Database-Driven High-Throughput Calculations and Machine Learning Models for Materials Design

17

Rickard Armiento

Abstract

This chapter reviews past and ongoing efforts in using high-throughput ab-initio calculations in combination with machine learning models for materials design. The primary focus is on bulk materials, i.e., materials with fixed, ordered, crystal structures, although the methods naturally extend into more complicated configurations. Efficient and robust computational methods, computational power, and reliable methods for automated database-driven high-throughput computation are combined to produce high-quality data sets. This data can be used to train machine learning models for predicting the stability of bulk materials and their properties. The underlying computational methods and the tools for automated calculations are discussed in some detail. Various machine learning models and, in particular, descriptors for general use in materials design are also covered.

17.1 Background

Design of new materials with desired properties is a crucial step in making many innovative technologies viable. The aim is to find materials that fulfill requirements on efficiency, cost, environmental impact, length of life, safety, and other properties. During the past decades, we have seen major progress in theoretical materials science due to the combination of improved computational methods and a massive increase in available computational power. It is now standard practice to obtain insights into the physics of materials by using supercomputers to find numerical solutions to the basic equations of quantum mechanics. When using the appropriate

R. Armiento (✉)
Department of Physics, Chemistry and Biology, Linköping University, Linköping, Sweden
e-mail: rickard.armiento@liu.se

K. T. Schütt et al. (eds.), *Machine Learning Meets Quantum Physics*,
Lecture Notes in Physics 968, https://doi.org/10.1007/978-3-030-40245-7_17

level of theory, these calculations can be robust enough to run in unsupervised high-throughput. Hence, materials design can be done via automated theoretical screening of candidate materials and substances, picking out those with desired properties. Early examples of this methodology include works in the fields of catalysts [1], battery materials [2, 3], detector materials for ionizing radiation [4], superconductivity [5], thermoelectricity [6, 7], piezoelectrics [8, 9], transparent conducting oxides [10], and two-dimensional materials [11]. There is a wealth of further examples in the literature, see, e.g., the reviews in Ref. [12–15].

Early adoption of high-throughput methodology for materials design has invoked the ambition that it may be possible to computationally predict the existence and basic properties of essentially *every single material*, i.e., any composition that, in principle, can be synthesized as a reasonably long-lived "stable" compound (in the context of an environment.) This ambition has been referred to as the *materials genome project* [13,16,17], which in 2011 was endorsed as a White House initiative [18]. The idea is that access to materials genome data with sufficient coverage would greatly accelerate materials design. It would be possible to perform queries against this data to pick out compositions that have some sought combination of desired properties for a specific application at, essentially, no additional computational cost [12, 13].

A large number of databases of materials-genome-type are now available, many of them open and free for access over the Internet. Some notable examples include: the *Electronic Structure Project* (http://gurka.physics.uu.se/esp/; 2002), the *Automatic FLOW repository* (aflowlib.org; 2011), *the Materials Project* (materialsproject.org; 2011), the *Open Materials Database* (openmaterialsdb.se; 2013), the *Open Quantum Materials Database* (oqmd.org; 2013), the *Theoretical Crystallographic Open Database* (www.crystallography.net/tcod; 2013), the *Novel Materials Discovery Repository* (nomad-repository.eu; 2014), the *High Performance Computing Center Materials Database—NREL MatDb* (materials.nrel.gov; 2015), and the *Materials Cloud* (materialscloud.org; 2017).

To use machine learning models for, e.g., molecular dynamics simulations of systems with up to a few chemical species has become increasingly popular (i.e., to accelerate simulations of the movement of some types of atoms in a material.) To train more general models with data from materials genome-type databases opens a way forward towards the vision of a complete coverage of materials and their properties. This chapter reviews the use of high-throughput techniques and tools to produce training data for these models and recent developments in the area of models with the aim of a general description of atomistic systems (i.e., molecules and materials.) This development is, at its core, the adoption of an informatics perspective to materials science and design, which has been referred to as *materials informatics* [19, 20].

It has been posed as a hypothesis that the progress of general AI methods will eventually reach "the singularity," a moment in time when self-improving AI methods set off a runaway technological development that fundamentally changes society.[1] One can, in a similar way, formulate the hypothesis that the development of increasingly sophisticated machine learning models for atomistic systems will reach a singularity of its own, i.e., a point in time of fundamental change in our theoretical description of physical matter. This change would happen when fully trained, general, machine learning models appear that are capable of predictions at the same accuracy as physics-based quantum mechanical simulations but at negligible computational effort. The result would turn the present materials genome-type databases obsolete and enable true inverse design of molecules and materials with desired properties across the full chemical space at near zero computational expense. Such a development would bring far-reaching changes across the natural sciences.

In conclusion, advancing the present state of materials design with machine learning models requires progress in three key areas: (1) progress in the theory and methods used in physics-based calculations that can be used to improve the quality of training data. This requires developing methods with improved accuracy without sacrificing the low computational demand and the high level of generality that are necessary for the methods to be useful for high-throughput calculations; (2) further improved methods and tools for running automated calculations in high throughput. While there are many software packages and solutions available today for running calculations in high-throughput, major work of both practical and theoretical nature remains to turn methods that were developed and tested only on a few systems into automated workflows capable of running unsupervised at large scale without human interference; and (3) further improved machine learning models for general atomistic systems.

17.2 Computational Methods

Kohn–Sham density functional theory [23, 24] (KS-DFT) is the standard theoretical framework for high-throughput computation in present materials property databases. There is a range of software implementations for performing the numerical solution of the basic equations of DFT. A few prominent examples include the *Vienna Ab-initio Simulation Package—VASP* (vasp.at), *ABINIT* (www.abinit.org), *Wien2K* (susi.theochem.tuwien.ac.at), and *Quantum ESPRESSO* (www.quantum-espresso.org). Of primary concern for these software packages is the numerical convergence towards the exact solution with respect to the approximations used. Most approximations are fairly straightforward to systematically improve towards a converged result, which has led to a number of standard practices for setting convergence parameters that are typically documented in relation to the respective database. See, e.g., Ref. [17] for the practices used in the Materials Project database.

[1] The term was recently popularized by a 2006 book by Kurzweil [21], but its use goes back to a 1958 account by Stanislaw Ulam of a discussion with John von Neumann that references a point in time of fundamental change due to runaway technological development [22].

One aspect of numerical convergence that frequently is in focus when discussing the accuracy of KS-DFT calculations in the context of chemistry-oriented calculations is the basis set used to represent the single particle wave functions (also known as the *KS orbitals.*) While more or less all basis sets can be systematically extended towards numerical convergence, this can be impractical for some choices. Nevertheless, in the context of materials design of bulk materials, we are mostly concerned with fully periodic crystals where the most common choice is a plane-wave basis set where systematic convergence is more straightforward.

In contrast to the numerical approximations that can be, at least in principle, systematically refined to arbitrary accuracy towards the solution of the KS-DFT equations, there is one aspect of the calculations where this is not possible. This is the choice of exchange-correlation density functional. This choice is crucial for the description of the physics of the system and, by extension, which properties are available in the output. The kind of systems and properties for which one can obtain reliable data is of key importance in the present context of using high-throughput computation to produce reliable training data for machine learning models. Hence, we will in the following review the important aspects of this choice in detail.

The level of theory that so far has been the standard for high-throughput computation in first-principles materials property databases is the semi-local, “second-rung” [25] level, which uses exchange-correlation functionals on the generalized gradient approximation (GGA) form. The most commonly used functional in the context of high-throughput calculations for materials databases is the one by Perdew, Burke, and Ernzerhof [26] (PBE) with the $+U$ correction [27]. This level of theory strikes a desirable balance between computational speed and accuracy while maintaining a high level of transferability. Nevertheless, the most popular GGA-type functionals, including PBE, have known shortcomings in their description of the electronic structure. The primary issues include: (1) a tendency to give energetics that in geometrical relaxations lead to a systematic over- or underestimation of bond lengths (the local density approximation, LDA, overbinds, whereas PBE underbinds); (2) an insufficient description of the physics of weak dispersion forces/van der Waals bonding; and (3) a systematically overdelocalized description of the KS orbitals that leads to inaccuracies in a number of properties that are derived from the orbitals. These three issues will be discussed in some more detail in the subsections below.

17.2.1 Overdelocalized Orbitals

The fundamental issue of overdelocalized KS orbitals is related to various aspects of the self-interaction error present at the semi-local exchange-correlation functional level of theory. A simplified picture is that the self-interaction introduces a repulsive electrostatic interaction of an electron with itself, leading to a delocalization that becomes more severe the more localized the correct representation of the orbital was supposed to be, i.e., the effect more severely impacts the more localized d-, and even more so, the f orbitals, compared to the less localized s and p states.

The result is a number of deficiencies in predicted materials properties. Examples of problematic properties include redox reaction energies [28, 29], the polarizability of extended systems [30, 31], and the silicon interstitial formation energy [32, 33].

In addition to these examples, issues are also seen in a number of properties calculated from the single-particle orbitals from the KS-DFT framework, where they are used as approximations of the "true quasi-electron orbitals" of the many-electron system (to the extent that such can be defined). However, from a fundamental perspective, the discussion of the accuracy of such properties is delicate because the DFT orbitals and the quasi-electron orbitals are not the same thing, even in theory for the exact exchange-correlation functional. Hence, one cannot a priori assume that an improved functional increases the agreement with the experimental values of, e.g., optical properties calculated from the KS band structure. Nevertheless, if one compares common GGA functionals to higher order methods that are still within the framework of KS-DFT (e.g., exact DFT exchange) one finds a qualitative difference in the orbital physics. This difference translates to that when materials properties which are directly associated with the electronic structure are calculated using higher-order theory, the results come out qualitatively closer to experiments than those calculated using standard GGAs. One can, therefore, take the position that it is a worthwhile improvement over standard semi-local functionals if improved functionals can make the orbitals to more closely mimic the orbital features given by higher order methods. This motivation is independent of the justification, or lack thereof, of using KS states to approximate quasi-particle bands for calculating materials properties. For an expanded discussion on this delicate topic, see, e.g., Ref. [34].

There are a range of well-known methods to address the description of localized states in semi-local DFT, (i) an explicit orbital-dependent correction that removes the surplus electrostatic term (sic correction) [35–37]; (ii) exact exchange DFT [38]; (iii) interpolating the DFT functional with Hartree–Fock exchange energy (hybrid functionals) [39–41]; (iv) use of the many-body Green's function for a more precise description of the localized quasi-particle orbitals (GW) [42]; (v) the DFT+U correction that adds an effective Hubbard-like term to the Hamiltonian to make selected localized orbitals energetically preferable [27]; and (vi) various attempts to modify the KS potential directly to make it reproduce essential features of exact exchange [31, 43–48]. All these methods, except for the last two (v, vi), require a vastly increased computational expense. Hybrid density functional methods (iii) are increasingly adopted for resolving these issues when the extra computational cost is acceptable. However, at a cost of roughly 50 times of that of standard GGAs, they are very inconvenient, or even completely unsuitable for, e.g., larger systems and high-throughput-type calculations.

Of the two less computationally expensive methods, DFT+U (v) is widely adopted as, arguably, the standard way of dealing with the issue of overdelocalized orbitals in high-throughput calculations and materials genome-type databases. However, DFT+U is not a highly transferable method; it requires attention in the assignment of site-specific "U-values." In setting the value of U, one selects how strongly a given localized orbital on a specific site prefers full occupation over partial occupation. In low throughput calculations, it is common to somewhat thoroughly investigate a system to arrive at a value of U that reasonably reproduces the expected physics of the system, but this is clearly not an option for high-throughput calculations. There are schemes to obtain sets of values that work well for systems with some specific type of physics, e.g., for typical oxides. However, in systems of mixed chemistries and intermixed types of bonding physics, the non-universality of U values becomes a serious problem. Energies obtained for different systems using different U-values for the same species cannot easily be mixed. Furthermore, since U values are usually only assigned to specific orbital projections on a pre-selected set of transition metal species, they cannot help with overdelocalized states of different origin, e.g., for defect states that are not atomic-orbital-like.

The second computationally less expensive method in the list above is (v) the approach to model the exchange-correlation potential directly to make it reproduce essential "non-local" features of exact exchange, instead of obtaining it as a functional derivative of an energy functional. Such potentials are known as *model potentials,* and have in some cases been quite successful [31, 43–48]. Some recent interest has been generated by the model potential of Becke and Johnson (BJ) [45], which was observed to mimic some of the crucial features of exact exchange for atoms. With various adjustments and extensions, it improves the polarization of hydrogen chains [31, 47], gives closer correspondence to experimental band gaps [48], and, to some extent, gives other improved properties [49, 50]. These model potentials seem promising for future adoption in high-throughput calculations to access properties that would otherwise be problematic because of orbital delocalization.

However, there are some fundamental issues with the general approach of model potentials. Since they directly model the exchange-correlation potential, the corresponding energy functionals are not merely unknown, they usually do not *exist* [46, 51, 52], and this deficiency cannot easily be corrected [53]. Since the KS equations are derived from a variational treatment of an energy equation, the use of such potentials has to be regarded on a weak formal basis, and are, strictly, outside the framework of KS-DFT. One cannot calculate any energy-derived properties from model potentials, e.g., one cannot do a geometry optimization that is consistent with the potential. Hence, if one starts from, e.g., theoretically generated structure candidates, one would have to use another method first to pre-relax the structure.

A closely related promising direction of functional development is the Armiento–Kümmel exchange functional (AK13) [54] (co-authored by the author of this chapter.) This is *a normal GGA exchange energy functional* that mimics the behavior of the BJ potential while avoiding the fundamental issues with model potentials. Similar to the modified BJ-based model potentials, the AK13 exchange energy functional gives qualitatively different orbitals from common GGA functionals. The results are a KS potential with improved atomic shell structure [54], improved ionization potentials from the highest eigenvalue [54] (but see the discussion in Ref. [55]), overall a KS band structure that better match that of higher order methods, including enlarged band gaps, and improved optical properties [34, 54, 56, 57].

As mentioned, the AK13 functional avoids the problem of undefined energies and energetics in model potentials. However, their values are not as accurate as those of commonly used GGAs and mostly insufficient. In addition, other issues appear from the AK13 construction that prevent its broader indiscriminate application [55,58,59]. We are hopeful that further research into modifications of the expression can overcome the difficulties while still retaining the favorable exchange potential features.

17.2.2 Under- and Overestimated Lattice Constants

On the issue of systematic under- and overestimation of lattice constants, this has mostly been resolved in functional development beyond PBE. The Armiento–Mattsson 2005 functional (AM05) [60, 61] is a semi-local functional with the same computational difficulty as PBE, but which gives roughly half the error for lattice constants. The comprehensive testing of Haas et al. finds for the lattice constants of 60 solids that the mean absolute error is 0.053 Å for PBE and 0.033 Å for AM05 [62–64]. Later functionals developed by Wu–Cohen in 2006 [65–67], SOGGA by Zaho, and Truhlar in 2008 [68], and PBEsol by Perdew et al. in 2009 [69–72] report similar improvements [63, 64, 70]. Further progress has been made by Perdew and coworkers on the meta-GGA level of theory, where, in addition to the electron density and its derivatives, a functional may also depend on the local value of the kinetic energy density of the KS particles. While meta-GGAs are technically more complex expressions than GGAs, implementations can be made that do not significantly increase the computational cost. The 2015 *Strongly Constrained and Appropriately Normed Semilocal Density Functional* (SCAN) meta-GGA [73] reportedly performs well for a wide range of properties for both solids and molecules, including lattice constants [74, 75]. However, some issues have recently been reported in the description of systems with itinerant magnetism [76].

17.2.3 Weak Dispersion Forces

On the topic of the description of van der Waals/London dispersion forces/weak interactions by semi-local DFT functionals, there exist a range of post-correction schemes of the energy to handle such interactions that can be deployed without any significant additional computational cost, see, e.g., Refs. [77–83]. Furthermore, there is a series of successful exchange-correlation functionals known as the vdW-DF from a collaboration between Chalmers University and Rutgers University [84–86] which allow a self-consistent treatment of these interactions. These functionals are not semi-local, but still fairly computationally inexpensive compared to, e.g., hybrid functionals. Furthermore, it has been shown that information about weak interactions can be extracted from local values of the kinetic energy density which are available to meta-GGAs [73, 87], at least to a level where the region around the equilibrium in van der Waals bonds can be described. This development has been incorporated in the SCAN functional [73].

17.3 Materials Properties

One of the central questions with the materials genome effort is what basic properties are within reach to be collected and included in these databases. This is determined by a combination of what can be described by the level of theory used for the computations (as carefully reviewed in the previous section), and what methods are available as automated workflows. The starting point, crucial for building any materials genome-type resource, is the crystal structures and corresponding formation energies. The importance of the formation energies is due to their use in creating composition phase diagrams to estimate the zero temperature thermodynamic stability of a material. The composition phase diagram gives the ground state crystal structure of a material at zero temperature as a function of composition. It is constructed by determining the convex hull of the predicted formation energies of all competing crystal structures in a chemical subspace [16, 88, 89]. A compound with a formation energy on the convex hull is stable, whereas a compound that ends up above the hull is unstable. The distance to the hull can be used as a rough estimate of the degree to which a material is unstable (i.e., how unlikely it is to be observed, and if observed, how quickly it would deteriorate into a combination of lower energy structures.) Crystal structures with a small hull distance (very roughly up to ∼50 meV) may still be regarded as candidates for materials that in practice may be stable since such an "error margin" can account for meta-stability, stability at limited elevated temperatures, and the computational inaccuracy of the methods.

Several works have investigated the accuracy of DFT calculations of formation energies. The standard deviation of formation energies calculated with PBE+U to experiments for the formation of ternary oxides from binary oxides was found to be 0.024 eV/atom; meaning 90% of the errors are within 0.047 eV/atom, which corresponds to a mean absolute error of approximately 0.02 eV/atom [90]. Kirklin et al. determined a mean absolute error of PBE formation energies of systems over all chemistries to be 0.136 eV/atom, but with energy corrections that are often used in high-throughput databases to some of the elemental phase energies, this lowers to 0.081 eV/atom [91]. However, the same paper notes that for 75 intermetallic structures they found experimental results from more than one source, giving an estimate for the mean absolute error in the experiments of 0.082 eV/atom. (Note that the latter estimate may be affected by selection bias, i.e., there may be a larger probability of finding multiple experimental values if the results are uncertain.)

Presently the set of materials properties beyond stability and formation energies available for large data sets is somewhat limited. There is an ongoing competition between the online materials genome-type databases to grow the data they provide both in terms of included structures and materials properties. There is a wealth of methods in the literature that could potentially be used to produce data for many different properties. However, to turn these methods into a form where they can run reliably in high-throughput is non-trivial. Among the available databases, the Materials Project is quite comprehensive in terms of properties. In addition to structural information and formation energies, they have over the years added the KS-DFT band structure (in some cases corrected using the GW approximation [42, 92]), elastic tensors [93], piezoelectric tensors [94], dielectric properties [95], phonon spectra [96], synthesis descriptors [97], and X-ray absorption spectra [98].

17.4 Database-Driven High-Throughput Calculations

A basic flowchart for materials design using database-driven high-throughput calculations is shown in Fig. 17.1. There are many software packages with partially overlapping aims for helping with the steps in the flowchart. Some recognized open source examples are the *atomic simulation environment—ASE* (wiki.fysik.dtu.dk/ase), *pymatgen, custodian, and fireworks* (pymatgen.org, see also the information at materialsproject.org/infrastructure; connected to the Materials Project), *aflow* (materials.duke.edu/AFLOW; connected to the AFLOW repository), *AiiDA* (aiida.net; connected to materials cloud), *qmpy* (connected to the open quantum materials database). The author is involved in the development of the open source *high-throughput toolkit—httk* (httk.org) framework, which we use extensively for high-throughput computation in our own research, and which provides the backend for the open materials database. This toolkit provides functionality for preparing and running unsupervised workflows of calculations (electronic structure, mostly targeted towards the software package VASP), analyzing the results, and storing

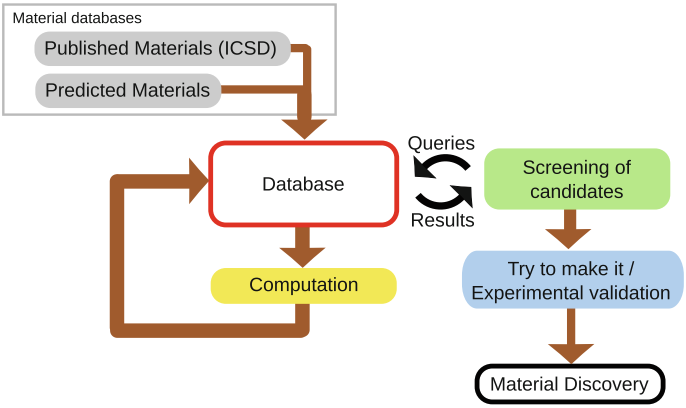

Fig. 17.1 A schematic flowchart representation of database-driven high-throughput materials design, largely inspired by the setup used in the Materials Project [16]. The steps on the right-hand sides represent the use of the database to find materials with desirable properties. In the context of machine learning models, the materials and materials properties in the database can be used for training and validation

them in a global and/or in a personalized database. The basic functionalities of these software packages are quite similar; in the following, we discuss the functionality of *httk*.

The primary focus of *httk* is for running automated calculations with as little human intervention as possible. This is crucial when working with large data sets, but can also be convenient when working with smaller projects. The toolkit consists of a software library developed in Python and a set of script programs that enable the interaction with supercomputers. The primary strengths of this framework compared to common alternatives are (1) the Python library provides a very integrated object-relational mapper, where classes in object-oriented Python are used to introduce abstractions that remove much of the difficulty in setting up a personal database of SQL type in which one can store, search, retrieve, and analyze results; (2) *httk* consistently allows the use of exact rational numbers in place of the more commonly used floating-point numbers. The exact rational numbers allow processing of crystal structures, application of transforms, etc., without the usual loss of precision. Hence, *httk* can deterministically produce an internal representation of structures read from a source file (e.g., on the cif file format), which is not the case in most other frameworks due to their use of floating-point numbers means the precise end result is influenced by the computer architecture.

The *httk* framework is distributed in several ways, including the PyPI service. Hence, it can easily be installed by issuing: `pip install httk` on a system with a modern distribution of Python. There is a set of tutorial steps and a large number of examples available to show how the framework can be utilized in the various steps of database-driven high-throughput as shown in Fig. 17.1. These are available via the project website (httk.openmaterialsdb.se).

17.5 Machine Learning Models for Materials Design

17.5.1 Models for Molecules

The primary focus in this chapter is on a type of machine learning models for use in materials design that can be said to begin with a 2012 paper by Rupp et al. on the use of kernel ridge regression for small molecules [99]. They define a matrix representation for molecules named the "Coulomb matrix." In this representation a system of N atoms generates an $N \times N$ matrix where the off-diagonal elements (i,j) are the Coulomb repulsion between the ith and jth bare atomic cores, and the diagonal elements are based on a polynomial fit to energies of free atoms to represent a static energy contribution pertaining to the ith atom,

$$C_{ij} = \begin{cases} 0.5Z_i^{2.4} & \text{if } i = j \\ Z_i Z_j/(\|\mathbf{r}_i - \mathbf{r}_j\|_2) & \text{if } i \neq j \end{cases} \tag{17.1}$$

One may note that the Coulomb interaction between the bare atomic cores is not a good indicator of the physics of the bonds in a system. However, the representation does not aim to push the machine learning model into a specific physics-based description, but just to constitute a well-formed way to represent the structural information (i.e., the positions of the atoms) so that the machine is free to learn the physics from the data. This model was trained on small organic molecules (with up to 7 atoms from the elements C, O, N, and S, and with the valencies satisfied by hydrogen atoms; this data set is named *qm7*.) It was shown in the original paper that the machine can be trained to predict atomization energies of molecules not in the training set down to a mean absolute error of 10 kcal/mol at a training set size of 7k. In units more common for materials, this model reaches 20 meV/atom at a training set of 3000 molecules from qm7 [100].

17.5.2 General Models for Periodic Systems

In a 2015 work Faber, Lindmaa, von Lilienfeld, and Armiento (the author of the present chapter) extended the Coulomb matrix construct into a suitable form for periodic crystals [100]. This extension is non-trivial, since there exist more than one way to choose a unit cell in a periodic system, and therefore representations based on the Coulomb matrix easily become non-unique. As pointed out in that paper, the aim when seeking a representation for atomistic systems is to find one that is (1) *complete*: incorporates all features of the structural information that are relevant for the underlying problem, but at the same time; (2) *compact*: avoids representation of features irrelevant for the underlying problem (e.g., static rotations); (3) *descriptive*: structural similarity should give representations that are close; and (4) *simple*: low computational effort to compute, and conceptually easy to understand.

The end result of Ref. [100] was three alternative Coulomb matrix inspired representations applicable to periodic crystals. The first one was based on replacing the bare Coulomb interactions in the off-diagonal matrix elements with the corresponding expression for fully periodic systems, i.e., the sum of the total Coulomb interaction energy per unit cell between the infinite periodic lattices of the bare cores of repetitions of two separate atoms in the unit cell. These expressions are evaluated via Ewald sums [101]. The issue with this expression is that it is somewhat computationally expensive and non-trivial to evaluate correctly. The second generalization of the Coulomb matrix was to duplicate the unit cell a number of times and then use the same expression as for the non-periodic Coulomb matrix, however, with a screened Coulomb interaction (i.e., where the interaction decays exponentially to give a finite reach.) This is very similar to just using the short range term in the Ewald sum. To get an even simpler descriptor, a third expression was invented. It was shown how the Ewald sum can be replaced by an expression that mimics the basic shape and periodicity of the Ewald expression, but which still remains on a simple closed form that is easy to evaluate. This expression was named the “sine” or “sinusoidal” descriptor, because of how it reproduced the periodicity over the unit cell via a sine function.

The three alternative extensions of the Coulomb matrix to periodic systems were tested on a data set that is now known as FLAA (from the authors' initials). It consists of structures with up to 25 atoms that were randomly selected out of the Materials Project database. In these structures most atomic species occur, in proportions roughly similar to their occurrence in structures published in the literature and extracted into the inorganic crystal structure database (ICSD) [102, 103] which is the main source of crystal structures for the Materials Project. The conclusion of the 2015 paper [100] was that all three alternative extensions of the Coulomb matrix to periodic systems performed approximately equal. The sine descriptor did slightly better than the others, with a 370 meV/atom mean absolute error for predicting formation energies when trained on 3k structures from the FLAA data set.

Two main conclusions follow from the above results. Firstly, the performance of kernel ridge regression-based machines for atomistic systems does not appear to be particularly sensitive to the exact details of how the generalized Coulomb matrix descriptors are constructed, as long as they reasonably well adhere to the aims for a good representation listed above. Secondly, at first glance it may appear as if the performance of the models for molecules far outperforms the corresponding ones for periodic crystals (20 meV/atom vs. 0.370 meV/atom). However, the sizes of the chemical space for the two cases are not comparable, and arguably the one used for crystals in Ref. [100] is far larger.

17.5.3 Crystal-Structure Specific Models

To demonstrate that these types of models are capable of reaching a level of accuracy directly useful for applications if one restricts the chemical space, the same authors investigated in 2016 a machine learning model operating on such a smaller space [104]. This work considered all substitutions of main group elements into four sites of one specific quaternary crystal structure, the elpasolite. This structure was selected because it is the quaternary crystal most frequently occurring with different substitutions in the ICSD database, indicating that this structure can accommodate many types of bonds and thus to be rewarding to characterize fully. High-throughput DFT calculations using the *httk* framework were used to produce data for ca 10k substitutions of elements into the elpasolite crystal structure out a total of two million possibilities. Furthermore, a subset of 12 main group elements was selected to give a reduced chemical space of 12k possible substitutions, which were run exhaustively.

A substitution into a fixed crystal structure can be uniquely specified by giving which chemical species are at which atomic site in the structure. Hence, the 2016 paper used a very straightforward representation of, essentially, a 2×4 matrix that specified the row and column in the periodic table of the atom species at each of the four sites in the elpasolite structure. This leaves out the precise structural information of the system from the descriptor, i.e., the bond lengths between the atoms. The 2×4 matrix descriptor should be understood to technically refer to the system relaxed while confined to the elpasolite crystal structure.

A kernel ridge regression machine learning model was trained using this descriptor on formation energies for structures in the elpasolite data set, and it was shown that (1) by training on a sufficiently large subset of the exhaustive 12k data set, the model can reach essentially any level of accuracy for predictions of structures outside the training set, at least below <10 meV/atom which is significantly less than the errors in the DFT data. (See the discussions of accuracy of DFT formation energies in Sect. 17.3.) This shows that the performance of this machine learning model is merely a question of having a large enough training set; (2) when training on data in the larger chemical space of two million possible substitutions of main group elements into the elpasolite structure, it was sufficient to train on about 10k structures to reach roughly the accuracy of the DFT calculations, 100 meV/atom. This result means that the machine learning model was capable of producing DFT-quality formation energies with a net $\times 200$ speedup, including all the time used to produce the training data. The resulting two million formation energies are illustrated in Fig. 17.2 reproduced from the original paper.

Furthermore, the 2016 paper also demonstrated a practical use of the large set of predicted formation energies. Phase diagrams were created for most of the elpasolite systems by using information about competing compositions from the Materials Project using the pymatgen Python library (some systems were outright dismissed on grounds of containing rare-gas elements). From these phase diagrams a number of candidates for thermodynamically stable materials were obtained by identifying compositions with a predicted formation energy on the convex hull. These candidates were validated by DFT calculations and 90 systems were confirmed to be thermodynamically stable within this level of theory. However, the compounds that passed validation only constituted a small fraction of the candidates. As explained in the paper, the reason is that the process of identifying structures on the convex hull is a screening for systems with the lowest formation energies, which are outliers in the full data set. The interpolative nature of machine learning models leads to them being significantly less accurate in predicting properties of outlier systems. Nevertheless, even with this limitation, the scheme far reduced the number of DFT calculations needed to identify thermodynamically stable elpasolite systems compared to just obtaining all formation energies from DFT calculations. The net result was a $\times 11$ speedup, including the full time spent both on the training set and the calculations used to validate the materials picked out as candidates for stability.

Hence, the crystal-structure-specific machine was demonstrated to be very successful for generating large amounts of formation energy data which is useful for greatly accelerating predictions of stable compounds in a considered crystal structure. The predictions allow extending the available data in materials genome-type databases. The structures identified as stable in the work discussed above are now available (with some singular exceptions) via the Materials Project and, e.g., enters the predictions of convex hulls for user-generated phase diagrams via their online service, thus contributing to the accuracy of those predictions.

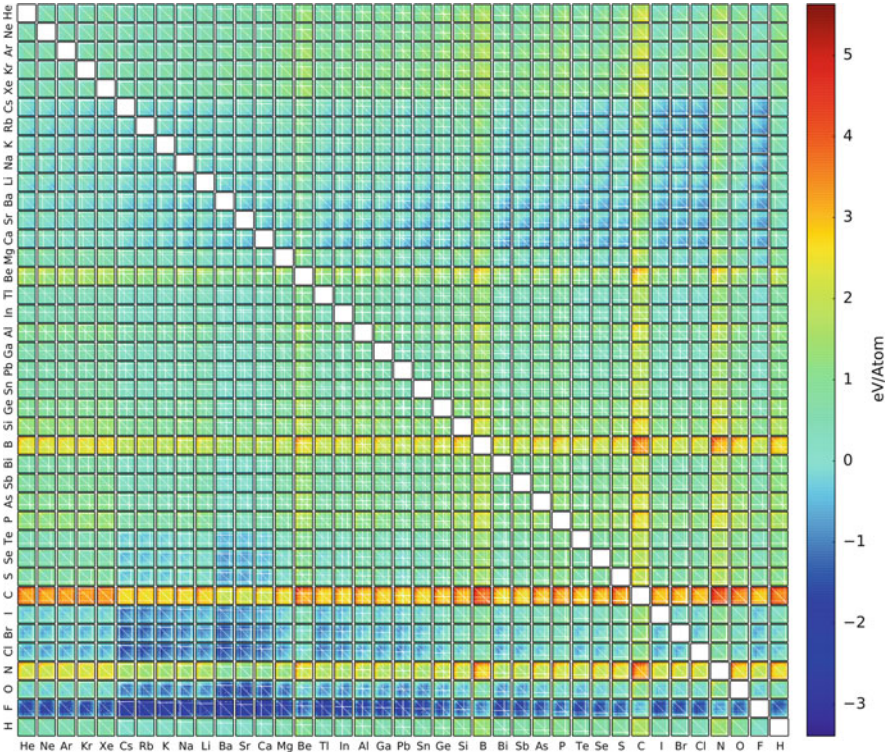

Fig. 17.2 Color matrix of the two million elpasolite energies predicted with the crystal-structure specific machine learning model of Faber et al. [104]. The x- and y-axes specify which atomic species sits on two of the four sites in the crystal structure. At those coordinates one finds a miniature diagram over the species at the remaining two sites. Every pixel in the miniature diagram shows a formation energy of the corresponding composition of four atomic species. The figure is reproduced from the original paper and is licensed under the Creative Commons Attribution 3.0 License

17.5.4 Models for Predicting Composition Phase Diagrams, Crystal Structures

The success of the crystal-structure-specific machine notwithstanding, it does not directly answer the most typical materials design problem. It is, arguably, more common to seek the stable crystal structures that can be formed from a given set of chemical species, rather than all the stable chemical compositions that share the same crystal structure. This is, in essence, the crystal structure prediction problem.

In 2016, Tholander, Andersson, Armiento, Tasnádi, and Alling [105] (TAATA) produced a data set by high-throughput calculations using the *httk* framework. The aim was to seek stable crystal structures in the ternary chemical systems Ti-Zn-N, Zr-Zn-N, and Hf-Zn-N for possible use in piezoelectrics. This high-throughput data set is a good real-world test case to evaluate the possible acceleration of the generation of phase diagrams for identifying stable structures using machine learning models.

The author of this chapter and coworkers have since then engaged in a project of trying out new machine learning models on this problem and to develop new ones for it; the progress on this was recently reported in, e.g., Ref. [106]. At the present stage, it appears the original Coulomb matrix-based descriptors from Ref. [100] perform similar on this data set as for the original FLAA data set, which is encouraging in establishing the generality of these models. However, the resulting accuracy is not sufficient to be useful for accelerating the production of the phase diagrams. Compared to the FLAA set, the TAATA data set has much fewer atomic species, but at the same time is comprised of structures over a very wide range of formation energies. The origin of the structures in the FLAA set is the Materials Project which, as explained above, are based on structures from the ICSD database. The ICSD primarily indexes materials seen in nature which means most are thermodynamically stable and have comparably low formation energies. This restriction lowers the dimensionality of the chemical space of FLAA relative to that of TAATA.

Other recent machine learning models perform better; e.g., in Ref. [106] it was found that a descriptor by Ward et al. that encodes structural information using a Voronoi tessellation reaches a mean absolute error of 0.28 eV/atom for 10k structures from the TAATA data set [107]. While errors on this level are not small enough to replace the need for DFT calculations with model predictions, one may still be able to use predictions to identify and remove competing structures that are highly unstable and therefore would not influence the phase diagram, thus reducing the number of DFT calculations necessary, giving an overall reduction in the effort of producing the phase diagram. The field moves rapidly forward, and some other interesting recent developments are found in Refs. [108–111].

17.6 Conclusions and Outlook

This chapter has reviewed several aspects of producing training data by database-driven high-throughput calculations, and the use of this data to train machine learning models with the aim of accelerating materials design. All these aspects are making rapid and encouraging progress. The research-front machine learning methods are now on the edge of producing results that are accurate and reliable enough to accelerate theoretical prediction of thermodynamic stability via the creation of convex hulls; i.e., the crystal prediction problem which arguably is the most important first step for materials design of bulk materials with desired properties. Further progress towards this goal, and for predicting other properties, is continuously being made. Looking forward, two crucial points can be raised: (1) further development of general machine learning models for atomistic systems with improved accuracy and a reduced need for training data is needed; but how far can that development go before it hits a fundamental wall where not enough information about the underlying physics is present in the data?; (2) the rapid development of machine learning models will drive a need for more accurate training data. Will the progress of physics-based computational methods be able to keep up with this need of methods with improved accuracy but low enough computational effort to be useful in high-throughput?; or will the lack of a sufficient amount of high quality training data become a major bottleneck for further progress? Future research needs to target both these areas.

Acknowledgments The author thanks Anatole von Lilienfeld and Felix Faber for many insightful discussions on topics in the overlap of machine learning and materials design. Joel Davidsson is acknowledged for help with supervising the master's thesis discussed in the text as Ref. [106]. The author acknowledges support from the Swedish e-Science Centre (SeRC), Swedish Research Council (VR) Grants No. 2016-04810, and the Centre in Nano science and Nanotechnology (CeNano) at Linköping University. Some of the discussed computations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at the National Supercomputer Centre (NSC) at Linköping University.

References

1. J. Greeley, T.F. Jaramillo, J. Bonde, I. Chorkendorff, J.K. Nørskov, Nat. Mater. **5**(11), 909 (2006)
2. K. Kang, Y.S. Meng, J. Bréger, C.P. Grey, G. Ceder, Science **311**(5763), 977 (2006)
3. S. Kirklin, B. Meredig, C. Wolverton, Adv. Energy Mater. **3**(2), 252 (2013)
4. C. Ortiz, O. Eriksson, M. Klintenberg, Comput. Mater. Sci. **44**(4), 1042 (2009)
5. M. Klintenberg, O. Eriksson, Comput. Mater. Sci. **67**, 282 (2013)
6. G.K.H. Madsen, J. Am. Chem. Soc. **128**(37), 12140 (2006)
7. S. Wang, Z. Wang, W. Setyawan, N. Mingo, S. Curtarolo, Phys. Rev. X **1**(2), 021012 (2011)
8. R. Armiento, B. Kozinsky, M. Fornari, G. Ceder, Phys. Rev. B **84**(1) (2011)
9. R. Armiento, B. Kozinsky, G. Hautier, M. Fornari, G. Ceder, Phys. Rev. B **89**(13), 134103 (2014)
10. G. Hautier, A. Miglio, G. Ceder, G.M. Rignanese, X. Gonze, Nat. Commun. **4**, 2292 (2013)
11. S. Lebègue, T. Björkman, M. Klintenberg, R.M. Nieminen, O. Eriksson, Phys. Rev. X **3**(3), 031002 (2013)
12. S. Curtarolo, G.L.W. Hart, M.B. Nardelli, N. Mingo, S. Sanvito, O. Levy, Nat. Mater. **12**(3), 191 (2013)
13. G. Ceder, K.A. Persson, Sci. Amer. **309**(6), 36 (2013)
14. K. Alberi, M.B. Nardelli, A. Zakutayev, L. Mitas, S. Curtarolo, A. Jain, M. Fornari, N. Marzari, I. Takeuchi, M.L. Green, M. Kanatzidis, M.F. Toney, S. Butenko, B. Meredig, S. Lany, U. Kattner, A. Davydov, E.S. Toberer, V. Stevanovic, A. Walsh, N.G. Park, A. Aspuru-Guzik, D.P. Tabor, J. Nelson, J. Murphy, A. Setlur, J. Gregoire, H. Li, R. Xiao, A. Ludwig, L.W. Martin, A.M. Rappe, S.-H. Wei, J. Perkins, J. Phys. D: Appl. Phys. **52**(1), 013001 (2019)
15. F. Oba, Y. Kumagai, Appl. Phys. Express **11**(6), 060101 (2018)
16. A. Jain, G. Hautier, C.J. Moore, S. Ping Ong, C.C. Fischer, T. Mueller, K.A. Persson, G. Ceder, Comput. Mater. Sci. **50**(8), 2295 (2011)
17. A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K.A. Persson, APL Mater. **1**(1), 011002 (2013)
18. Executive Office of the President National Science and Technology Council, Washington. Materials Genome Initiative for Global Competitiveness (2011). https://www.mgi.gov/sites/default/files/documents/materials_genome_initiative-final.pdf; https://www.mgi.gov/
19. K. Rajan, Mater. Today **8**(10), 38 (2005)

20. J.R. Rodgers, D. Cebon, MRS Bull. **31**(12), 975 (2006)
21. R. Kurzweil, *The Singularity Is Near: When Humans Transcend Biology* (Penguin Books, New York, 2006)
22. S. Ulam, Bull. Amer. Math. Soc. **64**(3), 1 (1958)
23. P. Hohenberg, W. Kohn, Phys. Rev. **136**(3B), B864 (1964)
24. W. Kohn, L.J. Sham, Phys. Rev. **140**(4A), A1133 (1965)
25. J.P. Perdew, K. Schmidt, in *AIP Conference Proceedings*, vol. 577 (AIP, College Park, 2001), pp. 1–20
26. J.P. Perdew, K. Burke, M. Ernzerhof, Phys. Rev. Lett. **77**(18), 3865 (1996)
27. V.I. Anisimov, F. Aryasetiawan, A.I. Lichtenstein, J. Phys. Condens. Matter **9**(4), 767 (1997)
28. F. Zhou, M. Cococcioni, C.A. Marianetti, D. Morgan, G. Ceder, Phys. Rev. B **70**(23), 235121 (2004)
29. V.L. Chevrier, S.P. Ong, R. Armiento, M.K.Y. Chan, G. Ceder, Phys. Rev. B **82**(7), 075122 (2010)
30. S. Kümmel, L. Kronik, J.P. Perdew, Phys. Rev. Lett. **93**(21), 213002 (2004)
31. R. Armiento, S. Kümmel, T. Körzdörfer, Phys. Rev. B **77**(16), 165106 (2008)
32. A.E. Mattsson, R.R. Wixom, R. Armiento, Phys. Rev. B **77**(15), 155211 (2008)
33. P. Rinke, A. Janotti, M. Scheffler, C.G. Van de Walle, Phys. Rev. Lett. **102**(2), 026402 (2009)
34. V. Vlček, G. Steinle-Neumann, L. Leppert, R. Armiento, S. Kümmel, Phys. Rev. B **91**(3), 035107 (2015)
35. J.P. Perdew, Chem. Phys. Lett. **64**(1), 127 (1979)
36. J.P. Perdew, A. Zunger, Phys. Rev. B **23**(10), 5048 (1981)
37. R.O. Jones, O. Gunnarsson, Rev. Mod. Phys. **61**(3), 689 (1989)
38. M. Städele, M. Moukara, J.A. Majewski, P. Vogl, A. Görling, Phys. Rev. B **59**(15), 10031 (1999)
39. A.D. Becke, J. Chem. Phys. **98**(7), 5648 (1993)
40. J. Heyd, G.E. Scuseria, M. Ernzerhof, J. Chem. Phys. **118**(18), 8207 (2003)
41. J. Heyd, G.E. Scuseria, M. Ernzerhof, J. Chem. Phys. **124**(21), 219906 (2006)
42. L. Hedin, Phys. Rev. **139**(3A), A796 (1965)
43. R. van Leeuwen, E.J. Baerends, Phys. Rev. A **49**(4), 2421 (1994)
44. O. Gritsenko, R. van Leeuwen, E. van Lenthe, E.J. Baerends, Phys. Rev. A **51**(3), 1944 (1995)
45. A.D. Becke, E.R. Johnson, J. Chem. Phys. **124**(22), 221101 (2006)
46. N. Umezawa, Phys. Rev. A **74**(3), 032505 (2006)
47. E. Räsänen, S. Pittalis, C.R. Proetto, J. Chem. Phys. **132**(4), 044112 (2010)
48. F. Tran, P. Blaha, Phys. Rev. Lett. **102**(22), 226401 (2009)
49. M.J.T. Oliveira, E. Räsänen, S. Pittalis, M.A.L. Marques, J. Chem. Theory Comput. **6**(12), 3664 (2010)
50. D.J. Singh, Phys. Rev. B **82**(20), 205102 (2010)
51. R. van Leeuwen, E.J. Baerends, Phys. Rev. A **51**(1), 170 (1995)
52. A.P. Gaiduk, V.N. Staroverov, Phys. Rev. A **83**(1), 012509 (2011)
53. A. Karolewski, R. Armiento, S. Kümmel, J. Chem. Theory Comput. **5**(4), 712 (2009)
54. R. Armiento, S. Kümmel, Phys. Rev. Lett. **111**(3), 036402 (2013)
55. T. Aschebrock, R. Armiento, S. Kümmel, Phys. Rev. B **96**(7), 075140 (2017)
56. T.F.T. Cerqueira, M.J.T. Oliveira, M.A.L. Marques, J. Chem. Theory Comput. **10**(12), 5625 (2014)
57. F. Tran, P. Blaha, M. Betzinger, S. Blügel, Phys. Rev. B **91**(16), 165121 (2015)
58. A. Lindmaa, R. Armiento, Phys. Rev. B **94**(15), 155143 (2016)
59. T. Aschebrock, R. Armiento, S. Kümmel, Phys. Rev. B **95**(24), 245118 (2017)
60. R. Armiento, A.E. Mattsson, Phys. Rev. B **72**(8), 085108 (2005)
61. A.E. Mattsson, R. Armiento, Phys. Rev. B **79**(15), 155101 (2009)
62. A.E. Mattsson, R. Armiento, J. Paier, G. Kresse, J.M. Wills, T.R. Mattsson, J. Chem. Phys. **128**(8), 084714 (2008)
63. P. Haas, F. Tran, P. Blaha, Phys. Rev. B **79**(8), 085104 (2009)
64. P. Haas, F. Tran, P. Blaha, Phys. Rev. B **79**(20), 209902 (2009)

65. Z. Wu, R.E. Cohen, Phys. Rev. B **73**(23), 235116 (2006)
66. Y. Zhao, D.G. Truhlar, Phys. Rev. B **78**(19), 197101 (2008)
67. Z. Wu, R.E. Cohen, Phys. Rev. B **78**(19), 197102 (2008)
68. Y. Zhao, D.G. Truhlar, J. Chem. Phys. **128**(18), 184109 (2008)
69. J.P. Perdew, A. Ruzsinszky, G.I. Csonka, O.A. Vydrov, G.E. Scuseria, L.A. Constantin, X. Zhou, K. Burke, Phys. Rev. Lett. **100**(13), 136406 (2008)
70. A.E. Mattsson, R. Armiento, T.R. Mattsson, Phys. Rev. Lett. **101**(23), 239701 (2008)
71. J.P. Perdew, A. Ruzsinszky, G.I. Csonka, O.A. Vydrov, G.E. Scuseria, L.A. Constantin, X. Zhou, K. Burke, Phys. Rev. Lett. **101**(23), 239702 (2008)
72. J.P. Perdew, A. Ruzsinszky, G.I. Csonka, O.A. Vydrov, G.E. Scuseria, L.A. Constantin, X. Zhou, K. Burke, Phys. Rev. Lett. **102**(3), 039902 (2009)
73. J. Sun, A. Ruzsinszky, J.P. Perdew, Phys. Rev. Lett. **115**(3), 036402 (2015)
74. J. Sun, R.C. Remsing, Y. Zhang, Z. Sun, A. Ruzsinszky, H. Peng, Z. Yang, A. Paul, U. Waghmare, X. Wu, M.L. Klein, J.P. Perdew, Nat. Chem. **8**(9), 831 (2016). https://doi.org/10.1038/nchem.2535. https://www.nature.com/articles/nchem.2535
75. Y. Zhang, D.A. Kitchaev, J. Yang, T. Chen, S.T. Dacek, R.A. Sarmiento-Pérez, M.A.L. Marques, H. Peng, G. Ceder, J.P. Perdew, J. Sun, npj Comput. Mater. **4**(1), 9 (2018). https://doi.org/10.1038/s41524-018-0065-z. https://www.nature.com/articles/s41524-018-0065-z
76. M. Ekholm, D. Gambino, H.J.M. Jönsson, F. Tasnádi, B. Alling, I.A. Abrikosov, Phys. Rev. B **98**(9), 094413 (2018). https://doi.org/10.1103/PhysRevB.98.094413. https://link.aps.org/doi/10.1103/PhysRevB.98.094413
77. S. Grimme, J. Comput. Chem. **27**(15), 1787 (2006)
78. S. Grimme, J. Antony, S. Ehrlich, H. Krieg, J. Chem. Phys. **132**(15), 154104 (2010)
79. A. Tkatchenko, M. Scheffler, Phys. Rev. Lett. **102**(7), 073005 (2009)
80. A. Tkatchenko, R.A. DiStasio, R. Car, M. Scheffler, Phys. Rev. Lett. **108**(23), 236402 (2012)
81. A. Ambrosetti, A.M. Reilly, R.A. DiStasio, A. Tkatchenko, J. Chem. Phys. **140**(18), 18A508 (2014)
82. S.N. Steinmann, C. Corminboeuf, J. Chem. Theory Comput. **7**(11), 3567 (2011)
83. S.N. Steinmann, C. Corminboeuf, J. Chem. Phys. **134**(4), 044117 (2011)
84. M. Dion, H. Rydberg, E. Schröder, D.C. Langreth, B.I. Lundqvist, Phys. Rev. Lett. **92**(24), 246401 (2004)
85. K. Lee, E.D. Murray, L. Kong, B.I. Lundqvist, D.C. Langreth, Phys. Rev. B **82**(8), 081101 (2010)
86. K. Berland, P. Hyldgaard, Phys. Rev. B **89**(3), 035412 (2014)
87. J. Sun, B. Xiao, Y. Fang, R. Haunschild, P. Hao, A. Ruzsinszky, G.I. Csonka, G.E. Scuseria, J.P. Perdew, Phys. Rev. Lett. **111**(10), 106401 (2013)
88. A.R. Akbarzadeh, V. Ozoliņš, C. Wolverton, Adv. Mater. **19**(20), 3233 (2007)
89. S.P. Ong, L. Wang, B. Kang, G. Ceder, Chem. Mater. **20**(5), 1798 (2008)
90. G. Hautier, S.P. Ong, A. Jain, C.J. Moore, G. Ceder, Phys. Rev. B **85**(15), 155208 (2012)
91. S. Kirklin, J.E. Saal, B. Meredig, A. Thompson, J.W. Doak, M. Aykol, S. Rühl, C.M. Wolverton, npj Comput. Mater. **1**, 15010 (2015)
92. I.E. Castelli, F. Hüser, M. Pandey, H. Li, K.S. Thygesen, B. Seger, A. Jain, K.A. Persson, G. Ceder, K.W. Jacobsen, Adv. Energy Mater. **5**(2), 1400915 (2015)
93. M. de Jong, W. Chen, T. Angsten, A. Jain, R. Notestine, A. Gamst, M. Sluiter, C. Krishna Ande, S. van der Zwaag, J.J. Plata, C. Toher, S. Curtarolo, G. Ceder, K.A. Persson, M. Asta, Sci. Data **2**, 150009 (2015)
94. M. de Jong, W. Chen, H. Geerlings, M. Asta, K.A. Persson, Sci. Data **2**, 150053 (2015)
95. I. Petousis, D. Mrdjenovich, E. Ballouz, M. Liu, D. Winston, W. Chen, T. Graf, T.D. Schladt, K.A. Persson, F.B. Prinz, Sci. Data **4**, 160134 (2017)
96. G. Petretto, S. Dwaraknath, H.P.C. Miranda, D. Winston, M. Giantomassi, M.J. van Setten, X. Gonze, K.A. Persson, G. Hautier, G.M. Rignanese, Sci. Data **5**, 180065 (2018)
97. E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder, E. Olivetti, Chem. Mater. **29**(21), 9436 (2017)

98. K. Mathew, C. Zheng, D. Winston, C. Chen, A. Dozier, J.J. Rehr, S.P. Ong, K.A. Persson, Sci. Data **5**, 180151 (2018)
99. M. Rupp, A. Tkatchenko, K.R. Müller, O.A. von Lilienfeld, Phys. Rev. Lett. **108**(5), 058301 (2012)
100. F. Faber, A. Lindmaa, O.A.V. Lilienfeld, R. Armiento, Int. J. Quantum Chem. **115**(16), 1094 (2015)
101. P.P. Ewald, Ann. Phys. **369**(3), 253 (1921)
102. G. Bergerhoff, R. Hundt, R. Sievers, I.D. Brown, J. Chem. Inf. Comput. Sci. **23**(2), 66 (1983)
103. A. Belsky, M. Hellenbrandt, V.L. Karen, P. Luksch, Acta Cryst. B **58**(3–1), 364 (2002)
104. F.A. Faber, A. Lindmaa, O.A.v. Lilienfeld, R. Armiento, Phys. Rev. Lett. **117**(13), 135502 (2016)
105. C. Tholander, C.B.A. Andersson, R. Armiento, F. Tasnádi, B. Alling, J. Appl. Phys. **120**(22), 225102 (2016)
106. C. Bratu, Machine Learning of Crystal Formation Energies with Novel Structural Descriptors. Master's Thesis, Linköping University, Sweden, 2017. http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-143203
107. L. Ward, R. Liu, A. Krishna, V.I. Hegde, A. Agrawal, A. Choudhary, C. Wolverton, Phys. Rev. B **96**(2), 024104 (2017)
108. F.A. Faber, A.S. Christensen, B. Huang, O.A. von Lilienfeld, J. Chem. Phys. **148**(24), 241717 (2018)
109. H. Huo, M. Rupp (2017). arXiv:1704.06439
110. K.T. Schütt, H.E. Sauceda, P.J. Kindermans, A. Tkatchenko, K.R. Müller, J. Chem. Phys. **148**(24), 241722 (2018)
111. W. Ye, C. Chen, Z. Wang, I.H. Chu, S.P. Ong, Nat. Commun. **9**(1), 3800 (2018)