scientific **data**

Check for updates

OPEN

DATA DESCRIPTOR

GEOM, energy-annotated molecular conformations for property prediction and molecular generation

Simon Axelrod[1,2] & Rafael Gómez-Bombarelli [ID][2 ✉]

Machine learning (ML) outperforms traditional approaches in many molecular design tasks. ML models usually predict molecular properties from a 2D chemical graph or a single 3D structure, but neither of these representations accounts for the ensemble of 3D conformers that are accessible to a molecule. Property prediction could be improved by using conformer ensembles as input, but there is no large-scale dataset that contains graphs annotated with accurate conformers and experimental data. Here we use advanced sampling and semi-empirical density functional theory (DFT) to generate 37 million molecular conformations for over 450,000 molecules. The Geometric Ensemble Of Molecules (GEOM) dataset contains conformers for 133,000 species from QM9, and 317,000 species with experimental data related to biophysics, physiology, and physical chemistry. Ensembles of 1,511 species with BACE-1 inhibition data are also labeled with high-quality DFT free energies in an implicit water solvent, and 534 ensembles are further optimized with DFT. GEOM will assist in the development of models that predict properties from conformer ensembles, and generative models that sample 3D conformations.

Background & Summary

Accurate and affordable prediction of molecular properties is a longstanding goal of computational chemistry. Predictions can be generated with rule-based[1] or physics-based[2] methods, which typically involve a trade-off between accuracy and speed. Machine learning offers an attractive alternative, as it is far quicker than physics-based methods and outperforms traditional rule-based baselines in many molecule-related tasks, including property prediction and virtual screening[3–5], inverse design using generative models[6–10], reinforcement learning[11–13], differentiable simulators[14,15], and synthesis planning and retrosynthesis[16,17].

Advances in molecular machine learning have been enabled by algorithmic improvements[18–22] and by reference datasets and tasks[23]. A number of reference datasets provide unlabeled molecules for generation tasks[7,24–27] or experimentally labeled molecules for property prediction[23,28–31]. The molecules are typically represented as SMILES[32] or InChi[33] strings, which can be converted into 2D graphs, or as single 3D structures. These representations can be used as input to machine learning models that predict properties or generate new compounds. However, these representations fail to capture the flexibility of molecules, which consist of atoms in continual motion on a potential energy surface (PES). Molecular properties are a function of the conformers accessible at finite temperature[34,35], which are not explicitly included in a 2D or single 3D representation (Fig. 1). Models that map conformer ensembles to experimental properties could be of interest, but they require a dataset with both conformers and experimental data.

Here we present the Geometric Ensemble Of Molecules (GEOM), a dataset of high-quality conformers for 317,928 mid-sized organic molecules with experimental data, and 133,258 molecules from the QM9 dataset[36]. 304,466 drug-like species and their biological assay results were accessed as part of AICures (https://www.aicures.mit.edu), an open machine learning challenge to predict which drugs can be repurposed to treat COVID-19 and related illnesses. 16,865 molecules are from the MoleculeNet benchmark[31]. They are labeled with experimental properties related to physical chemistry, biophysics, and physiology. Conformers were generated

[1]Harvard University, Department of Chemistry and Chemical Biology, Cambridge, MA, 02138, USA. [2]Massachusetts Institute of Technology, Department of Materials Science and Engineering, Cambridge, MA, 02139, USA. ✉e-mail: rafagb@mit.edu

CC(C)OC(=O)CCC/C=C\\C[C@H]1[C@@H](O)C[C@@H](O)[C@@H]1CC[C@@H](O)CCc1ccccc1
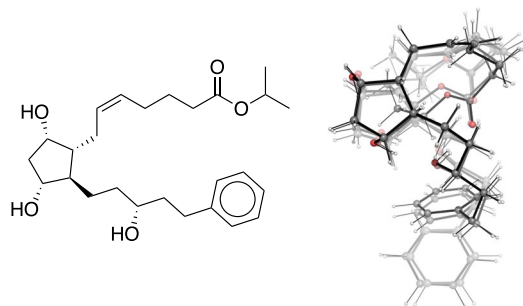

Fig. 1 Molecular representations of the latanoprost molecule. *top* SMILES string. *left* Stereochemical formula with edge features, including wedges for in- and out-of-plane bonds, and a double line for *cis* isomerism. *right* Overlay of conformers. Higher transparency corresponds to lower statistical weight.

AICures drug dataset (N = 304,466)			
	Mean	Standard deviation	Maximum
Number of atoms	44.4	11.3	181
Number of heavy atoms	24.9	5.7	91
Molecular weight (amu)	355.4	80.4	1549.7
Number of rotatable bonds	6.5	3.0	53
Stereochemistry (specified)	45,712	—	—
Stereochemistry (all)	83,326	—	—
QM9 dataset (N = 133,258)			
	Mean	Standard deviation	Maximum
Number of atoms	18.0	3.0	29
Number of heavy atoms	8.8	0.51	9
Molecular weight (amu)	122.7	7.6	152.0
Number of rotatable bonds	2.2	1.6	8
Stereochemistry (specified)	95,734	—	—
Stereochemistry (all)	95,734	—	—

Table 1. Molecular descriptor statistics for the QM9 and AICures molecules in the GEOM dataset.

with the CREST program[37], which uses extensive sampling based on the semi-empirical extended tight-binding method (GFN2-xTB[38]) to generate reliable and accurate structures. CREST ensembles from 1,511 species in the BACE dataset[39] were also labeled with high-accuracy single-point DFT energies and semi-empirical quasi-harmonic free energies. Of these ensembles, 534 were further refined with DFT geometry optimizations.

GEOM addresses two key gaps in the dataset literature. First, the data can be used to benchmark new models that take conformers as input to predict experimental properties, such as biological assay results for antiviral activity, or physicochemical and physiological properties. Such models could not be trained on the above molecular datasets, which contain only 2D graphs or single 3D structures. Some datasets provide single 3D structures for hundreds of thousands of molecules[36,40,41], but do not include a full ensemble for each species. Others contain a continuum of high-quality 3D structures for each species, but only contain hundreds of molecules[42–47]. Yet others contain conformers for tens of thousands of molecules with experimental data[48], but the conformers are of force-field quality (see below). GEOM is unique in its size, number of conformers per species, conformer quality, and connection with experiment.

Second, GEOM can be used to train generative models to predict conformers given an input molecular graph. This is an active area of research that seeks to lower the computation cost compared to exhaustive torsional approaches and to increase the speed, reliability and accuracy compared to stochastic approaches[49–55]. The size and simulation accuracy of the GEOM dataset make it an ideal training set and for pre-training generalizable models. Moreover, machine learning models for conformer generation are orders of magnitude faster than the methods used to generate GEOM. Hence models trained on GEOM may be able to reproduce its accuracy on unseen molecules at a fraction of the cost. As discussed below, the CREST ensembles have high coverage of the true thermally accessible conformers. Hence GEOM is an excellent benchmark for the recall and diversity of conformer generation methods. However, the CREST statistical weights for each conformer are rather inaccurate. Therefore, benchmarks that include conformer probabilities should use the DFT weights provided in GEOM.

Table 1 provides summary statistics of the molecules that make up the dataset. The drug-like molecules from AICures are generally medium-sized organic compounds, containing an average of 44.4 atoms (24.9 heavy

Target	Species	Hits	Sources
SARS-CoV-2	5,832	101	[78]
SARS-CoV-2 3CL protease	817	78	[79]
SARS-CoV 3CL protease	289,808	447	[80]
SARS-CoV PL protease	232,708	696	[56,57]
E. Coli	2,186	111	[3,81]
Pseudomonas aeruginosa	1,968	48	[78]

Table 2. Experimental data for GEOM species from AICures.

Category	Dataset	Property	Tasks	Species	Recovered	Sources
Physical chemistry	ESOL	Water solubility	1	1,113	99.6%	[28]
	FreeSolv	Hydration free energy	1	642	100.0%	[29]
	Lipophilicity	log $K_{octanol-water}$	1	4,194	99.9%	[24,101]
Biophysics	BACE	BACE-1 inhibition	1	1,511	99.9%	[39]
Physiology	BBBP	Blood-brain barrier penetration	1	1,959	99.2%	[102]
	Tox21	Qualitative toxicity	12	7,677	98.0%	[103]
	ToxCast	Qualitative toxicity	617	8,405	98.0%	[104]
	SIDER	Drug side effects	27	1,356	95.1%	[105]
	ClinTox	Toxicity of failed, approved drugs	2	1,438	98.7%	[106,107]

Table 3. Experimental data for GEOM species from MoleculeNet[31]. "Species" denotes the number of MoleculeNet compounds that have CREST CREs in vacuum. "Recovered" gives this quantity as a percentage of the original number of compounds in MoleculeNet. The original numbers in each dataset, used to compute the "recovered" percentage, are slightly different than in ref. [31]. This is because several of the original compounds were found to be identical after SMILES pre-processing and conversion to InChi keys. Note that 1,511 BACE species (99.9%) also have CREST CREs in water.

atoms), up to a maximum of 181 atoms (91 heavy atoms). They contain a large variance in flexibility, as demonstrated by the mean (6.5) and maximum (53) number of rotatable bonds. 15% (45,712) of the molecules have specified stereochemistry, while 27% (83,326) have stereocenters but may or may not have specified stereochemistry. The QM9 dataset is limited to 9 heavy atoms (29 total atoms), with a much smaller molecular mass and few rotatable bonds. 72% (95,734) of the species have specified stereochemistry.

Table 2 summarizes the experimental properties in the GEOM dataset from the AICures dataset. Of note is data for the inhibition of SARS-CoV-2, and for the specific inhibition of the SARS-CoV-2 3CL protease. The 3CL protease has high sequence similarity to its SARS-CoV 3CL counterpart, for which there is significantly more experimental data. The similarity of the two proteases means that CoV-2 models may benefit from pre-training with CoV data, so GEOM can also be used to benchmark transfer learning methods. Another target of interest is the SARS-CoV PL protease[56,57]. The dataset also contains molecules screened for growth inhibition of *E. Coli* and *Pseudomonas aeruginosa*, both of which can cause secondary infections in COVID-19 patients.

Table 3 shows the species from MoleculeNet[31] that are included in GEOM. We used every compound from the physical chemistry and physiology categories. These molecules have experimental data for three physical chemistry tasks and 659 physiology tasks. The latter include blood-brain barrier penetration, qualitative toxicity, and whether a drug fails in clinical trials due to toxicity. GEOM also contains the BACE dataset[39], which is part of the biophysics category of MoleculeNet. Each BACE molecule has an experimental binding affinity for human β-secretase 1 (BACE-1). The remaining biophysics datasets were excluded because of size, and because the AICures drug dataset is already sufficiently large. The "recovered" column in Table 3 shows that vacuum conformer-rotamer ensembles (CREs) were generated for over 98% of the molecules in each dataset other than SIDER. CREST CREs were also generated with an implicit solvent model of water for 99.9% of the BACE compounds. As mentioned above, these conformers were further annotated with single-point DFT energies and xTB quasi-harmonic free energies.

GEOM contains vacuum CREs for 98% of the original molecules in all but one of the datasets within MoleculeNet. This means that future models using the CREs can be benchmarked against past predictions from 2D and single-conformer models[31]. Care should still be taken when making such comparisons, as the missing molecules have similar characteristics, and may therefore bias the resulting data. For example, many missing compounds are extremely flexible. For most of these compounds, the CREST calculations ran for several days with 40 cores and did not finish. Other missing compounds failed during initial xTB optimization, often because of unusual topologies; this was most common in the SIDER dataset.

Methods

CREST. Generation of conformers ranked by energy is computationally complex. Many exhaustive, stochastic, and Bayesian methods have been developed to generate conformers[58–65]. The exhaustive method is to enumerate all the possible rotations around every bond, but this approach has prohibitive exponential scaling with the number of rotatable bonds[60,66]. Stochastic algorithms available in cheminformatics packages such as RDKit[64] suffer from two flaws. First, they explore conformational space very sparsely through a combination of pre-defined distances and stochastic samples[67] and can miss many low-energy conformations. Second, in most standalone applications, conformer energies are determined with classical force fields, which are rather inaccurate[47]. Enhanced molecular dynamics simulations, such as metadynamics (MTD), can sample conformational space more exhaustively, but need to evaluate an energy function many times. *Ab initio* methods, such as DFT, can assign energies to conformers more accurately than force fields, but are also orders of magnitude more computationally demanding.

An efficient balance between speed and accuracy is offered by the newly developed CREST software[37]. This program uses semi-empirical tight-binding DFT to calculate the energy. The predicted energies are significantly more accurate than classical force fields, accounting for electronic effects, rare functional groups, and bond-breaking/formation of labile bonds, but are computationally less demanding than full DFT. Moreover, the search algorithm is based on MTD, a well-established thermodynamic sampling approach that can efficiently explore the low-energy search space. Finally, the CREST software identifies and groups rotamers, conformers that are identical except for atom re-indexing. It then assigns each conformer a probability through

$$p_i^{\text{CREST}} = \frac{d_i \exp(-E_i/k_{\text{B}}T)}{\sum_j d_j \exp(-E_j/k_{\text{B}}T)}. \tag{1}$$

Here p_i is the statistical weight of the ith conformer, d_i is its degeneracy (i.e., how many chemically and permutationally equivalent rotamers correspond to the same conformer), E_i is its energy, k_{B} is the Boltzmann constant, T is the temperature, and the sum is over all conformers. Equation (1) is an approximation to the true probability, $p_i \propto \exp(-G_i/k_{\text{B}}T)$, where G is the free energy [Eqs. (3–4)]. The solvation free energy can be incorporated into E with a solvent model, but the translation, rotation, and vibrational free energies are missing. The addition of these terms is discussed below.

To generate conformers and rotamers, CREST takes a geometry as input and uses its flexibility to determine an MTD simulation time t_{max} (between 5 and 200 ps). The initial structure is deformed by propagating Newton's equations of motion with an NVT thermostat[68] from time $t = 0$ to t_{max}. The potential at each time step is given by the sum of the GFN2-xTB potential energy and a bias potential,

$$V_{\text{bias}} = \sum_i^n k_i \exp(-\alpha_i \Delta_i^2), \tag{2}$$

which forces the molecule into new conformations. The collective variables Δ_i are the root-mean-square displacements (RMSDs) of the structure with respect to the ith reference structure, n is the number of reference structures, k_i is the pushing strength and α_i determines the potentials' shapes. A new reference structure from the trajectory is added to V_{bias} every 1.0 ps, driving the molecule to explore new conformations. Different molecules require different (k_i, α_i) pairs to produce best results, so twelve different MTD runs are used with different settings for the V_{bias} parameters.

Conformers are defined by rotation about dihedral angles. In MTD simulations with RMSD collective variables, the biasing potential in Eq. (2) generates energy for overcoming torsional barriers. Since it takes less energy to cross a rotational barrier than to break a covalent bond, the biasing term leads to exploration of conformational space through rotation, rather than trivial fragmentation[37,68]. Indeed, the bare energy without the biasing term keeps the molecule from exploring ultra-high energy regions, and thus reduces the size of the $3N$-6-dimensional PES to be explored, where N is the number of atoms. This also makes it more efficient at finding accessible minima than an exhaustive enumeration of dihedral angles, since the latter would include high-energy, thermally inaccessible structures.

Geometries from the MTD runs are then optimized with GFN2-xTB. Conformers are identified as structures with $\Delta E > E_{\text{thr}}$, RMSD > RMSD$_{\text{thr}}$, and $\Delta B_e > B_{\text{thr}}$, where ΔE is the energy difference between structures, ΔB_e is the difference in their rotational constants, and thr denotes a threshold value. Rotamers are identified through $\Delta E > E_{\text{thr}}$, RMSD > RMSD$_{\text{thr}}$, and $\Delta B_e < B_{\text{thr}}$. Duplicates are identified through $\Delta E < E_{\text{thr}}$, RMSD < RMSD$_{\text{thr}}$, and $\Delta B_e < B_{\text{thr}}$. The defaults, which are used in this work, are $E_{\text{thr}} = 0.1$ kcal/mol, RMSD$_{\text{thr}} = 0.125$ Å, and $B_{\text{thr}} = 15.0$ MHz. Conformers and rotamers are added to the CRE and duplicates are discarded.

If a new conformer has a lower energy than the input structure, the procedure is restarted using the conformer as input, and the resulting structures are added to the CRE. The procedure is restarted between one and five times. The three conformers of lowest energy then undergo two normal molecular dynamics (MD) simulations at 400 K and 500 K. These are used to sample low-energy barrier crossings, such as simple torsional motions, which are needed to identify the remaining rotamers. Conformers and rotamers are once again identified and added to the CRE. All accumulated structures are then used as inputs to a genetic Z-matrix crossing algorithm[68,69], the results of which are also added to the CRE. All geometries accumulated throughout the sampling process are optimized with a tight convergence threshold, identified as conformers, rotamers or duplicates, and sorted to yield the final set of structures. The process is restarted after the regular MD runs or the tight optimization if any conformers have lower energy than the input, with no limit to the number of restarts. The final CRE contains conformers and rotamers up to a maximum energy E_{win}. The default $E_{\text{win}} = 6.0$ kcal/mol provides a

safety net around errors in the xTB energies, as only conformers with $E \lesssim 2.5$ kcal/mol have significant population at room temperature.

CREST generates ensembles with good coverage of the true CREs. For example, ref. [37] compared the experimental conformations of gas-phase citronellal, inferred through microwave spectroscopy[70], with computational predictions. Each of the 15 lowest-energy experimental conformers was found in the CREST ensemble. The ^{1}H-NMR spectrum was then computed in chloroform using CREST conformers, together with DFT for energy re-ranking and computation of the coupling and shielding constants. The spectrum with the ensemble matched experiment far better than with only one conformer[37]. Investigation of macrocycles, a protonated peptide, metal-organic systems, and the 1-Naphthol dimer yielded similarly good results.

DFT. CREST offers an excellent balance between cost and accuracy for generating an initial CRE. The GFN2-xTB method is fast enough to be used in long MTD runs, and its conformational energies are accurate to within 2 kcal/mol (see Technical Validation). The number of energy and force calculations can easily reach into the millions for a single CREST run, making full DFT prohibitive and xTB quite practical. Further, the CREST safety window of 6.0 kcal/mol ensures that the vast majority of accessible conformers should be present in the CRE. However, the typical xTB errors of 2 kcal/mol are too large for the accurate ranking of the conformers by statistical weight. This is because p is exponential in $\Delta E/k_B T$, and at room temperature $k_B T = 0.59$ kcal/mol, which is 3.4 times smaller than the average error. Further, the weights do not take into account the zero-point energy or the roto-translational and vibrational entropy (see below). Each of these contributions to the free energy is conformation-dependent, and can lead to non-negligible changes in statistical weight.

DFT can be used to optimize conformers and compute their relaxed energies. However, each ensemble can contain hundreds of conformers, which makes DFT optimization extremely resource-intensive. Further, a Hessian calculation is required to compute the zero-point energy and entropic corrections to the free energy. Such calculations are among the most computationally demanding in quantum chemistry. Thus a full DFT optimization of each ensemble, together with an accurate free energy calculation, is a daunting task.

To address these issues, the developers of CREST recently introduced the CENSO program[71]. CENSO uses a series of optimizations at increasingly accurate levels of DFT theory. The free energy cutoff for discarding conformers is reduced at each stage, leading to fewer conformers in each successive round. Further, CENSO uses the recently developed r2scan-3c meta-GGA functional[72] for the final optimization. r2scan-3c with the custom-made mTZVPP basis set is extremely accurate, yielding conformational energies that are within 0.3 kcal/mol of the CCSD(T) complete basis set limit[72]. It is also quite affordable given its accuracy, with a cost that is 100–1000 times lower than hybrid functionals with large basis sets[72]. The optimization is further accelerated by discarding duplicate conformers and high-energy geometries that are close to converged[71]. Lastly, CENSO computes entropic and zero-point corrections using the new biased Hessian method[73]. This technique uses xTB, which is quite computationally affordable, together with an extra biasing potential. The biasing potential accounts for energy differences between xTB and DFT, which allows xTB Hessians to be computed for DFT-optimized geometries.

The statistical weight computed by CENSO for the ith conformer is

$$p_i^{\text{CENSO}} = \frac{\exp(-G_i/k_B T)}{\sum_j \exp(-G_j/k_B T)},$$

(3)

where G_i is the conformation-dependent free energy. Note that unlike CREST, CENSO does not include rotamer degeneracy in the calculation of p. The reason is that accounting for all rotamers, though attempted by CREST, is still a difficult task, and the difference in rotamers among different conformers should be small. The free energy is given by[71]:

$$G_i = E_{\text{gas}}^{(i)} + \delta G_{\text{solv}}^{(i)}(T) + G_{\text{trv}}^{(i)}(T).$$

(4)

Here E_{gas} is the gas phase energy, $\delta G_{\text{solv}}(T)$ is the solvation free energy, and $G_{\text{trv}}(T)$ is the free energy due to translation, rotation and vibration. δG_{solv} can be calculated with implicit solvent methods such as COSMO-RS[74,75] or C-PCM[76]. The solvation free energy predicted by r2scan-3c/COSMO-RS is typically accurate to within 0.5 kcal/mol[71]. Given the Hessian matrix and the associated normal modes, $G_{\text{trv}}(T)$ can be computed within the standard modified rigid-rotor harmonic-oscillator approximation[77]. This term can be predicted quite accurately, with sub-chemical accuracy attainable even for semi-empirical methods[71].

CENSO qualitatively reproduces the optical rotation of organic molecules measured in solution, which is a challenging task that depends sensitively on the CRE[71]. Further, it makes very accurate predictions of the octanol-water partition coefficients and pK_a values of various organic molecules[71]. Conformers and statistical weights generated by CENSO are thus quite reliable.

In this work we apply CENSO to 534 species, yielding the highest-accuracy ensembles ever generated for drug-like molecules. Calculations are performed in implicit water solvent for 35% of the molecules in the BACE dataset[31,39], which contains experimental binding affinities for inhibitors of BACE-1 (Table 3). Binding affinity models that incorporate CREs can be trained with this data. Models trained on a single conformer can also benefit from the CENSO ensembles. Since many of the drug-like molecules are quite flexible, the typical approach of optimizing a single force field conformer with DFT is likely to miss the true lowest-energy structure. Thus the lowest-energy CENSO structures are far more reliable inputs to single-conformer models. Lastly, the ensembles can be used for transfer learning (TL), so that generative models trained on the large CREST dataset can be fine-tuned with the CENSO data.

In addition to the fully optimized CREs, we provide single-point DFT energies for all 1.3 million CREST conformers in 1,511 out of 1,513 BACE species (99.9%). We also provide xTB vibrational frequencies to complete the calculation of G. Together, these calculations give statistical weights that are much more accurate than those of CREST, and somewhat less accurate than CENSO. Since nearly all BACE species have single-point calculations, future binding affinity models using the re-ranked CREs can be benchmarked against predictions from past 2D and 3D models[31]. All geometries with DFT energies are also annotated with DFT dipole moments, partial charges, and molecular orbital energies. This data can be used for multi-task learning to improve TL for conformer generation.

Conformer generation.

SMILES pre-processing. SMILES strings from the QM9 dataset were used as given. SMILES strings and properties of the drug-like molecules were accessed from ref. [78] and https://github.com/yangkevin2/coronavirus_data/tree/master/data (original sources are[3,56,57,79–81]). Each SMILES string was converted to its canonical form using RDKit. This allowed us to assign multiple properties from multiple sources to a single species, even if different non-canonical SMILES strings were used in the original sources.

3.9% of the drug molecules accessed (11,886 total) were given as clusters, either with a counterbalancing ion (e.g. ".[Na+]", ".[Cl-]") or with an acid to represent the protonated salt (e.g. ".Cl"). For non acid-base clusters we identified the compound of interest as the heaviest component of the cluster. For the acid/base SMILES strings, used reaction SMARTS in RDKit to generate the protonated molecule and counterion. This product SMILES was used in place of the original SMILES. Original SMILES strings are available in the dataset with the key uncleaned_smiles (see https://github.com/learningmatter-mit/geom for details). Not only does de-salting identify the drug-like compound in each cluster and correct its ionization state, it also homogenizes the molecular representations in the drug datasets. For MoleculeNet we also selected the heaviest component from each cluster SMILES, but did not perform protonation.

Initial structure generation. To generate conformers with CREST one must provide an initial guess geometry, ideally optimized at the same level of theory as the simulation (GFN2-xTB). For the drug molecules we therefore used RDKit to generate initial conformers from SMILES strings, optimized each conformer with GFN2-xTB, and used the lowest energy conformer as input to CREST.

Conformers were generated in RDKit using the EmbedMultipleConfs command with 50 conformers (numConfs = 50), a pruning threshold of similar conformers of 0.01 Å (pruneRmsThresh = 0.01), a maximum of five embedding attempts per conformer (maxAttempts = 5), coordinate initialization from the eigenvalues of the distance matrix (useRandomCoords = False), and a random seed. If no conformers were successfully generated then numConfs was increased to 500. Each conformer was then optimized with the MMFF force field[82] in RDKit using the default arguments. Duplicate conformers, identified as those with an RMSD below 0.1 Å, were removed after optimization. Optimization was skipped for any molecules with *cis/trans* stereochemistry (indicated by "\" or "/" in the SMILES string), as such stereochemistry is not always maintained during RDKit optimization.

The ten MMFF-optimized conformers with the lowest energy were further optimized with xTB using Orca 4.2.0[83,84]. The conformer with the lowest xTB energy was selected as the seed geometry for CREST. The QM9 molecules are already optimized with DFT, and so in principle did not need to be optimized further for CREST. However, since it is recommended to seed CREST with a structure optimized at the GFN2-xTB level of theory, we re-optimized each QM9 geometry with xTB before using it in CREST.

CREST simulation. A single xTB-optimized structure was used as input to the CREST simulation of each species. Default values were used for all CREST arguments, except for the charge of each geometry. CREST runs on the AICures drug dataset took an average of 2.8 hours of wall time on 32 cores on Knights Landing (KNL) nodes (89.1 core hours), and 0.63 hours on 13 cores on Cascade Lake and Sky Lake nodes (8.2 core hours). QM9 jobs were only performed on the latter two nodes, and took an average of 0.04 wall hours on 13 cores (0.5 core hours). 13 million KNL core hours and 1.2 million Cascade Lake/Sky Lake core hours were used in total.

CREST calculations on MoleculeNet species were run across several compute clusters, each with various node types and different core counts per node. KNL nodes were not used. Excluding species already present in the AICures dataset, each MoleculeNet job took 6.3 hours of wall time using 18.1 cores on average. These values are skewed by extremely flexible molecules whose CREST jobs took several days to finish: the *median* wall time was 1.4 hours, and the median core count was 12.0. 1.5 million CPU hours were used in total.

Graph re-identification. It was necessary to re-identify the graph of each conformer generated by CREST, for the following reasons. First, stereochemistry may not have been specified in the original SMILES string, but necessarily existed in each of the generated 3D structures. Second, reactivity such as dissociation or tautomerization may have occurred in the CREST simulations (CREST has specific commands to generate tautomers, but they were not used here). This would also lead to conformers with different graphs.

To re-identify the graphs we used xyz2mol[85] (code accessed from https://github.com/jensengroup/xyz-2mol) to generate an RDKit mol object. These mol objects were used to assign graph features to each conformer (see Data Records). It should be noted that xyz2mol sometimes assigned resonance structure graphs instead of the original graphs. In some cases this caused different conformers of the same species to have different graphs. This happened, for example, when the conformers had different *cis/trans* isomerism about a double bond that was only present because of the resonance structure used (see the RDKit tutorial at https://github.com/learningmatter-mit/geom). This is conceptually different from species whose conformer graphs differ because of reactivity. One may want to distinguish these two cases when analyzing the conformer mol objects. We also note that CREST changed the atom ordering of the input geometry, and hence of the subsequent conformers. This

AICures drug dataset			
	Mean	**Std. deviation**	**Maximum**
S (cal/mol K)	8.2	2.6	16.8
-G (kcal/mol)	2.4	0.8	5.0
$\langle E \rangle$ (kcal/mol)	0.4	0.2	2.4
Conformers	102.6	159.1	7,451
QM9 dataset			
	Mean	**Std. deviation**	**Maximum**
S (cal/mol K)	3.9	2.8	14.2
-G (kcal/mol)	1.2	0.8	4.2
$\langle E \rangle$ (kcal/mol)	0.2	0.2	2.2
Conformers	13.5	42.2	1,101

Table 4. CREST-based statistics for the QM9 and AICures drug datasets.

means that, even if a conformer did not react, we could not simply create an RDKit mol object with its canonical SMILES and set its coordinates.

CENSO simulation. 534 molecules from the BACE dataset (35%) were optimized with CENSO. Initial CREs were generated with CREST using the ALPB model for water[86]. The CREs were refined with CENSO 1.1.2, using Orca 5.0.1[87] to perform the DFT calculations. The C-PCM[76] model of water was used for DFT and the ALPB model was used for xTB. Conformer and rotamer duplicates were removed throughout the optimization using CREST (`crestcheck = "on"`). Default values were used for all other parameters. We used the same clusters and nodes for CENSO as for CREST with MoleculeNet species. The average CENSO job took 1 day and 4 hours of wall time using 54 cores. 781,000 CPU hours were used in total.

Single point calculations. We performed single-point DFT calculations on all CREST conformers in the BACE dataset without further optimization. We used Orca version 5.0.2 and the same level of theory as in CENSO optimization (r2scan-3c functional, mTZVPP basis, C-PCM model of water, and default grid 2). The average run took 6.4 minutes of wall time using 8 cores. Calculations took a total of 1.1 million CPU hours for 1.3 million conformers.

Hessian calculations. We performed Hessian calculations on all CREST conformers in the BACE dataset, using xTB with the ALPB model for water. The average run took 41 seconds of wall time using 4 cores. Calculations took a total of 63,000 CPU hours for 1.3 million conformers.

Conformational property prediction. The GEOM dataset is significant because it allows for the training of conformer-based property predictors and generative models to predict new conformations. The first application will be explored in a future publication. The second application is necessary for using conformer-based ML models in practice, since generating CREST structures from scratch is too costly for the virtual screening of new species. Such work is already underway[88], paving the way for `graph → conformer ensemble → property` models that can be trained end-to-end. Here we give an example of a simpler application in the same vein, benchmarking methods to predict *summary statistics* of each conformer ensemble, rather than the conformers themselves. Our proposed tasks are similar to the benchmark QM9 tasks, which measure a model's ability to predict properties that are uniquely determined by geometry. Here, since we provide conformer ensembles for each species, we measure a model's ability to predict properties defined by the ensemble. Because one chemical graph spawns a unique conformer ensemble, these tasks are also a metric of the performance of graph-based models to infer properties mediated through conformational flexibility.

We trained different models to predict three quantities related to conformational information. A summary of these quantities can be found in Table 4 and Fig. 2. The first quantity is the conformational free energy, $G = -TS$, where the ensemble entropy is $S = -R\sum_i p_i \log p_i$[37]. Here the sum is over the statistical probabilities p_i of the ith conformer, and R is the gas constant. The conformational entropy is a measure of the conformational degrees of freedom available to a molecule. A molecule with only one conformer has an entropy of exactly 0, while a molecule with equal statistical weight for an infinite number of conformers has infinite conformational entropy. The conformational Gibbs free energy is an important quantity for predicting the binding affinity of a drug to a target. The affinity is determined by the change in Gibbs free energy of the molecule and protein upon binding, which includes the loss of molecular conformational free energy[89]. The second quantity is the average conformational energy. The average energy is given by $\langle E \rangle = \sum_i p_i E_i$, where E_i is the energy of the ith conformer. Each energy is defined with respect to the lowest-energy conformer. The third quantity is the number of unique conformers for a given molecule, as predicted by CREST within the default maximum energy window[37].

We trained a kernel ridge regression (KRR) model[90], a random forest[91], and three different neural networks to predict conformer properties. The random forest, KRR and feed-forward neural network (FFNN) models were trained on Morgan fingerprints[92] generated through RDKit. Two different message-passing neural networks[93] were trained. The first, called ChemProp, has achieved state-of-the-art performance on a number of benchmarks[20]. The second is based on the SchNet force field model[94,95]. We call it SchNetFeatures, as it learns
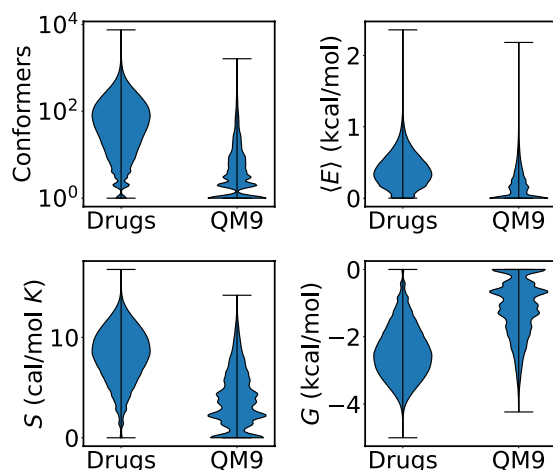
Fig. 2 Violin plots of CREST-based statistics for the QM9 and AICures drug datasets.

Model	G (kcal/mol)	E (kcal/mol)	ln(unique conformers)
Random Forest	0.406	0.166	0.763
KRR	0.289	0.131	0.484
FFNN	0.274	0.119	0.455
ChemProp	0.225	**0.110**	0.380
SchNetFeatures	**0.203**	0.113	**0.363**

Table 5. Prediction mean absolute error (MAE) for three conformer-related properties. Models were trained and tested on the AICures drug dataset.

from 3D geometries using the SchNet architecture, but also incorporates graph-based node and bond features. The SchNetFeatures models were trained on the highest-probability conformer of each species.

100,000 species were sampled randomly from the AICures drug subset of GEOM. We used the same 60-20-20 train-validation-test split for each model. The splits, trained models, and log files can be found at[96], under the heading "synthetic". Hyperparameters were optimized for each model type and for each task using the hyperopt package. Details of the hyperparameter searches, optimal parameters, and network architectures can be found in the same location as the models. Source code is available at https://github.com/learningmatter-mit/NeuralForceField.

Results are shown in Table 5. ChemProp and SchNetFeatures are the strongest models overall, followed in order by FFNN, KRR, and random forest. Of the three models that use fixed 2D fingerprints, we see that the FFNN is best able to map these non-learnable representations to properties. ChemProp has the added flexibility of learning an ideal molecular representation directly from the graph, and so performs even better than the FFNN. The SchNetFeatures model retains this flexibility while incorporating extra information from one 3D structure. Compared to ChemProp, its prediction error is 10% lower for G, nearly equal for $\langle E \rangle$, and 5% lower for ln(unique conformers). The small improvement in performance is not surprising, as the ensemble properties are mainly determined by molecular flexibility, which is a function of the graph through the number of rotatable bonds. A single 3D geometry would not provide extra information about this flexibility.

We see that various models can accurately predict conformer properties when trained on the GEOM dataset. With access to the dataset, researchers will therefore be able to predict results of expensive simulations without performing them directly. This has implications beyond ensemble-averaged properties, as generative models trained on the GEOM dataset will also be able to produce the conformers themselves[88].

Data Records

The dataset is available online at[97], and detailed tutorials for loading and analyzing the data can be found at https://github.com/learningmatter-mit/geom.

The data is available either through MessagePack, a language-agnostic binary serialization format, or through Python pickle files. There are two MessagePack files for the AICures drug dataset and two for QM9. Each of the two files contains a dictionary, where the keys are SMILES strings and the values are sub-dictionaries. In the file with suffix `crude`, the sub-dictionaries contain both species-level information (experimental binding data, average conformer energy, etc.) and a list of dictionaries for each conformer. Each conformer dictionary has its own conformer-level information (geometry, energy, degeneracy, etc.). In the file with suffix `featurized`, each conformer dictionary contains information about its molecular graph.

The Python pickle files are organized in a different fashion. The main folder is divided into sub-folders for QM9, AICures, and MoleculeNet data, plus separate folders for BACE calculations in water and with CENSO.
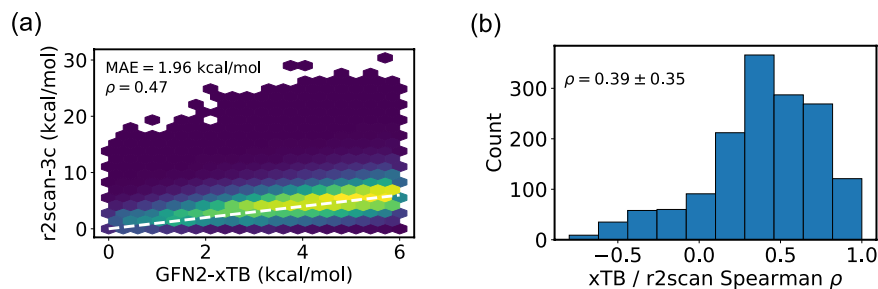
(a)

(b)

Fig. 3 Comparison of GFN2-xTB (CREST) energies and single-point r2scan-3c (DFT) energies. (**a**) xTB vs. r2scan-3c energies for all geometries in the BACE-1 dataset. The ideal correlation is shown with a dashed white line. (**b**) Distribution of Spearman rank correlation coefficients ρ, measuring the accuracy of xTB energy ranking for each of the ensembles.

Each sub-folder contains one pickle file for each species. Each pickle file contains both summary information and conformer information for its species. Each conformer is stored as an RDKit `mol` object, so that it contains both the geometry and graph features. One may only want to load the pickle files of species with specific properties (e.g., those with experimental data for SARS-CoV-2 inhibition); for this one can use the summary `JSON` file. This file contains all summary information along with the path to the pickle file, but without the list of conformers. It is therefore lightweight and quick to load, and can be used to choose species before loading their pickles.

Technical Validation

The quality of the data was validated in three different ways. First, we checked that the conformer data was accurately parsed from the CREST calculations. To do so we randomly sampled one conformer from 20 different species and manually confirmed that its data matched the data in the CREST output files.

Second, we re-identified the graphs of the conformers generated by CREST using `xyz2mol`. The graph re-attribution procedure succeeded for 88.4% of the QM9 molecules and 94.7% of the drug molecules, recovering the original molecular graph that was used to generate each conformer. Note that to compare graphs we removed stereochemical indicators from the original and the re-generated graph. This was done because of cases in which stereochemistry was not specified originally but was specified in the generated conformers. All of the failed QM9 graphs underwent some sort of reaction, which can be explained by the presence of highly strained and unstable molecules. However, manual inspection of 53 cases in the AICures drug dataset suggests that 70% of the drug graphs failed only because of poor handling of resonance forms by `xyz2mol` (see above). This means that the original graph was likely recovered for 98.4% of all drugs. 21% of cases failed because of tautomerization (1% of all cases), and 9.4% failed because of a different reaction (usually dissociation or ring formation; 0.5% of all cases). The high success rate of the graph re-identification indicates that, in the vast majority of cases, the geometries generated by CREST were actual conformers of the species.

Third, we compared the CREST energies and coordinates to those from higher levels of theory. Figure 3 compares the GFN2-xTB calculations of CREST with single-point r2scan-3c calculations, both performed in water for 1,511 species in the BACE dataset. Panel (a) shows the relative energies of the two methods. The mean absolute error (MAE) of xTB is 1.96 kcal/mol, which is similar to reported values in conformational energy benchmarks[38]. The ranking accuracy can be measured with the Spearman correlation coefficient ρ, which lies between 1 and −1 (perfect correlation and anti-correlation, respectively). The Spearman coefficient is 0.47 when using all geometries from all species. However, it is more meaningful to judge the energy rankings among different conformers in a single species. Computing ρ separately for each species yields the distribution in panel (b). The distribution of ρ is quite wide, with an average value of 0.39 and a standard deviation of 0.35. The mean value of ρ indicates moderate correlation between the methods. The correlation is significantly better than for classical force fields such as MMFF94[82], UFF[98], and GAFF[99]: For instance, the median ρ between MMFF94 and single-point DFT for drug-like molecules is between −0.1 and −0.45, meaning that the two methods are actually weakly *anti-correlated* (Supporting Information of ref. [47]).

Figure 4 compares single-point DFT calculations on CREST geometries ("SP") with DFT results on fully optimized geometries ("CENSO"). Panel (a) shows the distribution of ρ for conformer energies. The average Spearman correlation is 0.69 and the standard deviation is 0.27, indicating good agreement between the two methods. Indeed, the MAE between optimized and single-point relative energies is 0.54 kcal/mol, which is 3.6 times lower than the xTB error (the MAE of the absolute energy, equal to the average energy released after optimization, is 5.74 kcal/mol). Panel (b) shows that the geometries change very little during optimization, with a mean RMSD of only 0.36 Å. This shows that the CREST geometries are quite good, thus validating the quality of the GEOM ensembles. The median RMSD among heavy atoms is 0.25 Å; this is 2.4 times lower than the value of 0.6 Å between MMFF94 and PM7 geometries[100] for drug-like molecules[47].

Similar comparisons can be made between CENSO geometries and their most similar CREST counterparts (i.e., the CREST geometry with the lowest RMSD relative to a CENSO geometry). These may not be the same as the CREST geometries used to seed the optimization. We have found that using the most similar geometries does not significantly affect the results; for example, the Spearman coefficient only climbs to 0.72 ± 0.27, while the RMSD only drops to 0.33 ± 0.19. Note also that the comparison of the methods only includes conformers with non-negligible weight after optimization ($\Delta G \leq 2.5$ kcal/mol), since CENSO discards high-energy
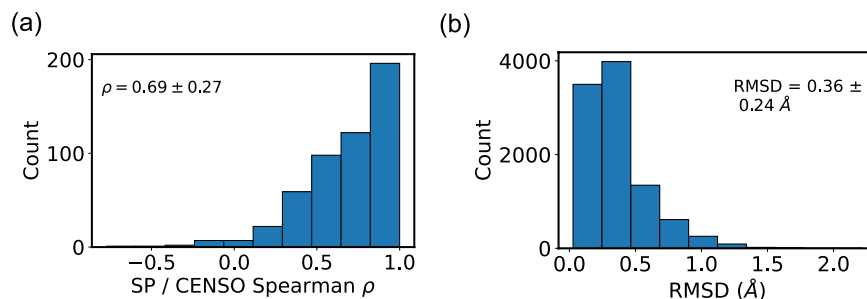
(a)

(b)

Fig. 4 Comparison of CENSO and single-point DFT calculations. (**a**) Distribution of Spearman coefficients, measuring the accuracy of single-point ranking for each of the ensembles. (**b**) RMSDs between CREST geometries and DFT-optimized geometries.
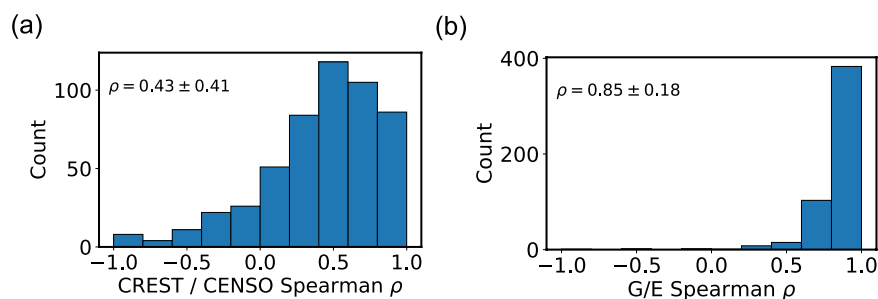
(a)

(b)

Fig. 5 (**a**) Comparison of CENSO and CREST calculations. The distribution of Spearman coefficients shows the accuracy of CREST ranking for each of the ensembles. (**b**) Comparison of energy and free-energy ranking with CENSO. The distribution of Spearman coefficients shows the accuracy of energy ranking for each of the ensembles.

conformers during optimization. Hence high-energy conformers were not fully optimized and thus not included in the comparison.

Figure 5(a) compares the ordering of geometries with CREST and with CENSO. The Spearman correlation is $\rho = 0.43 \pm 0.41$, which is similar to the correlation between CREST and single-point energies. This result should be interpreted with caution, however, since only the lowest-energy CENSO geometries are included in the comparison, whereas the rank correlation in Fig. 3 includes all CREST conformers. Lastly, Fig. 5(b) compares the ordering of CENSO geometries by energy and by free energy. The correlation is quite high ($\rho = 0.85 \pm 0.18$), and the MAE between energies and free energies is only 0.33 kcal/mol. Hence energies alone can be quite good for ordering conformers by statistical weight. This also means that the statistical weight errors in GEOM are dominated by xTB errors, and that the quasi-harmonic errors are comparably negligible.

Usage Notes
Researchers are encouraged to use the data-loading tutorials given in https://github.com/learningmatter-mit/geom. We suggest loading the data through the RDKit pickle files, as RDKit mol objects are easy to handle and their properties can be readily analyzed. The MessagePack files, while secure and accessible in all languages, represent graphs through their features rather than objects with built-in methods, and are thus more difficult to analyze. To train 3D-based models we suggest following the tutorial and README file at https://github.com/learningmatter-mit/NeuralForceField.

Code availability
Tutorials for loading the dataset and code for training 3D-based neural network models are publicly available without restriction (https://github.com/learningmatter-mit/geom and https://github.com/learningmatter-mit/NeuralForceField). CREST and xTB are both freely available online (https://github.com/grimme-lab/crest/releases and https://github.com/grimme-lab/xtb/releases). CREST version 2.9 was used with xTB version 6.2.3 to generate the initial CREs. CENSO 1.1.2 was used with Orca 5.0.1[87] and xTB 6.4.1 to refine the ensembles. Orca 5.0.2 was used for all single-point calculations. A race condition bug in version 5.0.1 meant that some CENSO energies were clearly incorrect (conformational energies above 1,000 kcal/mol), while some energy calculations failed to converge for reasonable geometries. Therefore, we discarded ensembles with failed energy calculations or conformational energy ranges exceeding 30 kcal/mol at any stage of the optimization. We also performed new single-point calculations on all converged CENSO geometries with Orca 5.0.2; 0.44% of the energies were found to be incorrect and were replaced.

References

1. Norinder, U., Lidén, P. & Boström, H. Discrimination between modes of toxic action of phenols using rule based methods. *Molecular diversity* **10**, 207–212, https://doi.org/10.1007/s11030-006-9019-3 (2006).
2. Durrant, J. D. & McCammon, J. A. Molecular dynamics simulations and drug discovery. *BMC biology* **9**, 1–9, https://doi.org/10.1186/1741-7007-9-71 (2011).
3. Stokes, J. M. *et al.* A deep learning approach to antibiotic discovery. *Cell* **180**, 688–702, https://doi.org/10.1016/j.cell.2020.01.021 (2020).
4. Gómez-Bombarelli, R. *et al.* Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nature Materials* **15**, 1120–1127, https://doi.org/10.1038/nmat4717 (2016).
5. Zhavoronkov, A. *et al.* Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature biotechnology* **37**, 1038–1040, https://doi.org/10.1038/s41587-019-0224-x (2019).
6. Schwalbe-Koda, D. & Gómez-Bombarelli, R. Generative models for automatic chemical design. In *Machine Learning Meets Quantum Physics*, 445–467 https://doi.org/10.1007/978-3-030-40245-7_21 (Springer, 2020).
7. Gómez-Bombarelli, R. *et al.* Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Central Science* **4**, 268–276, https://doi.org/10.1021/acscentsci.7b00572 (2018).
8. Jin, W., Barzilay, R. & Jaakkola, T. Junction Tree Variational Autoencoder for Molecular Graph Generation. In *International Conference on Machine Learning*, https://proceedings.mlr.press/v80/jin18a.html (2018).
9. Dai, H., Tian, Y., Dai, B., Skiena, S. & Song, L. Syntax-directed variational autoencoder for structured data. In *International Conference on Learning Representations*, https://openreview.net/forum?id=SyqShMZRb (2018).
10. Noé, F., Olsson, S., Köhler, J. & Wu, H. Boltzmann generators: Sampling equilibrium states of many-body systems with deep. *Science* **365**, eaaw1147, https://doi.org/10.1126/science.aaw1147 (2019).
11. Olivecrona, M., Blaschke, T., Engkvist, O. & Chen, H. Molecular de-novo design through deep reinforcement learning. *Journal of cheminformatics* **9**, 1–14, https://doi.org/10.1186/s13321-017-0235-x (2017).
12. Gottipati, S. K. *et al.* Learning to navigate the synthetically accessible chemical space using reinforcement learning. In *International Conference on Machine Learning*, 3668–3679, https://proceedings.mlr.press/v119/gottipati20a.html (PMLR, 2020)
13. Popova, M., Isayev, O. & Tropsha, A. Deep reinforcement learning for de novo drug design. *Science Advances* **4**, eaap7885, https://doi.org/10.1126/sciadv.aap7885 (2018).
14. AlQuraishi, M. End-to-end differentiable learning of protein structure. *Cell Systems* **8**, 292–301.e3, https://doi.org/10.1016/j.cels.2019.03.006 (2019).
15. Ingraham, J., Riesselman, A., Sander, C. & Marks, D. Learning protein structure with a differentiable simulator. In *International Conference on Learning Representations*, https://openreview.net/forum?id=Byg3y3C9Km (2019).
16. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610, https://doi.org/10.1038/nature25978 (2018).
17. Coley, C. W., Barzilay, R., Jaakkola, T. S., Green, W. H. & Jensen, K. F. Prediction of organic reaction outcomes using machine learning. *ACS Central Science* **3**, 434–443, https://doi.org/10.1021/acscentsci.7b00064 (2017).
18. Duvenaud, D. K. *et al.* Convolutional networks on graphs for learning molecular fingerprints. In *Advances in Neural Information Processing Systems*, 2215–2223, https://proceedings.neurips.cc/paper/2015/file/f9be311e65d81a9ad8150a60844bb94c-Paper.pdf (2015).
19. Kearnes, S., McCloskey, K., Berndl, M., Pande, V. & Riley, P. Molecular graph convolutions: moving beyond fingerprints. *Journal of Computer-Aided Molecular Design* **30**, 595–608, https://doi.org/10.1007/s10822-016-9938-8 (2016).
20. Yang, K. *et al.* Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling* **59**, 3370–3388, https://doi.org/10.1021/acs.jcim.9b00237 (2019).
21. Anderson, B., Hy, T. S. & Kondor, R. Cormorant: Covariant molecular neural networks. In *Advances in Neural Information Processing Systems*, 14537–14546, https://proceedings.neurips.cc/paper/2019/file/03573b32b2746e6e8ca98b9123f2249b-Paper.pdf (2019).
22. Klicpera, J., Groß, J. & Günnemann, S. Directional message passing for molecular graphs. In *International Conference on Learning Representations*, https://openreview.net/forum?id=B1eWbxStPH (2019).
23. Ramsundar, B. *et al. Deep Learning for the Life Sciences* (O'Reilly Media, 2019).
24. Mendez, D. *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research* **47**, D930–D940, https://doi.org/10.1093/nar/gky1075 (2018).
25. Sterling, T. & Irwin, J. J. ZINC 15–Ligand discovery for everyone. *Journal of chemical information and modeling* **55**, 2324–37, https://doi.org/10.1021/acs.jcim.5b00559 (2015).
26. Brown, N., Fiscato, M., Segler, M. H. S. & Vaucher, A. C. GuacaMol: Benchmarking models for de novo molecular design. *J. Chem. Inf. Model.* **59**, 1096–1108, https://doi.org/10.1021/acs.jcim.8b00839 (2019).
27. Polykovskiy, D. *et al.* Molecular sets (MOSES): A benchmarking platform for molecular generation models. *Frontiers in Pharmacology* **11**, https://doi.org/10.3389/fphar.2020.565644 (2020).
28. Delaney, J. S. ESOL: Estimating aqueous solubility directly from molecular structure. *Journal of Chemical Information and Computer Sciences* **44**, 1000–1005, https://doi.org/10.1021/ci034243x (2004).
29. Mobley, D. L. & Guthrie, J. P. FreeSolv: A database of experimental and calculated hydration free energies, with input files. *Journal of Computer-Aided Molecular Design* **28**, 711–720, https://doi.org/10.1007/s10822-014-9747-x (2014).
30. Wang, R., Fang, X., Lu, Y. & Wang, S. The PDBbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *Journal of Medicinal Chemistry* **47**, 2977–2980, https://doi.org/10.1021/jm030580l (2004).
31. Wu, Z. *et al.* MoleculeNet: a benchmark for molecular machine learning. *Chemical science* **9**, 513–530, https://doi.org/10.1039/C7SC02664A (2018).
32. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling* **28**, 31–36, https://doi.org/10.1021/ci00057a005 (1988).
33. Heller, S. R., McNaught, A., Pletnev, I., Stein, S. & Tchekhovskoi, D. InChI, the IUPAC international chemical identifier. *Journal of cheminformatics* **7**, 23, https://doi.org/10.1186/s13321-015-0068-4 (2015).
34. Kuhn, B. *et al.* A real-world perspective on molecular design: Miniperspective. *Journal of medicinal chemistry* **59**, 4087–4102, https://doi.org/10.1021/acs.jmedchem.5b01875 (2016).
35. Hawkins, P. C. Conformation generation: The state of the art. *Journal of chemical information and modeling* **57**, 1747–1756, https://doi.org/10.1021/acs.jcim.7b00221 (2017).
36. Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data* **1**, 140022, https://doi.org/10.1038/sdata.2014.22 (2014).
37. Pracht, P., Bohle, F. & Grimme, S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Physical Chemistry Chemical Physics* **22**, 7169–7192, https://doi.org/10.1039/C9CP06869D (2020).

38. Bannwarth, C., Ehlert, S. & Grimme, S. GFN2-xTB—An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *Journal of chemical theory and computation* **15**, 1652–1671, https://doi.org/10.1021/acs.jctc.8b01176 (2019).

39. Subramanian, G., Ramsundar, B., Pande, V. & Denny, R. A. Computational modeling of β-secretase 1 (BACE-1) inhibitors using ligand based approaches. *Journal of chemical information and modeling* **56**, 1936–1949, https://doi.org/10.1021/acs.jcim.6b00290 (2016).

40. Gražulis, S. *et al.* Crystallography Open Database–an open-access collection of crystal structures. *Journal of applied crystallography* **42**, 726–729, https://doi.org/10.1107/S0021889809016690 (2009).

41. Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. The Cambridge structural database. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* **72**, 171–179, https://doi.org/10.1107/S2052520616003954 (2016).

42. Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1, A data set of 20 million calculated off-equilibrium conformations for organic molecules. *Scientific Data* **4**, 170193, https://doi.org/10.1038/sdata.2017.193 (2017).

43. Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chemical Science* **8**, 3192–3203, https://doi.org/10.1039/C6SC05720A (2017).

44. Smith, J. S., Nebgen, B., Lubbers, N., Isayev, O. & Roitberg, A. E. Less is more: Sampling chemical space with active learning. *Journal of Chemical Physics* **148**, 241733, https://doi.org/10.1063/1.5023802 (2018).

45. Chmiela, S. *et al.* Machine learning of accurate energy-conserving molecular force fields. *Science Advances* **3**, e1603015, https://doi.org/10.1126/sciadv.1603015 (2017).

46. Simm, G. & Hernandez-Lobato, J. M. A generative model for molecular distance geometry. In *International Conference on Machine Learning*, 8949–8958, https://proceedings.mlr.press/v119/simm20a.html (PMLR, 2020).

47. Kanal, I. Y., Keith, J. A. & Hutchison, G. R. A sobering assessment of small-molecule force field methods for low energy conformer predictions. *International Journal of Quantum Chemistry* **118**, e25512, https://doi.org/10.1002/qua.25512 (2018).

48. Bolton, E. E., Kim, S. & Bryant, S. H. PubChem3D: conformer generation. *Journal of cheminformatics* **3**, 4, https://doi.org/10.1186/1758-2946-3-4 (2011).

49. Simm, G., Pinsler, R. & Hernández-Lobato, J. M. Reinforcement learning for molecular design guided by quantum mechanics. In *International Conference on Machine Learning*, 8959–8969 https://proceedings.mlr.press/v119/simm20b.html (PMLR, 2020).

50. Stieffenhofer, M., Wand, M. & Bereau, T. Adversarial reverse mapping of equilibrated condensed-phase molecular structures. *Machine Learning: Science and Technology* **1**, 045014, https://doi.org/10.1088/2632-2153/abb6d4 (2020).

51. Imrie, F., Bradley, A. R., van der Schaar, M. & Deane, C. M. Deep generative models for 3D linker design. *Journal of chemical information and modeling* **60**, 1983–1995, https://doi.org/10.1021/acs.jcim.9b01120 (2020).

52. Mansimov, E., Mahmood, O., Kang, S. & Cho, K. Molecular geometry prediction using a deep generative graph neural network. *Scientific Reports* **9**, 1–13, https://doi.org/10.1038/s41598-019-56773-5 (2019).

53. Chan, L., Hutchison, G. R. & Morris, G. M. Bayesian optimization for conformer generation. *Journal of Cheminformatics* **11**, 32, https://doi.org/10.1186/s13321-019-0354-7 (2019).

54. Gebauer, N., Gastegger, M. & Schütt, K. Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules. In *Advances in neural information processing systems*, **32**, https://proceedings.neurips.cc/paper/2019/file/a4d8e2a7e0d0c102339f97716d2fdfb6-Paper.pdf (2019).

55. Wang, W. & Gómez-Bombarelli, R. Coarse-graining auto-encoders for molecular dynamics. *npj Computational Materials* **5**, 125, https://doi.org/10.1038/s41524-019-0261-5 (2019).

56. Engel, D. qHTS of yeast-based assay for SARS-CoV PLP. https://pubchem.ncbi.nlm.nih.gov/bioassay/485353.

57. Engel, D. qHTS of yeast-based assay for SARS-CoV PLP: Hit validation. https://pubchem.ncbi.nlm.nih.gov/bioassay/652038.

58. Vainio, M. J. & Johnson, M. S. Generating conformer ensembles using a multiobjective genetic algorithm. *Journal of chemical information and modeling* **47**, 2462–2474, https://doi.org/10.1021/ci6005646 (2007).

59. Puranen, J. S., Vainio, M. J. & Johnson, M. S. Accurate conformation-dependent molecular electrostatic potentials for high-throughput in silico drug discovery. *Journal of computational chemistry* **31**, 1722–1732, https://doi.org/10.1002/jcc.21460 (2010).

60. O'Boyle, N. M., Vandermeersch, T., Flynn, C. J., Maguire, A. R. & Hutchison, G. R. Confab-Systematic generation of diverse low-energy conformers. *Journal of cheminformatics* **3**, 1–9, https://doi.org/10.1186/1758-2946-3-8 (2011).

61. Miteva, M. A., Guyon, F. & Pierre, T. Frog2: Efficient 3D conformation ensemble generator for small compounds. *Nucleic acids research* **38**, W622–W627, https://doi.org/10.1093/nar/gkq325 (2010).

62. Vilar, S., Cozza, G. & Stefano, M. Medicinal chemistry and the molecular operating environment (MOE): application of QSAR and molecular docking to drug discovery. *Current topics in medicinal chemistry* **8**, 1555–1572, https://doi.org/10.2174/156802608786786624 (2008).

63. Hawkins, P. C., Skillman, A. G., Warren, G. L., Ellingson, B. A. & Stahl, M. T. Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. *Journal of chemical information and modeling* **50**, 572–584, https://doi.org/10.1021/ci100031x (2010).

64. RDKit: Open-source cheminformatics. http://www.rdkit.org.

65. Chan, L., Hutchison, G. R. & Morris, G. M. Bayesian optimization for conformer generation. *Journal of cheminformatics* **11**, 1–11, https://doi.org/10.1186/s13321-019-0354-7 (2019).

66. Schwab, C. H. Conformations and 3D pharmacophore searching. *Drug Discovery Today: Technologies* **7**, e245–e253, https://doi.org/10.1016/j.ddtec.2010.10.003 (2010).

67. Spellmeyer, D. C., Wong, A. K., Bower, M. J. & Blaney, J. M. Conformational analysis using distance geometry methods. *Journal of Molecular Graphics and Modelling* **15**, 18–36, https://doi.org/10.1016/S1093-3263(97)00014-4 (1997).

68. Grimme, S. Exploration of chemical compound, conformer, and reaction space with meta-dynamics simulations based on tight-binding quantum chemical calculations. *Journal of chemical theory and computation* **15**, 2847–2862, https://doi.org/10.1021/acs.jctc.9b00143 (2019).

69. Grimme, S. *et al.* Fully automated quantum-chemistry-based computation of spin–spin-coupled nuclear magnetic resonance spectra. *Angewandte Chemie International Edition* **56**, 14763–14769, https://doi.org/10.1002/anie.201708266 (2017).

70. Domingos, S. R., Pérez, C., Medcraft, C., Pinacho, P. & Schnell, M. Flexibility unleashed in acyclic monoterpenes: Conformational space of citronellal revealed by broadband rotational spectroscopy. *Physical Chemistry Chemical Physics* **18**, 16682–16689, https://doi.org/10.1039/c6cp02876d (2016).

71. Grimme, S. *et al.* Efficient quantum chemical calculation of structure ensembles and free energies for nonrigid molecules. *The Journal of Physical Chemistry A* **125**, 4039–4054, https://doi.org/10.1021/acs.jpca.1c00971 (2021).

72. Grimme, S., Hansen, A. & Ehlert, J.-M. & Mewes, J.-M. r2SCAN-3c: A "Swiss army knife" composite electronic-structure method. *The Journal of Chemical Physics* **154**, 064103, https://doi.org/10.1063/5.0040021 (2021).

73. Spicher, S. & Grimme, S. Single-point Hessian calculations for improved vibrational frequencies and rigid-rotor-harmonic-oscillator thermodynamics. *Journal of Chemical Theory and Computation* **17**, 1701–1714, https://doi.org/10.1021/acs.jctc.0c01306 (2021).

74. Klamt, A. Conductor-like screening model for real solvents: a new approach to the quantitative calculation of solvation phenomena. *The Journal of Physical Chemistry* **99**, 2224–2235, https://doi.org/10.1021/j100007a062 (1995).

75. Klamt, A., Jonas, V., Bürger, T. & Lohrenz, J. C. Refinement and parametrization of COSMO-RS. *The Journal of Physical Chemistry A* **102**, 5074–5085, https://doi.org/10.1021/jp980017s (1998).

76. Barone, V. & Cossi, M. Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model. *The Journal of Physical Chemistry A* **102**, 1995–2001, https://doi.org/10.1021/jp9716997 (1998).

77. Grimme, S. Supramolecular binding thermodynamics by dispersion-corrected density functional theory. *Chemistry–A European Journal* **18**, 9955–9964, https://doi.org/10.1002/chem.201200497 (2012).

78. Open Source Data. https://www.aicures.mit.edu/data. Accessed: 2020-05-22 (2020).

79. Main protease structure and XChem fragment screen. https://www.diamond.ac.uk/covid-19/for-scientists/Main-protease-structure-and-XChem.html. Accessed: 2020-05-22.

80. Tokars, V. & Mesecar, A. QFRET-based primary biochemical high throughput screening assay to identify inhibitors of the SARS coronavirus 3C-like Protease (3CLPro). https://pubchem.ncbi.nlm.nih.gov/bioassay/1706.

81. Zampieri, M., Zimmermann, M., Claassen, M. & Sauer, U. Nontargeted metabolomics reveals the multilevel response to antibiotic perturbations. *Cell reports* **19**, 1214–1228, https://doi.org/10.1016/j.celrep.2017.04.002 (2017).

82. Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of computational chemistry* **17**, 490–519, 10.1002/(SICI)1096-987X(199604)17:5/6<490::AID-JCC1>3.0.CO;2-P (1996).

83. Neese, F. The ORCA program system. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2**, 73–78, https://doi.org/10.1002/wcms.81 (2012).

84. Neese, F. Software update: the ORCA program system, version 4.0. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **8**, e1327, https://doi.org/10.1002/wcms.1327 (2018).

85. Kim, Y. & Kim, W. Y. Universal structure conversion method for organic molecules: from atomic connectivity to three-dimensional geometry. *Bulletin of the Korean Chemical Society* **36**, 1769–1777, https://doi.org/10.1002/bkcs.10334 (2015).

86. Ehlert, S., Stahn, M., Spicher, S. & Grimme, S. A robust and efficient implicit solvation model for fast semiempirical methods. *Journal of Chemical Theory and Computation* **17**, 4250–4261, https://doi.org/10.1021/acs.jctc.1c00471 (2021).

87. Neese, F., Wennmohs, F., Becker, U. & Riplinger, C. The ORCA quantum chemistry program package. *The Journal of Chemical Physics* **152**, 224108, https://doi.org/10.1063/5.0004608 (2020).

88. Xu, M., Luo, S., Bengio, Y., Peng, J. & Tang, J. Learning neural generative dynamics for molecular conformation generation. In *International Conference on Learning Representations* https://openreview.net/forum?id=pAbm1qfheGk (2021).

89. Frederick, K. K., Marlow, M. S., Valentine, K. G. & Wand, A. J. Conformational entropy in molecular recognition by proteins. *Nature* **448**, 325–329, https://doi.org/10.1038/nature05959 (2007).

90. Murphy, K. P. *Machine learning: a probabilistic perspective* (MIT press, 2012).

91. Breiman, L. Random forests. *Machine learning* **45**, 5–32, https://doi.org/10.1023/A:1010933404324 (2001).

92. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *Journal of chemical information and modeling* **50**, 742–754, https://doi.org/10.1021/ci100050t (2010).

93. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, **70**, 1263–1272, https://proceedings.mlr.press/v70/gilmer17a.html (PMLR, 2017)

94. Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A. & Müller, K.-R. SchNet–A deep learning architecture for molecules and materials. *The Journal of Chemical Physics* **148**, 241722, https://doi.org/10.1063/1.5019779 (2018).

95. Schütt, K. *et al.* SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in neural information processing systems*, 991–1001, https://proceedings.neurips.cc/paper/2017/file/303ed4c69846ab36c2904d3ba8573050-Paper.pdf (2017).

96. Axelrod, S. & Gomez-Bombarelli, R. Conformer models and training datasets. *Harvard Dataverse* https://doi.org/10.7910/DVN/N4VLQL (2021).

97. Axelrod, S. & Gomez-Bombarelli, R. GEOM. *Harvard Dataverse* https://doi.org/10.7910/DVN/JNGTDF (2021).

98. Rappé, A. K., Casewit, C. J., Colwell, K., Goddard, W. A. III & Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American chemical society* **114**, 10024–10035, https://doi.org/10.1021/ja00051a040 (1992).

99. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *Journal of computational chemistry* **25**, 1157–1174, https://doi.org/10.1002/jcc.20035 (2004).

100. Stewart, J. J. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. *Journal of molecular modeling* **19**, 1–32, https://doi.org/10.1007/s00894-012-1667-x (2013).

101. Wenlock, M. & Tomkinson, N. Experimental *in vitro* DMPK and physicochemical data on a set of publicly disclosed compounds. https://doi.org/10.6019/CHEMBL3301361.

102. Martins, I. F., Teixeira, A. L., Pinheiro, L. & Falcao, A. O. A Bayesian approach to in silico blood-brain barrier penetration modeling. *Journal of chemical information and modeling* **52**, 1686–1697, https://doi.org/10.1021/ci300124c (2012).

103. Tox21 challenge. http://tripod.nih.gov/tox21/challenge/. Accessed 2017-09-27.

104. Richard, A. M. *et al.* ToxCast chemical landscape: paving the road to 21st century toxicology. *Chemical research in toxicology* **29**, 1225–1251, https://doi.org/10.1021/acs.chemrestox.6b00135 (2016).

105. Kuhn, M., Letunic, I., Jensen, L. J. & Bork, P. The SIDER database of drugs and side effects. *Nucleic acids research* **44**, D1075–D1079, https://doi.org/10.1093/nar/gkv1075 (2016).

106. Novick, P. A., Ortiz, O. F., Poelman, J., Abdulhay, A. Y. & Pande, V. S. SWEETLEAD: an in silico database of approved drugs, regulated chemicals, and herbal isolates for computer-aided drug discovery. *PloS one* **8**, e79568, https://doi.org/10.1371/journal.pone.0079568 (2013).

107. Aggregate Analysis of ClincalTrials.gov (AACT) Database. https://aact.ctti-clinicaltrials.org/. Accessed 2017-09-27.

Acknowledgements

Author contributions

R.G.-B. conceived the project and S.A. performed the calculations. Both authors wrote and revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to R.G.-B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.