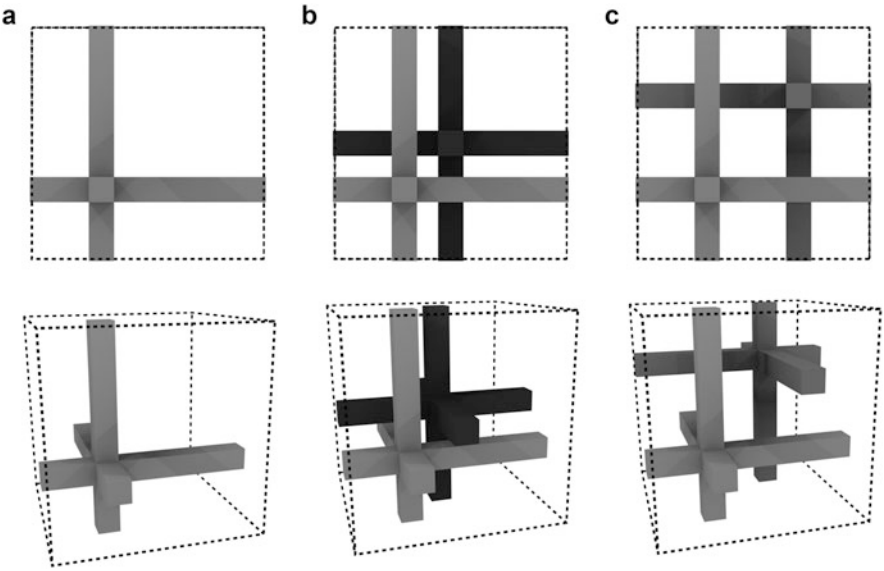


Fig. 6 Based on the choice of building blocks, it is sometimes easy to predict the structure of the framework **(a)**, but not whether multiple frameworks will interpentrate **(b**, **c)**. Additionally, the spacing between the interpenetrated frameworks, which can be small **(b)** or large **(c)**, can be difficult to predict

is inefficient. The discovery of structure–property relationships, ideally ones that could be expressed as analytical equations, would be the ultimate tool in designing optimal MOF structures. Unfortunately, only a few such relationships have been tentatively proposed in the literature [42, 68, 69], and those only apply to specific gases under a limited range of conditions (see Fig. 7.)

As will be discussed later in the chapter, the use of high throughput computational screening methods has provided unprecedented clarity in the form of highly resolved structure–property relationships for certain gases. However, this is still an area where much work needs to be done.

3 Strategies for MOF Design

3.1 High Throughput Experimental Synthesis

Although the focus of this chapter is on computational methods, it is worth mentioning that a valid approach to finding useful MOFs from the vast sea of possibilities is high throughput experimental synthesis using sophisticated robotic equipment [33, 34, 70]. While robotic equipment presents a tremendous improvement in speed

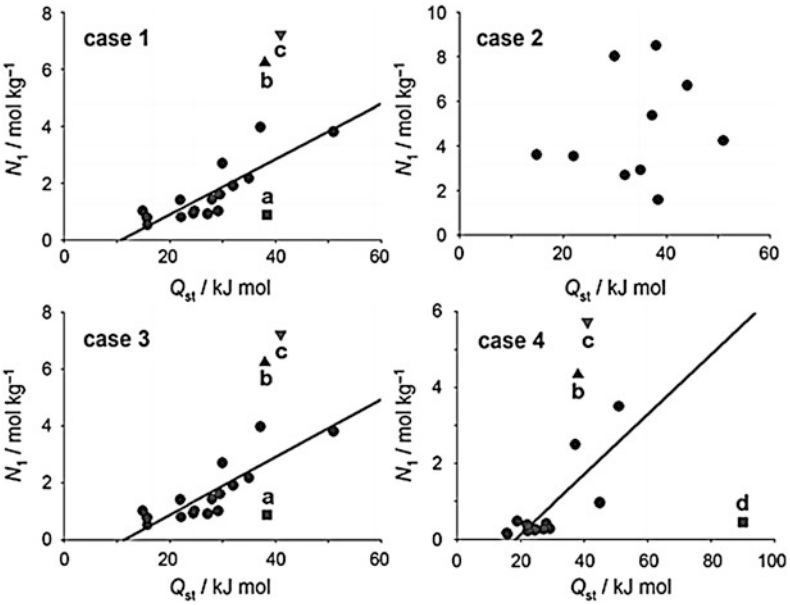


Fig. 7 The relationship between CO_2 adsorption, N_1, and the heat of adsorption, Q_{st}, in over 40 different MOFs in 4 cases corresponding to different pressures. The *points* represent experimental measurements and the *solid black lines* represent linear fits to the data, which are drawn for the purpose of identifying structure–property relationships. Cases 1–4 correspond to pressures of 0.5, 2.5, 0.5, and 0.1 bar, respectively. Materials labelled *a–d* were not used in the linear fit. Figure obtained from [68]. Copyright 2011 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

over traditional synthesis workflows, it is unlikely to become fast enough to cover a significant fraction of the space of possible MOF structures. To truly harness the vast space of MOFs, high speed computational methods are needed.

3.2 *Simulating the Self-Assembly of Hypothetical MOF Structures*

An intuitive computational approach to exploring hypothetical MOF structures is to model the self-assembly process that takes place during synthesis for many combinations of molecular building blocks. Based on what is known today, molecular simulations that take into account chemical reactions amongst thousands of molecules simultaneously are either prohibitively costly or insufficiently accurate. However, if only the final crystal structure is desired, rather than the intermediate details, then the self-assembly process can be modeled in a simpler way that ignores much of the unimportant physics.

This is the approach taken by Mellot-Draznieks et al. in what is called the "automated assembly of secondary building units" (AASBU) approach [71]. Here, building blocks referred to as SBUs, are treated as rigid bodies that are assigned "sticky sites" that have no physical significance except to cause SBUs to bind to one another via a Lennard–Jones interaction potential. The SBUs are initially distributed randomly in a periodic unit cell and then allowed to settle into an ordered crystal structure by simulated annealing Monte Carlo [72]. This procedure can generate thousands of plausible structures in a short period of time.

The AASBU approach allows one to explore the phase space of MOFs orders of magnitude more quickly than by high throughput experimental methods. However, the iterative nature of energy minimization schemes requires a baseline level of computational expense that can potentially be avoided by other approaches. In the following sections we describe non-iterative methods of generating hypothetical MOF structures.

3.3 Non-iterative Generation of Hypothetical MOF Structures

Rather than arranging building blocks randomly in space and then minimizing the energy of the system, one can potentially choose the initial configuration in such a way that minimization is practically unnecessary. Here, the computational complexity is shifted from iterative energy minimization to searching the space of logical arrangements of chemical building blocks (i.e., candidate "initial configurations") based on geometrical, topological, and chemical considerations.

While there are potentially many ways to approach non-iterative MOF generation, we have broadly defined two categories: "top-down" and "bottom-up" generation.

3.3.1 "Top-Down" Generation: From Network Topology to Hypothetical MOF

For over a century there has been interest in categorizing crystalline structures by various mathematical descriptions [73–75]. One approach, in which crystal structures are described by periodic graphs called nets, has played a central role in MOF design. Each node and edge in a net represents a group of atoms, with the atoms of the inorganic building blocks often grouped as nodes (see Fig. 8). The assignment of atoms to nodes and edges is subjective, but the study of nets has nevertheless been very significant in predicting how building blocks can connect into periodic structures.

Many experimentally synthesized zeolites and MOF crystal structures are described a posteriori as corresponding to a particular net. Therefore, a potential

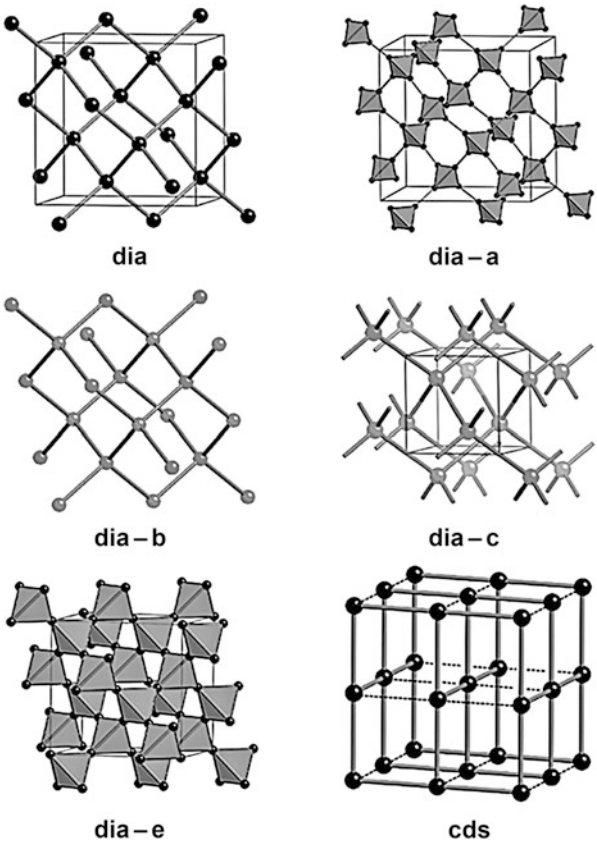


Fig. 8 Examples of six different periodic graphs, called nets, that can represent the underlying topology of a MOF. Reprinted with permission from [75]. Copyright 2008 American Chemical Society

strategy for generating a candidate MOF structure computationally is to begin with a known net, and then substitute chemical building blocks in place of the nodes and edges as appropriate. For this approach, however, a database of nets or a net generation algorithm is needed.

It is possible to use nets from experimentally discovered MOFs, such as those kept in the Reticular Chemistry Structure Resource (RCSR) database [75]. However, the number of experimentally synthesized MOFs thus far is a negligible fraction of the total MOF space (there are 2,031 nets available as of writing in the RCSR database). There may also be specific interest in designing MOFs corresponding to new nets, which is fundamentally not possible if only the known nets are used to generate MOF structures.

There have recently been significant advances in the mathematical understanding of nets, and this has led to the development of algorithms that can systematically enumerate them (see Fig. 9) [73, 76–78]. An exciting property of enumerative algorithms is that there is the possibility of *comprehensively* generating all possible nets (within certain constraints on complexity). In the context of computational screening in the search of optimal materials, it is reassuring to know that no structure, or subset of structures, was missed that might have had better properties.

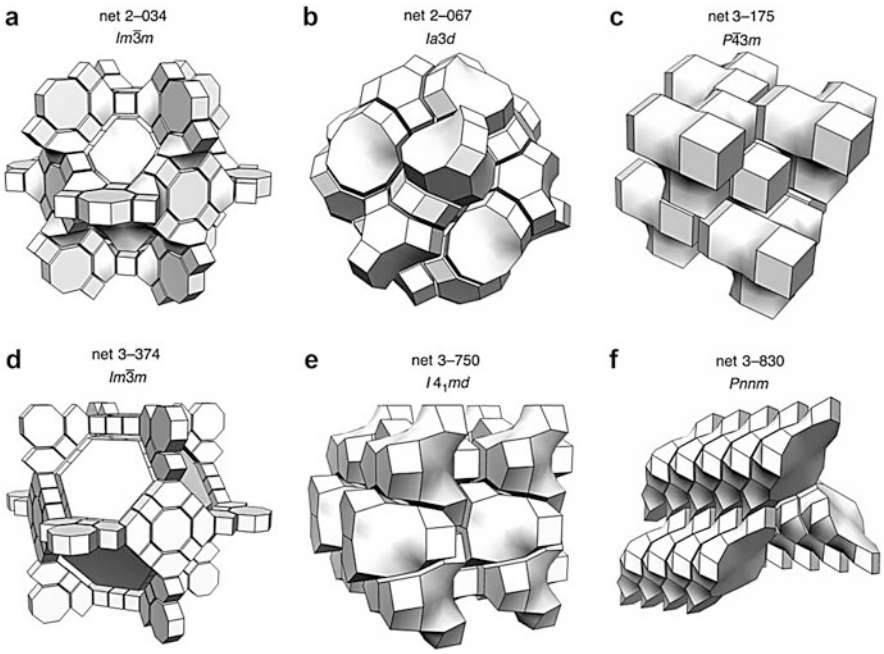


Fig. 9 Six examples of enumerated nets from the algorithm of Delgado Friedrichs et al. [73]. Below each net ID is the space group. Reprinted with permission from [73]. Copyright 1999 Nature Publishing Group

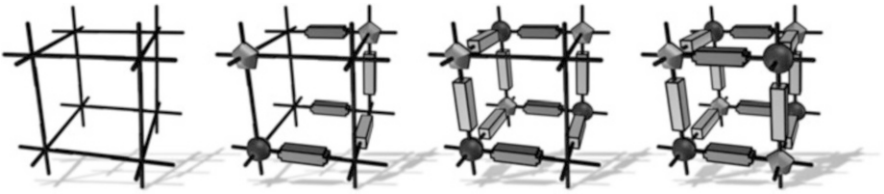


Fig. 10 Schematic illustration of the top-down approach to generating hypothetical MOFs. First a net is chosen (*far left*), and then chemical building blocks are appended to the net, resulting in a chemically detailed MOF structure (*far right*)

Such algorithms are potentially ideal for generating hypothetical MOF structures since they provide a large and potentially comprehensive set of nets onto which chemical building blocks can be added (see Fig. 10.) This strategy has been used to generate zeolite-like microporous solids systematically [79], and recently Bureekaew and Schmid reported a hypothetical covalent organic framework (COF) generation scheme that follows a top-down approach [80]. It is also worth pointing out that non-systematic top-down approaches have been used to design new MOFs one at a time, such as with the design of **NU-100** [81].

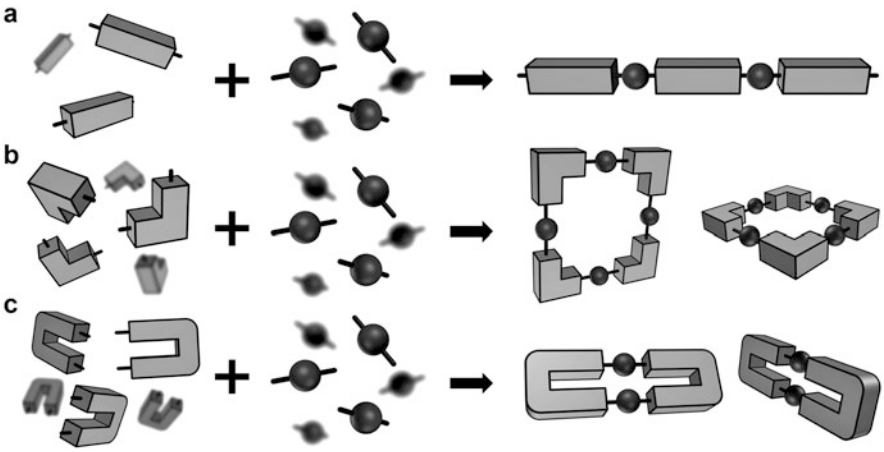


Fig. 11 The shape of the building block, in addition to its connectivity, plays a critical role in determining the shape and even dimensionality of the final self-assembled structure. An organic 2-connected building block can self-assemble with a 2-connected inorganic building block to form **(a)** linear chains, **(b)** squares, or **(c)** rods, depending on the organic building block's shape

For a given net, only certain combinations of chemical building blocks will match the topology of the net (e.g., a net may have both vertices with three edges and vertices with two edges, which would require a combination of 2-connected and 3-connected chemical building blocks). Since the number of possible nets is very large, a large fraction of nets may not be compatible with the building blocks in one's library. Another challenge is that the geometry of the building blocks, which is not explicitly specified by the net, can affect the topology of the crystal structure. In the supramolecular chemistry literature it is known that the assembly of building blocks into larger structures depends on the shape and size of the building blocks, in addition to the degree of connectivity [82]. A 2-connected organic building block self-assembling with a 2-connected metal node may form an infinite chain, square, or rod, depending on the bend angle of the linker, as shown in Fig. 11.

Therefore, the top-down approach requires matching a net to a set of compatible building blocks in terms of both connectivity and shape. An interesting strategy may be to generate de novo building blocks that are compatible with a given net and are constrained to be reasonable candidates for chemical synthesis.

3.3.2 "Bottom-Up" Generation: Connecting Building Blocks into Crystals

Another approach, which does not make use of nets explicitly, is to begin with the building blocks and connect them together sequentially until they form a logical periodic structure (i.e., where all connection points are satisfied) (see Fig. 12).

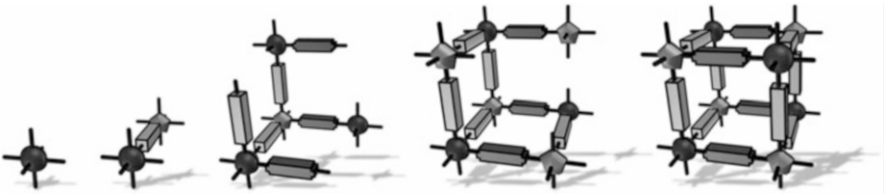


Fig. 12 A schematic of bottom-up generation of hypothetical MOF structures. Chemical building blocks are connected together until they form a periodic, chemically detailed structure

Although it is not known a priori whether any particular combination of building blocks is able to form a plausible MOF structure, the scale of the computational search is a function of the size of the building block library (whereas the space of possible nets is independent of the building block library size). It is not clear whether the top-down or bottom-up approach is computationally more efficient. As will be described below, in our first implementation of a bottom-up strategy, over 100,000 hypothetical MOF structures could be generated in less than 24 h on a single CPU [57].

4 Bottom-Up MOF Generation Details

4.1 Creating MOF Building Blocks

We created a library of building blocks by extracting fragments from experimentally determined MOF crystal structures. These building blocks were later recombined in various ways to form new hypothetical MOF structures, as shown in Fig. 13. Although it is beyond the scope of this chapter, it is worth noting that the partitioning of a crystal structure into fragments in such a way that they can be recombined into many different structures is a challenging problem; we relied on human intuition and manual inspection, but potentially pattern recognition and other techniques from computer science could have been used. An important aspect to creating these fragments such that they were *modular* was to partition the MOF at junctures that occurred frequently in MOF materials (e.g., the point where carboxyl-terminated ligands coordinate to a metal).

Each extracted fragment is assigned a number of connection sites, and each site contains information about the chemical details at the fragment boundary, such as which building blocks can combine with each other. Information about the relative spatial arrangement of two fragments before they were extracted is preserved by a set of vectors associated with each connection site, as shown in Fig. 14.

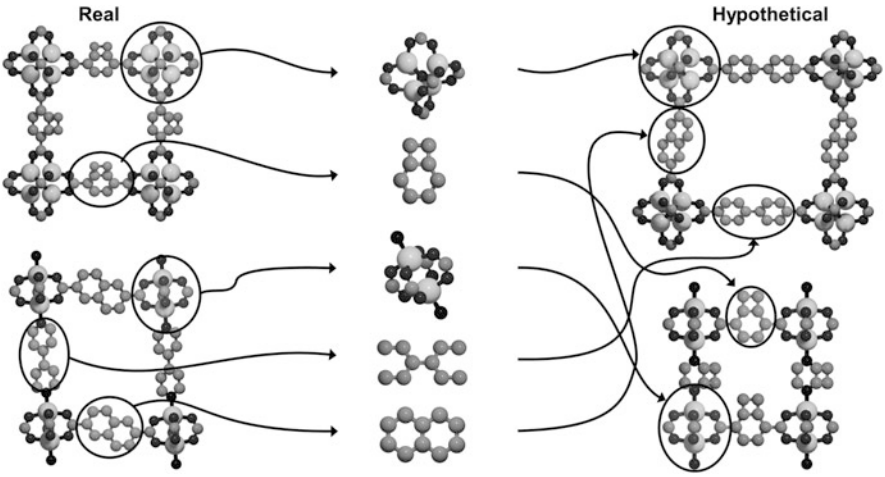


Fig. 13 Visual summary of the bottom-up hypothetical MOF generation strategy. Crystal structures of existing MOFs are obtained from X-ray diffraction data (*left*), and are subsequently divided into building blocks (*middle*) that can then be recombined to form new, hypothetical MOFs (*right*). Figure adapted and reprinted with permission from [57]. Copyright 2012 Nature Publishing Group

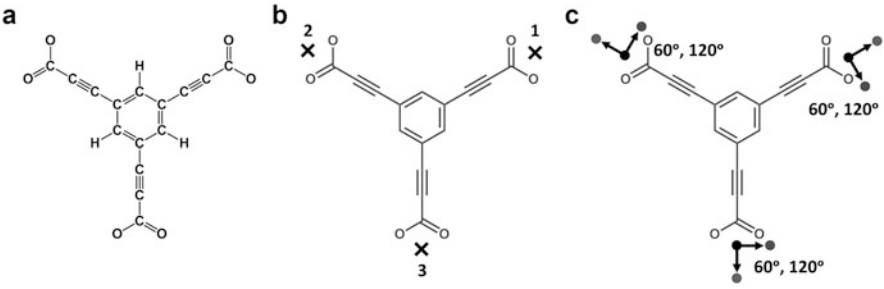


Fig. 14 Encoded in each fragment (i.e., building block) are the **(a)** atom composition and geometry, **(b)** topological information via numbered connection sites, and **(c)** geometrical information in the form of orientation vectors. To allow for rotational degrees of freedom (i.e., for building blocks that can rotate relative to one another), a list of angles for alternative orientations is also included. Reprinted with permission [57]. Copyright 2012 Nature Publishing Group

4.2 Assembling MOFs Block by Block

Generating a MOF can be described as a sequence of decisions, starting with choosing a building block in the library, then choosing a second building block and how it will connect to the first one, and so on until a complete structure is formed. This process is depicted in Fig. 15. Building blocks are combined stepwise, and if an atomic overlap occurs at a particular step, a different building block is

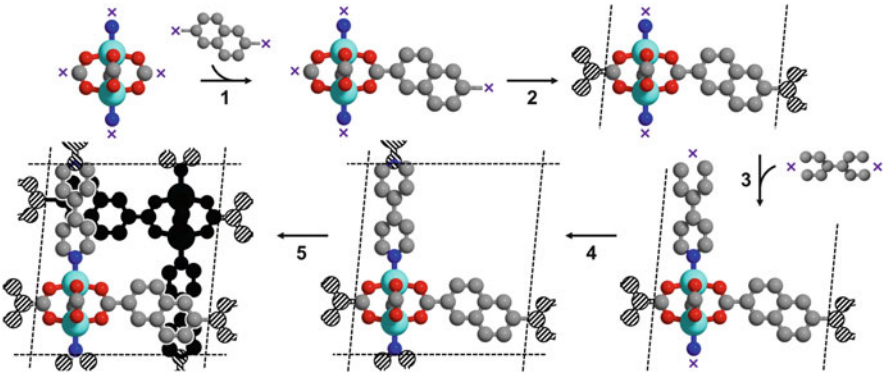


Fig. 15 The assembly process occurs by stepwise addition of building blocks (**1**), which are attached at their connection sites (*purple Xs*). Building blocks are also connected across periodic boundaries (**2,** *hashed circles indicate mirror images*). The process repeats (**3**, **4**) until all connection sites are utilized. An interpenetrated MOF may be generated if enough space exists (**5,** *black circles indicate atoms belonging to one of two interpenetrated frameworks*). *Gray*, *red*, *blue*, and *turquoise spheres* represent carbon, oxygen, nitrogen, and zinc atoms, respectively. Hydrogen atoms have been omitted for clarity. Reprinted with permission from [57]. Copyright 2012 Nature Publishing Group

chosen or a different connection site, until all possibilities are exhausted. While the total number of steps in each generation process can vary, there are always three steps when, instead of adding a building block, a periodic boundary is imposed by connecting any two building blocks (see steps 2 and 4 in Fig. 15). When no more building blocks can be added, the crystal generation procedure is complete. (Note that no force field or quantum mechanical energy minimizations are involved.)

Since after deciding on the *first* building block there are many distinct *second* decisions, the space of all MOF generation attempts can be described by a decision tree. The number of branches that lead to failed attempts (i.e., illogical structures with only partially connected building blocks or building blocks that overlap sterically) is vastly greater than the number of successful structures (i.e., plausible hypothetical MOFs).

Here we attempt to crudely quantify lower bounds on the size of this decision tree. The number of possible hypothetical MOFs (where we consider every decision sequence a "possible" hypothetical MOF) can be estimated based on the size of the library of modular building blocks (from here on assumed to be 100) and a few simplifying assumptions.

Let's consider the case of MOFs composed of only one type of inorganic building block and one type of organic building block. Let L be the number of organic building blocks (L as in "linkers") and C be the number of inorganic building blocks to choose from (C as in "corners"). Linkers may only connect with corners, and vice versa. The number of possible MOFs, N, is simply $N = L \times C$, which corresponds, for example, to 900 for $L = 90$ and $C = 10$.

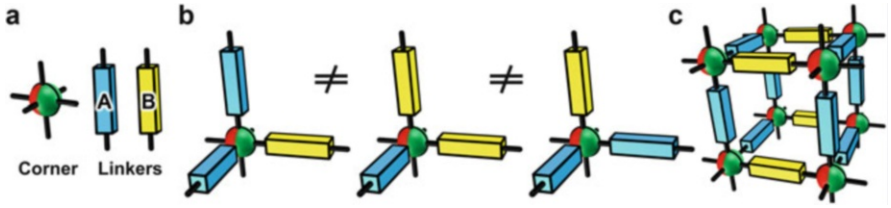


Fig. 16 MOFs that contain two distinct linkers (A-type, *blue* and B-type, *yellow*) (**a**) may vary in the ratio of A to B linkers (**b** – *left vs middle*) or in the arrangement of those linkers at a fixed ratio (**b** – *left vs right*). A larger fragment of the schematic MOF framework is shown in (**c**) for clarity. Reprinted with permission from [57]. Copyright 2012 Nature Publishing Group

Now we can consider the case where a unit-cell of a MOF contains M linkers (not to be confused with L: the number of linker types), which can be either of two types: A or B. Here the diversity of possible structures spans two dimensions: the ratio of A-linkers to B-linkers, and the number of possible arrangements of A and B linkers at a fixed ratio (see Fig. 16).

We can estimate a lower bound on the number of unique MOFs by the number of ratios of component types (i.e., a unit-cell with two A-linkers and one B-linker cannot be the same crystal as one with one A-linker and 2 B-linkers). Calculating this lower bound is equivalent to finding the number of unordered sets of M balls of L colors (the answer is: $M + L - 1$ choose $L - 1$). However, two crystals, both with two A-linkers and one B-linker but in different positions, can either be physically identical (i.e., related by a symmetry operation) or unique (for example, if the corner is asymmetrical as in Fig. 16). Thus, we can set an upper bound on the number of possible crystals by forming strings such as "BBA," "BAA," "BAB," and so forth. Thus, with a meager library of one corner and two linkers, the number, N, of possible MOFs is

$$\begin{gathered}\binom{M+L-1}{L-1} = \binom{3+2-1}{2-1} < N < 2^3 = L^{\mathrm{M}} \\ 4 < N < 8\end{gathered} \tag{1}$$

If we allow for more corners and linkers in our library (for example, $C = 10$, $L = 90$) but keep the constraint that MOFs may only use two linkers simultaneously, then we arrive at the modified expression

$$\begin{gathered}C \times L \times \left[\left(\frac{L-1}{2}\right)\binom{3+2-1}{2-1} - L + 2\right] < N < C \times L \times \left[\left(\frac{L-1}{2}\right)2^3 - L + 2\right] \\ 81,000 < N < 241,200\end{gathered} \tag{2}$$

This analysis underestimates the size of the decision tree because, among other things, conformational degrees of freedom and topological variations were

neglected (which would have a further multiplying effect on the possibility space). We have also neglected to consider decision sequences that are not equal, but that can lead to the same MOF structure (e.g., by changing the order in which two building blocks were added). Collapsing all such "degenerate" decision sequences into one is itself a computationally expensive procedure. Therefore it is possible that searching a larger decision tree with degeneracies is comparable to the computation cost of searching a reduced non-degenerate decision tree.

Finding plausible hypothetical MOFs in this decision tree can be carried out using a depth-first or breadth-first search along with established optimization techniques, such as various branch pruning methods. In our first implementation of this approach we used a depth-first search and implemented a fail-safe measure that would abort a branch of the tree and jump ahead by a random increment to another branch to prevent getting stuck. One should be cautious about the use of a random increment without precise control of the random number generator because it can hinder reproducibility of generated results.

Using the above approach, we were able to efficiently create a database of hypothetical MOFs that could be screened (or searched through) for applications in natural gas storage [57], xenon/krypton separations [83], and CO_2 separation and capture [84]. These investigations have largely served as proof-of-concept demonstrations of how a database of hypothetical MOFs can help find candidate MOFs for synthesis and also help elucidate structure–property relationships.

5 Large-Scale Screening of Hypothetical MOFs for Gas Storage and Separations

5.1 Motivations

Independent of the method, generating and screening hypothetical MOF structures computationally en masse serves two distinct purposes: helping to identify MOFs that can be synthesized and tested experimentally, and identifying structure–property relationships that can also reveal important physical limits on gas adsorption.

5.1.1 Identifying Promising Candidates for Experimental Synthesis

Generating hypothetical MOF structures to find promising candidates for experimental synthesis saves significant time and resources. There are already a few reported cases where a de novo MOF was designed, simulated, and later synthesized and found to have properties in nearly perfect agreement with the simulation predictions. Notably, Farha et al. demonstrated this approach for the MOF NU-100 [81], which had nearly the highest BET surface area for any porous material at the time of publication (6,143 m^2/g, second only to MOF-210, reported at almost the

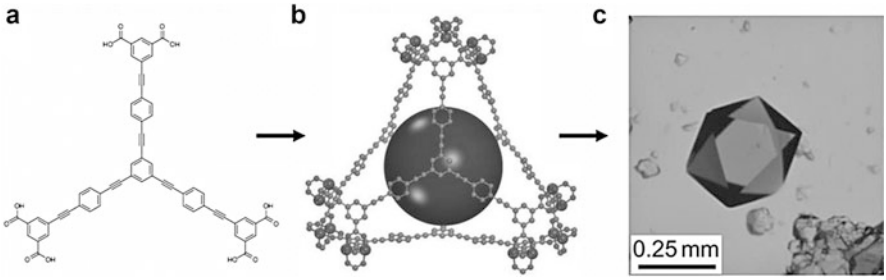


Fig. 17 A new organic linker (**a**) was designed and the resulting crystal, the MOF NU-100, (**b**) was predicted computationally. NU-100 was then synthesized (**c**), and the measured crystal structure and gas adsorption properties were in excellent agreement with the computational predictions. Figure adapted and reprinted with permission from [81]. Copyright 2010 Nature Publishing Group

same time [85]) (see Fig 17). We recently reported the synthesis of a MOF, which had been output by our bottom-up generator, whose crystal structure and methane adsorption agreed well with our computational predictions [57], but had been previously synthesized by Lin et al. under the name NOTT-107 [86]. In the future, we anticipate that more MOFs generated in silico will be subsequently synthesized experimentally, but it is worth expanding on the notion that significant insight can be obtained from large-scale screening even without going to the final synthesis step.

5.1.2 Discovering Performance Limits and Structure–property Relationships

A natural question when designing a new MOF for a particular application is what is the best possible performance outcome (e.g., what is the highest possible methane storage density at 35 bar and 298 K? [15, 16, 60]). This is a question for which generating and screening hypothetical MOFs on a large scale is ideal. However, it is difficult to know whether the best structure from any particular set of hypothetical MOFs represents the limit of the class of MOFs as a whole. Occasionally it is easier to address a question that appears at first more ambitious: what are the performance limits for any material whatsoever (i.e., beyond just MOFs)? When considering all physical arrangements of matter, any discovered performance limit will necessarily be an upper bound on what is possible to achieve with MOFs. This abstract notion is, in fact, an important consideration when creating a database of hypothetical MOFs. One can, for example, choose to use exotic building blocks to generate a more diverse database but at the risk of inadvertently creating hypothetical materials that are unlikely to be synthesized (or entirely unphysical). The benefit of a more diverse database is that it is more likely to span the full range of possible properties (including perhaps inadvertently unobtainable properties), which is helpful for determining fundamental performance limits (see Fig. 18).

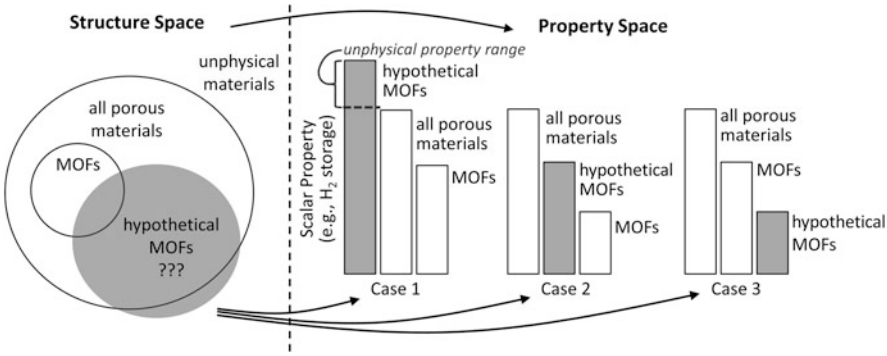


Fig. 18 A database of computationally generated materials might contain structures that correspond to *i*) MOFs that have already been or can be synthesized, (*ii*) physically plausible materials that are not MOFs (e.g., no inorganic building blocks), and (*iii*) unphysical structures (i.e., thermodynamically unstable, artifically raised binding interactions). In the process of mapping the properties of the hypothetical MOFs, three possible cases can result: (*1*) the hypothetical MOF with the best performance exceeds that of all physically realizable materials and thus reveals an upper bound on what is possible with real materials, (*2*) the span of property values of hypothetical MOFs contains all realizable MOFs, but a non-MOF material is potentially better, and (*3*) the properties exhibited by the hypothetical MOFs are a subset within the space of MOFs. It may not always be possible to know which case is most applicable in a given scenario, but it is a function of both the particular hypothetical MOF database and property being considered

Beyond discovering performance limits, the large data sets generated from simulating gas adsorption in thousands of hypothetical MOFs can be mined to find new structure–property relationships. Structure–property relationships are useful in establishing design rules for MOFs that can guide subsequent synthesis efforts. However, when experimental data from only a handful of MOFs is considered, it can be difficult to resolve the trends. As with the performance limits, it is not necessary for every hypothetical MOF to be a good candidate for chemical synthesis in order to be useful in identifying structure–property trends.

For both performance limits and structure–property relationships to yield accurate insights, not only must the set of hypothetical MOFs considered be large and diverse, but the model used to calculate gas adsorption must also be accurate. Below, we describe the commonly used classical atomistic simulation method to predict gas adsorption behavior in MOFs.

5.2 The Simulation Model

Gas adsorption properties of MOFs are usually calculated using Monte Carlo methods, which use random moves to sample a statistical mechanical ensemble and determine average quantities such as the equilibrium uptake and enthalpy [87, 88]. Typically the MOF framework is assumed to be rigid, but the gas

molecules can move and take on many possible configurations. For each configuration, the energy is calculated using a classical interaction potential. For simulations of CH_4, CO_2, and N_2 adsorption (pertinent to the examples described later in this chapter) the interaction energies between non-bonded atoms were computed through a Lennard-Jones (LJ) plus Coulomb potential:

$$v_{ij} = 4\epsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right) + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}$$

where i and j are interacting atoms, r_{ij} is the distance between atoms i and j, ϵ_{ij} and σ_{ij} are the LJ well depth and diameter, respectively, q_i and q_j are the partial charges of the interacting atoms, and ϵ_0 is the dielectric constant.

The LJ parameters for the gas molecules are often taken from the TraPPE force field, which stands for transferable potentials for phase equilibria [89, 90]. The TraPPE force field was developed to reproduce vapor–liquid coexistence curves for pure components of various classes of molecules. For framework atoms in MOFs, the LJ parameters are usually taken from rather general force fields, such as the Universal Force Field (UFF) [91]. Partial charges for the framework atoms have in the past (when only a handful of MOFs were being investigated) been derived from quantum chemistry calculations [46, 92]. For large databases it is impractical to use quantum chemistry-based methods, and so we applied a semi-empirical charge equilibration method that estimates partial charges based on known ionization energies [47, 93].

Adsorption isotherms are typically calculated using grand canonical Monte Carlo (GCMC) simulations. In this method an adsorbate phase at constant temperature T, volume V, and chemical potential μ is allowed to equilibrate with a gas phase (which is not simulated). The number of molecules N in the adsorbate phase is allowed to fluctuate so that the chemical potentials of the two phases are equal. For more details on our simulation method, the reader is referred to [35].

5.3 *Selected Gas Storage and Separations Applications*

5.3.1 Natural Gas Storage in Vehicles

Natural gas is an abundant fuel that can be used to power transportation vehicles [94]. It is less expensive and generates less CO_2 per mile travelled than liquid petroleum-based fuels [16]. However, it is a challenge to store natural gas in sufficient quantities in light-duty vehicles in a compact form. Compressed natural gas (CNG) has less than a third of the volumetric energy density of gasoline, and in the United States CNG tanks are pressurized to 3,600 psi, resulting in fuel tanks that are costly, heavy, and require capitally intensive compression equipment to refill [16, 95]. An alternative to storing natural gas at high pressures is to use a porous adsorbent that can store methane at similar concentrations to CNG tanks but at

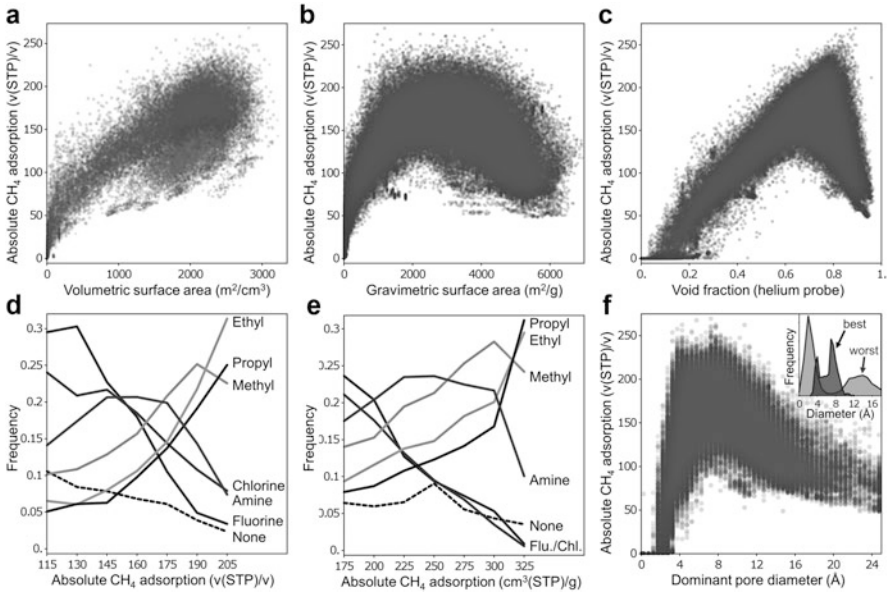


Fig. 19 Structure–property relationships obtained from the database of hypothetical MOFs. Reprinted with permission from [57]. Copyright 2012 Nature Publishing Group

reduced pressures. Over a decade ago, the Department of Energy put forth a methane storage target for adsorbent materials aimed at vehicular natural gas storage: 180 v(STP)/v at 35 bar and 298 K. This target has since been met, but given the high concentration of methane in CNG tanks (~260 v(STP)/v), there is significant interest in an even better adsorbent [16].

Both to find a MOF that could outperform existing adsorbents and to learn about the underlying structure–property relationships of methane storage in porous materials, we generated and screened a database of 137,953 hypothetical MOFs from 102 building blocks [57]. From this screening effort we found, in addition to a promising MOF for methane storage (that we subsequently synthesized), several revealing structure–property relationships (see Fig. 19). We found that volumetric methane storage density increased linearly with volumetric surface area, but that there was an optimal gravimetric surface area in the 2,000–3,000 m^2/g range. This latter observation ran counter to conventional wisdom that higher BET surface areas were always better for gas storage. We also found that the best MOFs in our database for storing methane at 35 bar shared a remarkably narrow range of void fraction values. The void fraction is the fraction of empty space within a porous material, and the highest densities of methane were found in MOFs that had a void fraction of almost exactly 80% (see Fig. 19c).

It was also possible to correlate the effects of choosing specific building blocks with methane storage ability. For example, we found that MOFs in our database with short alkyl functional groups (i.e., methyl, ethyl, and propyl groups) were found in over 75% of MOFs with methane storage capacities above 205 v(STP)/v

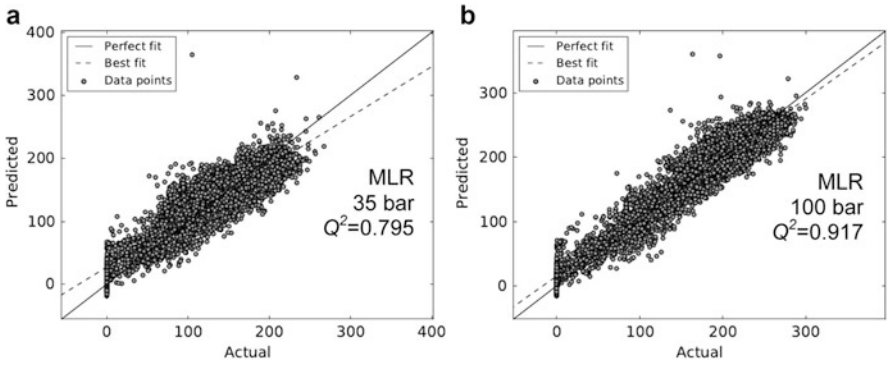


Fig. 20 A comparison of GCMC simulation data ("Actual") to data mining-based predictions using a multi-linear regression (MLR) fit ("Predicted"). The MLR fit used a training set of 10,000 MOFs to estimate methane storage at (**a**) 35 bar and (**b**) 100 bar on the remaining 127,953 MOFs based on three structure variables: the void fraction, dominant pore diameter, and gravimetric surface area. Reprinted with permission [69]. Copyright 2013 American Chemical Society

(see Fig. 19). This observation is relatively straightforward to test experimentally by incorporating specific functional groups into new MOFs that may or may not be in our generated database.

From the mountain of generated data, there are potentially many patterns within that may be hard to detect. Rather than merely plotting combinations of variables against each other and discovering patterns by visual inspection, rigorous data mining techniques can be used. Although data mining is common in large-scale drug and catalyst design, these methods have not yet been significantly applied to MOFs. This is likely because, until very recently, there were no databases of MOFs of suitable size.

Fernandez et al. recently used data mining techniques to investigate systematically our database of 137,953 hypothetical MOFs to determine the relative importance of various properties for predicting methane storage [69] (see Fig. 20). They applied a range of established data mining algorithms (e.g., multi-linear regression, decision trees, supported vector machines) to see which were the most suitable for extracting trends in MOF-based gas adsorption data.

Fernandez et al. found that the methane storage at 35 bar, $U^{35\text{bar}}$, or at 100 bar, $U^{100\text{bar}}$, could be estimated reasonably well by knowledge of just the void fraction (VF), dominant pore diameter (DP, which is the diameter that corresponds to the tallest peak in a pore size distribution of a porous material), and gravimetric surface area (Sg):

$$U^{35\text{bar}} = 391.6180 \times VF - 9.3361 \times DP - 0.0161 \times Sg + 1.4954$$
$$U^{100\text{bar}} = 390.9582 \times VF - 6.1908 \times DP - 0.0044 \times Sg - 3.2607$$
$$U^{35,100\text{bar}}[=]\ v(STP)/v,\ \ VF\ [=]\ \text{dimensionless},\ \ DP\ [=]\ \text{Å},\ \ Sg\ [=]\ m^2/g$$

Using supported vector machines, they were able to identify combinations of parameters that suggest there are MOFs with greater methane storage capacities than those present in the database itself.

This preliminary illustration indicates that data mining is likely to become commonplace in future MOF research. Such systematic methods will be particularly helpful for problems where computational simulations are very costly, such as in MOF-based catalysis or low pressure CO_2 capture where water and other trace gases can play a significant role.

5.3.2 Carbon Dioxide Separation and Capture

Due to both rising global greenhouse gas emissions [96] and an increased worldwide demand for natural gas [97], there is significant interest in the development of porous materials to separate carbon dioxide (CO_2) from mixtures of gases, such as the exhaust of fossil-fuel-based power plants (flue gas) and gases that are rich sources of methane (CH_4).

Porous materials like MOFs can be used to separate CO_2 from these mixtures via pressure-swing adsorption (PSA) or vacuum-swing adsorption (VSA), where the material is exposed to impure gas at a high(er) pressure and then regenerated by lowering (i.e., releasing or "swinging") the pressure. The effectiveness of a MOF for either PSA or VSA depends on how well it adsorbs CO_2 at the higher pressure and then how easily it releases the CO_2 at the lower pressure.

Creating a MOF that is optimal for a CO_2 separation process requires that we are able to synthesize a structure with pores that selectively bind CO_2 either much more strongly, or more weakly, than other gases in the mixture. This, in turn, requires that we determine (independent of our synthesis capabilities) what the optimal shape, size, and chemistry of the pores *ought* to be for CO_2 separation. To address this need we used the same database of 137,953 hypothetical MOFs as was used for methane storage screening, but instead ran molecular simulations of CO_2 and N_2 adsorption (as well as CH_4 adsorption, but at lower pressures than in the earlier work that focused on compressed methane storage in vehicles). The objective was solely to determine structure–property relationships, which were unclear at the time, rather than find a particular MOF candidate to synthesize [84].

In this study we considered every MOF in the database in four distinct cases corresponding to separating CO_2 from either N_2 or CH_4 at pressures and compositions selected for their industrial relevance, namely: (1) natural gas purification using PSA, (2) landfill gas separation using PSA, (3) landfill gas separation using VSA, and (4) flue gas separation using VSA. See Table 1 for gas phase mixture compositions and pressures that approximate each of these four cases (at temperatures of 298 K).

With this large dataset we were able to observe sharply defined correlations between the properties of the MOFs, such as the pore diameter, surface area, pore volume, and chemical functionality, and their usefulness for CO_2 separations in each of the four cases (see Fig. 21). As with the case of high pressure methane storage, we observed what could be described as structure–property domains whose boundaries arise from either the limited diversity of our database or from fundamental physical limits (as discussed earlier).

Table 1 Evaluation criteria used by Bae and Snurr to assess the effectiveness of porous materials for CO_2 separation and capture [68]

Case	Application	Mixture composition	Adsorption and desorption pressures (p^{ads} and p^{des})
1)	Natural gas purification using PSA	$CO_2/CH_4 = 10{:}90$	$p^{ads} = 5$ bar, $p^{des} = 1$ bar
2)	Landfill gas separation using PSA	$CO_2/CH_4 = 50{:}50$	$p^{ads} = 5$ bar, $p^{des} = 1$ bar
3)	Landfill gas separation using VSA	$CO_2/CH_4 = 50{:}50$	$p^{ads} = 1$ bar, $p^{des} = 0.1$ bar
4)	Flue gas separation using VSA	$CO_2/N_2 = 10{:}90$	$p^{ads} = 1$ bar, $p^{des} = 0.1$ bar

The four mixture compositions and adsorption/desorption conditions considered are for: (1) natural gas purification using PSA, (2) landfill gas separation using PSA, (3) landfill gas separation using VSA, and (4) flue gas separation using VSA. Temperature is 298 K in all cases

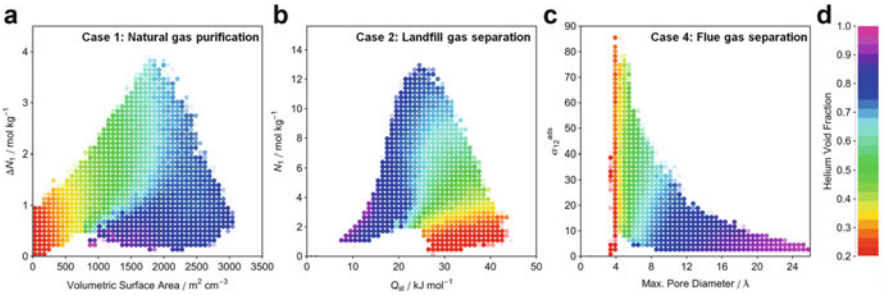


Fig. 21 A sample of structure–property relationships derived from simulated CO_2, CH_4, and N_2 adsorption in over 130,000 hypothetical MOFs. Clear relationships can be discerned between **(a)** CO_2 working capacity (ΔN_1) and surface area, **(b)** CO_2 uptake (N_1) at 2.5 bar and CO_2 heat of adsorption (Q_{st}), and **(c)** selectivity of CO_2 over N_2 ($\alpha_{12}{}^{ads}$) and maximum pore diameter. Q_{st} values are determined from CO_2 adsorption at the lowest simulated pressure, 0.01 bar. Each plot is divided into 50 × 50 regions that are represented by a *filled circle* if more than 25 structures exist within the region. The *color* of each circle represents the average **(d)** helium void fraction of all structures in that plot region. Figure obtained from [84] and reprinted with permission from The Royal Society of Chemistry

Our work complements several recently reported large-scale computational screening efforts focused on CO_2 separations [98–102]. Lin et al. [100] screened hundreds of thousands of hypothetical zeolite and zeolitic imidazolate framework structures for their application to CO_2 capture from flue gas [100]. In this study, each structure was measured on its ability to reduce the "parasitic load," which is the amount of energy a fossil fuel-based power plant would need to spend on capturing CO_2 (instead of delivering electrical power) (see Fig. 22). Similarly, Haldoupis et al. [98] computationally screened ~500 MOFs for their ability to separate CO_2 from N_2, which was the largest set of predictions for CO_2 adsorption in MOFs at the time it was reported. In their work, Henry's constants were correlated with pore diameters, but similar comparisons with other structural characteristics (e.g., surface area, void fraction) or other adsorption properties (e.g., working capacity, selectivity) were not reported. Wu et al. [102] recently

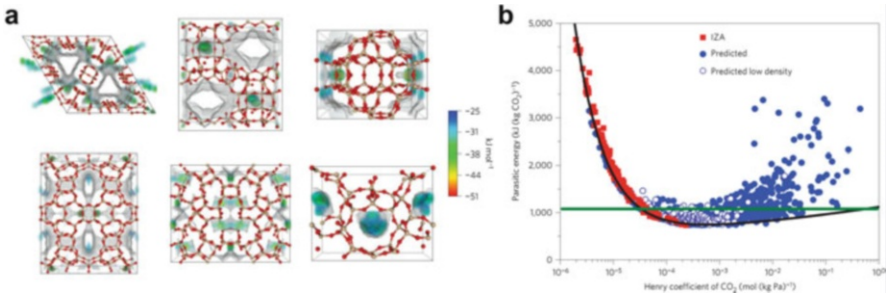


Fig. 22 (**a**) Six example hypothetical zeolites that Lin et al. [100] found performed well for CO_2 capture. Materials shown as ball and stick (O,*red*; Si,*tan*), with colored surfaces indicating the local free energies of binding CO_2. An important focus in this screening study was the (**b**) parasitic energy that each porous material could potentially reduce. The *green line* indicates the parasitic energy load of existing monoethanolamine CO_2 capture technology and the *black line* indicates a minimum parasitic energy threshold. Reprinted with permission from [100]. Copyright 2012 Nature Publishing Group

examined 105 MOFs for CO_2/N_2 separations and discovered that simultaneously increasing Q_{st} values while decreasing the void fraction was a useful design rule for increasing the selectivity.

6 Conclusions

Large-scale computational methods have great potential to accelerate the development of new materials. MOFs provide a potentially ideal platform for applying computational crystal engineering methods due to their predictable structures *and* predictable gas adsorption behaviors. We have shown how large-scale simulations can reveal structure–property insights in MOFs for important gas storage and separation applications such as natural gas storage in vehicles and CO_2 separation and capture.

In the future we will likely see experimentally synthesized MOFs that were designed and optimized entirely in silico. This will depend, however, not only on the ability to generate hypothetical MOFs that can be readily synthesized but also on further improvements in our simulation models used to predict gas adsorption behavior.

It is also exciting to consider the possibility of high throughput computational methods working in tandem with high throughput robotic synthesis equipment. This would enable researchers to spend less time attempting to discover promising new materials serendipitously and more time testing material hypotheses and creating the next generation of MOFs.

Acknowledgements The authors thank Profs. Omar Farha and Joseph Hupp for stimulating discussions and the Defense Threat Reduction Agency (HDTRA – 10 – 1 – 0023) for financial support.

References

1. Kitagawa S, Kitaura R, Noro S (2004) Functional porous coordination polymers. Angew Chem Int Ed 43:2334
2. Zhou H-C, Long JR, Yaghi OM (2012) Introduction to metal-organic frameworks. Chem Rev 112:673
3. Yaghi OM, Li G, Li H (1995) Selective binding and removal of guests in a microporous metal-organic framework. Nature 378:703
4. Li H, Eddaoudi M, O'Keeffe M, Yaghi OM (1999) Design and synthesis of an exceptionally stable and highly porous metal-organic framework. Nature 402:276
5. Subramanian S, Zaworotko MJ (1995) Porous solids by design: $[Zn(4,4'\text{-bpy})_2(SiF_6)]_n \cdot xDMF$, a single framework octahedral coordination polymer with large square channels. Angew Chem Int Ed Engl 34:2127
6. Robson R (2000) A net-based approach to coordination polymers. J Chem Soc Dalton Trans 3735
7. Spokoyny AM, Kim D, Sumrein A, Mirkin CA (2009) Infinite coordination polymer nano- and microparticle structures. Chem Soc Rev 38:1218
8. Kitaura R, Seki K, Akiyama G, Kitagawa S (2003) Porous coordination-polymer crystals with gated channels specific for supercritical gases. Angew Chem Int Ed 42:428
9. Uemura T, Yanai N, Kitagawa S (2009) Polymerization reactions in porous coordination polymers. Chem Soc Rev 38:1228
10. Chui SS-Y, Lo SM-F, Charmant JPH, Orpen AG, Williams ID (1999) A chemically functionalizable nanoporous material $[Cu_3(TMA)_2(H_2O)_3]n$. Science 283:1148
11. Noro S, Kitagawa S, Kondo M, Seki K (2000) A new, methane adsorbent, porous coordination polymer $[\{CuSiF_6(4,4'\text{-bipyridine})_2\}_n]$. Angew Chem Int Ed 39:2081
12. Czaja AU, Trukhan N, Müller U (2009) Industrial applications of metal-organic frameworks. Chem Soc Rev 38:1284
13. Farha OK, Eryazici I, Jeong NC, Hauser BG, Wilmer CE, Sarjeant AA, Nguyen ST, Snurr RQ, Yazaydin AÖ, Hupp JT (2012) Metal-organic framework materials with ultrahigh surface areas: is the sky the limit? J Am Chem Soc 134:15016
14. Millward AR, Yaghi OM (2005) Metal-organic frameworks with exceptionally high capacity for storage of carbon dioxide at room temperature. J Am Chem Soc 127:17998
15. Düren T, Sarkisov L, Yaghi OM, Snurr RQ (2004) Design of new materials for methane storage. Langmuir 20:2683
16. Wilmer CE, Farha OK, Krungleviciute V, Eryazici I, Sarjeant AA, Yildirim T, Snurr RQ, Hupp JT (2013) Gram-scale, high-yield synthesis of a robust metal-organic framework for methane storage. Energy Environ Sci 6:1158
17. Rosi NL, Eckert J, Eddaoudi M, Vodak DT, Kim J, O'Keeffe M, Yaghi OM (2003) Hydrogen storage in microporous metal-organic frameworks. Science 300:1127
18. Murray L, Dinca M, Long J (2009) Hydrogen storage in metal-organic frameworks. Chem Soc Rev 38:1294
19. Li J-R, Kuppler RJ, Zhou H-C (2009) Selective gas adsorption and separation in metal-organic frameworks. Chem Soc Rev 38:1477
20. Li J-R, Sculley J, Zhou H-C (2012) Metal-organic frameworks for separations. Chem Rev 112:869

21. Guo H, Zhu G, Hewitt IJ, Qiu S (2009) "Twin Copper Source" growth of metal-organic framework membrane: $Cu_3(BTC)_2$ with high permeability and selectivity for recycling H_2. J Am Chem Soc 131:1646
22. Finsy V, Ma L, Alaerts L, De Vos DE, Baron GV, Denayer JFM (2009) Separation of CO_2/CH_4 mixtures with the MIL-53(Al) metal-organic framework. Microporous Mesoporous Mater 120:221
23. Maes M, Alaerts L, Vermoortele F, Ameloot R, Couck S, Finsy V, Denayer JFM, De Vos DE (2010) Separation of C5-hydrocarbons on microporous materials: complementary performance of MOFs and zeolites. J Am Chem Soc 132:2284
24. Nicolau MPM, Bárcia PS, Gallegos JM, Silva JAC, Rodrigues AE, Chen B (2009) Single- and multicomponent vapor-phase adsorption of xylene isomers and ethylbenzene in a microporous metal-organic framework. J Phys Chem C 113:13173
25. Gu Z-Y, Jiang D-Q, Wang H-F, Cui X-Y, Yan X-P (2010) Adsorption and separation of xylene isomers and ethylbenzene on two Zn–terephthalate metal−organic frameworks. J Phys Chem C 114:311
26. Seo J, Whang D, Lee H, Jun S, Oh J, Jeon Y, Kim K (2000) A homochiral metal-organic porous material for enantioselective separation and catalysis. Nature 404:982
27. Bradshaw D, Prior TJ, Cussen EJ, Claridge JB, Rosseinsky MJ (2004) Permanent microporosity and enantioselective sorption in a chiral open framework. J Am Chem Soc 126:6106
28. Düren T, Snurr RQ (2004) Assessment of isoreticular metal-organic frameworks for adsorption separations: a molecular simulation study of methane/n-butane mixtures. J Phys Chem B 108:15703
29. Watanabe T, Keskin S, Nair S, Sholl DS (2009) Computational identification of a metal organic framework for high selectivity membrane-based CO_2/CH_4 separations: Cu(hfipbb)$(H_2hfipbb)_{0.5}$. Phys Chem Chem Phys 11:11389
30. Liu B, Yang Q, Xue C, Zhong C, Chen B, Smit B (2008) Enhanced adsorption selectivity of hydrogen/methane mixtures in metal-organic frameworks with interpenetration: a molecular simulation study. J Phys Chem C 112:9854
31. Yaghi OM, O'Keeffe M, Ockwig NW, Chae HK, Eddaoudi M, Kim J (2003) Reticular synthesis and the design of new materials. Nature 423:705
32. Ockwig NW, Delgado Friedrichs O, O'Keeffe M, Yaghi OM (2005) Reticular chemistry: occurrence and taxonomy of nets and grammar for the design of frameworks. Acc Chem Res 38:176
33. Banerjee R, Phan A, Wang B, Knobler C, Furukawa H, O'Keeffe M, Yaghi OM (2008) High-throughput synthesis of zeolitic imidazolate frameworks and application to CO_2 capture. Science 319:939
34. Sumida K, Horike S, Kaye SS, Herm ZR, Queen WL, Brown CM, Grandjean F, Long GJ, Dailly A, Long JR (2010) Hydrogen storage and carbon dioxide capture in an iron-based sodalite-type metal-organic framework (Fe-BTT) discovered via high-throughput methods. Chem Sci 1:184
35. Getman RB, Bae Y-S, Wilmer CE, Snurr RQ (2011) Review and analysis of molecular simulations of methane, hydrogen, and acetylene storage in metal-organic frameworks. Chem Rev 112:703
36. Snurr RQ, Yazaydin AO, Dubbeldam D, Frost H (2010) In: MacGillivray LR (ed) Metal–organic frameworks: design and application. Wiley, Hoboken, p 313
37. Walton KS, Millward AR, Dubbeldam D, Frost H, Low JJ, Yaghi OM, Snurr RQ (2008) Understanding inflections and steps in carbon dioxide adsorption isotherms in metal-organic frameworks. J Am Chem Soc 130:406
38. Peng Y, Srinivas G, Wilmer CE, Eryazici I, Snurr RQ, Hupp JT, Yildirim T, Farha OK (2013) Simultaneously high gravimetric and volumetric methane uptake characteristics of the metal-organic framework NU-111. Chem Commun 49:2992
39. Sese L (1995) Feynman–Hibbs potentials and path integrals for quantum Lennard–Jones systems: theory and Monte Carlo simulations. Mol Phys 85:931

40. Wang Q, Johnson JK (1999) Molecular simulation of hydrogen adsorption in single-walled carbon nanotubes and idealized carbon slit pores. J Chem Phys 110:577
41. Farha OK, Wilmer CE, Eryazici I, Hauser BG, Parilla PA, O'Neill K, Sarjeant AA, Nguyen ST, Snurr RQ, Hupp JT (2012) Designing higher surface area metal-organic frameworks: are triple bonds better than phenyls? J Am Chem Soc 134:9860
42. Frost H, Snurr RQ (2007) Design requirements for metal-organic frameworks as hydrogen storage materials. J Phys Chem C 111:18794
43. Muller E, Rull L, Vega L, Gubbins K (1996) Adsorption of water on activated carbons: a molecular simulation study. J Phys Chem 100:1189
44. Ramachandran CE, Chempath S, Broadbelt LJ, Snurr RQ (2006) Water adsorption in hydrophobic nanopores: Monte Carlo simulations of water in silicalite. Microporous Mesoporous Mater 90:293
45. Paranthaman S, Coudert F-X, Fuchs AH (2010) Water adsorption in hydrophobic MOF channels. Phys Chem Chem Phys 12:8124
46. Yazaydin AÖ et al (2009) Screening of metal-organic frameworks for carbon dioxide capture from flue gas using a combined experimental and modeling approach. J Am Chem Soc 131:18198
47. Wilmer CE, Kim K-C, Snurr RQ (2012) An extended charge equilibration method. J Phys Chem Lett 3:2506
48. Xiang SC, Zhou W, Zhang ZJ, Green MA, Liu Y, Chen BL (2010) Open metal sites within isostructural metal-organic frameworks for differential recognition of acetylene and extraordinarily high acetylene storage capacity at room temperature. Angew Chem Int Ed 49:4615
49. Chen B, Ockwig NW, Millward AR, Contreras DS, Yaghi OM (2005) High H_2 adsorption in a microporous metal-organic framework with open metal sites. Angew Chem 117:4823
50. Getman RB, Miller JH, Wang K, Snurr RQ (2011) Metal alkoxide functionalization in metal−organic frameworks for enhanced ambient-temperature hydrogen storage. J Phys Chem C 115:2066
51. Haldoupis E, Nair S, Sholl DS (2010) Efficient calculation of diffusion limitations in metal organic framework materials: a tool for identifying materials for kinetic separations. J Am Chem Soc 7258
52. Eddaoudi M, Kim J, Rosi N, Vodak D, Wachter J, O'Keeffe M, Yaghi OM (2002) Systematic design of pore size and functionality in isoreticular MOFs and their application in methane storage. Science 295:469
53. Zhao D, Timmons DJ, Yuan D, Zhou H-C (2011) Tuning the topology and functionality of metal-organic frameworks by ligand design. Accounts Chem Res 44:123
54. Wang Z, Cohen SM (2009) Postsynthetic modification of metal-organic frameworks. Chem Soc Rev 38:1315
55. Karagiaridi O, Bury W, Sarjeant AA, Stern CL, Farha OK, Hupp JT (2012) Synthesis and characterization of isostructural cadmium zeolitic imidazolate frameworks via solvent-assisted linker exchange. Chem Sci 3:3256
56. Takaishi S, DeMarco EJ, Pellin MJ, Farha OK, Hupp JT (2013) Solvent-assisted linker exchange (SALE) and post-assembly metallation in porphyrinic metal-organic framework materials. Chem Sci 4:1509
57. Wilmer CE, Leaf M, Lee C- Y, Farha OK, Hauser BG, Hupp JT, Snurr RQ (2012) Large-scale screening of hypothetical metal-organic frameworks. Nat Chem 4:83
58. Cook TR, Zheng Y-R, Stang PJ (2013) Metal-organic frameworks and self-assembled supramolecular coordination complexes: comparing and contrasting the design, synthesis, and functionality of metal-organic materials. Chem Rev 113:734
59. Gould SL, Tranchemontagne D, Yaghi OM, Garcia-Garibay MA (2008) Amphidynamic character of crystalline MOF-5: rotational dynamics of terephthalate phenylenes in a free-volume, sterically unhindered environment. J Am Chem Soc 130:3246

60. Ma S, Sun D, Simmons JM, Collier CD, Yuan D, Zhou HC (2008) Metal-organic framework from an anthracene derivative containing nanoscopic cages exhibiting high methane uptake. J Am Chem Soc 130:1012
61. Amirjalayer S, Schmid R (2008) Conformational isomerism in the isoreticular metal organic framework family: a force field investigation. J Phys Chem C 112:14980
62. Rowsell JLC, Yaghi OM (2006) Effects of functionalization, catenation, and variation of the metal oxide and organic linking units on the low-pressure hydrogen adsorption properties of metal-organic frameworks. J Am Chem Soc 128:1304
63. Farha OK, Malliakas CD, Kanatzidis MG, Hupp JT (2010) Control over catenation in metal-organic frameworks via rational design of the organic building block. J Am Chem Soc 132:950
64. Ryan P, Broadbelt LJ, Snurr RQ (2008) Is catenation beneficial for hydrogen storage in metal-organic frameworks? Chem Commun 4132
65. Eddaoudi M, Moler DB, Li H, Chen B, Reineke TM, O'Keeffe M, Yaghi OM (2001) Modular chemistry: secondary building units as a basis for the design of highly porous and robust metal-organic carboxylate frameworks. Accounts Chem Res 34:319
66. Tranchemontagne DJ, Mendoza-Cortés JL, O'Keeffe M, Yaghi OM (2009) Secondary building units, nets and bonding in the chemistry of metal-organic frameworks. Chem Soc Rev 38:1257
67. Deng H, Doonan CJ, Furukawa H, Ferreira RB, Towne J, Knobler CB, Wang B, Yaghi OM (2010) Multiple functional groups of varying ratios in metal-organic frameworks. Science 327:846
68. Bae Y-S, Snurr RQ (2011) Development and evaluation of porous materials for carbon dioxide separation and capture. Angew Chem Int Ed 50:11586
69. Fernandez M, Woo TK, Wilmer CE, Snurr RQ (2013) Large-scale quantitative structure–property relationship (QSPR) analysis of methane storage in metal-organic frameworks. J Phys Chem C 117:7681
70. Bauer S, Serre C, Devic T, Horcajada P, Marrot J, Férey G, Stock N (2008) High-throughput assisted rationalization of the formation of metal organic frameworks in the iron(III) aminoterephthalate solvothermal system. Inorg Chem 47:7568
71. Mellot Draznieks C, Newsam JM, Gorman AM, Freeman CM, Férey G (2000) De novo prediction of inorganic structures developed through automated assembly of secondary building units (AASBU method). Angew Chem Int Ed 39:2270
72. Kirkpatrick S, Gelatt C, Vechhi M (1983) Optimization by simulated annealing. Science 220:671
73. Delgado Friedrichs O, Dress AWM, Huson DH, Klinowski J, Mackay AL (1999) Systematic enumeration of crystalline networks. Nature 400:644
74. Wells AF (1977) Three dimensional nets and polyhedra. Wiley, New York
75. O'Keeffe M, Peskov MA, Ramsden SJ, Yaghi OM (2008) The Reticular Chemistry Structure Resource (RCSR) database of, and symbols for, crystal nets. Accounts Chem Res 41:1782
76. Hyde ST, Delgado Friedrichs O, Ramsden SJ, Robins V (2006) Towards enumeration of crystalline frameworks: the 2D hyperbolic approach. Solid State Sci 8:740
77. Ramsden SJ, Robins V, Hyde ST (2009) Three-dimensional Euclidean nets from two-dimensional hyperbolic tilings: kaleidoscopic examples. Acta Crystallogr A 65:81
78. McColm GL, Clark WE, Eddaoudi M, Wojtas L, Zaworotko M (2011) Crystal engineering using a "Turtlebug" algorithm: a de novo approach to the design of binodal metal-organic frameworks. Cryst Growth Des 11:3686
79. Thomas JM, Klinowski J (2007) Systematic enumeration of microporous solids: towards designer catalysts. Angew Chem Int Ed 46:7160
80. Bureekaew S, Schmid R (2013) Hypothetical 3D-periodic covalent organic frameworks: exploring the possibilities by a first principles derived force field. CrystEngComm 15:1551

81. Farha OK, Yazaydın AÖ, Eryazici I, Malliakas CD, Hauser BG, Kanatzidis MG, Nguyen ST, Snurr RQ, Hupp JT (2010) De novo synthesis of a metal-organic framework material featuring ultrahigh surface area and gas storage capacities. Nat Chem 2:944
82. Chakrabarty R, Mukherjee PS, Stang PJ (2011) Supramolecular coordination: self-assembly of finite two- and three-dimensional ensembles. Chem Rev 111:6810
83. Sikora BJ, Wilmer CE, Greenfield ML, Snurr RQ (2012) Thermodynamic analysis of Xe/Kr selectivity in over 137000 hypothetical metal-organic frameworks. Chem Sci 3:2217
84. Wilmer CE, Farha OK, Bae Y-S, Hupp JT, Snurr RQ (2012) Structure–property relationships of porous materials for carbon dioxide separation and capture. Energy Environ Sci 5:9849
85. Furukawa H, Ko N, Go YB, Aratani N, Choi SB, Choi E, Yazaydin AÖ, Snurr RQ, O'Keeffe M, Kim J, Yaghi OM (2010) Ultrahigh porosity in metal-organic frameworks. Science 329:424
86. Lin X et al (2009) High capacity hydrogen adsorption in Cu(II) tetracarboxylate framework materials: the role of pore size, ligand functionalization, and exposed metal sites. J Am Chem Soc 131:2159
87. Metropolis N, Ulam S (1949) The Monte Carlo method. J Am Stat Assoc 44:335
88. Leach AR (2001) Molecular modelling: principles and applications. Prentice Hall, Upper Saddle River
89. Martin MG, Siepmann JI (1998) Transferable potentials for phase equilibria. 1. United-atom description of N-alkanes. J Phys Chem B 102:2569
90. Martin MG, Siepmann JI (1999) Novel configurational-bias Monte Carlo method for branched molecules. Transferable potentials for phase equilibria. 2. United-atom description of branched alkanes. J Phys Chem B 103:4508
91. Rappé AK, Casewit CJ, Colwell KS, Goddard WA III, Skiff WM (1992) UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. J Am Chem Soc 114:10024
92. Breneman CM, Wiberg KB (1990) Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis. J Comput Chem 11:361
93. Rappé AK, Goddard WA III (1991) Charge equilibration for molecular dynamics simulations. J Phys Chem 95:3358
94. Flynn PC (2002) Commercializing an alternate vehicle fuel: lessons learned from natural gas for vehicles. Energy Policy 30:613
95. Brown RN (2005) Compressors: selection and sizing. Elsevier, Oxford
96. Chu S (2009) Carbon capture and sequestration. Science 325:1599
97. Tagliabue M, Farrusseng D, Valencia S, Aguado S, Ravon U, Rizzo C, Corma A, Mirodatos C (2009) Natural gas treating by selective adsorption: material science and chemical engineering interplay. Chem Eng J 155:553
98. Haldoupis E, Nair S, Sholl DS (2012) Finding MOFs for highly selective CO_2/N_2 adsorption using materials screening based on efficient assignment of atomic point charges. J Am Chem Soc 134:4313
99. Wu D, Wang C, Liu B, Liu D, Yang Q, Zhong C (2012) Large-scale computational screening of metal-organic frameworks for CH_4/H_2 separation. Aiche J 58:2078
100. Lin L-C, Berger AH, Martin RL, Kim J, Swisher JA, Jariwala K, Rycroft CH, Bhown AS, Deem MW, Haranczyk M, Smit B (2012) In silico screening of carbon-capture materials. Nat Mater 11:633
101. Krishna R, van Baten JM (2011) In silico screening of metal-organic frameworks in separation applications. Phys Chem Chem Phys 13:10593
102. Wu D, Yang Q, Zhong C, Liu D, Huang H, Zhang W, Maurin G (2012) Revealing the structure–property relationships of metal-organic frameworks for CO_2 capture from flue gas. Langmuir 28:12094

Index