SCIENTIFIC DATA

OPEN

DATA DESCRIPTOR

QM7-X, a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules

Johannes Hoja [1,2], Leonardo Medrano Sandonas[1], Brian G. Ernst[3], Alvaro Vazquez-Mayagoitia[4], Robert A. DiStasio Jr. [3✉] & Alexandre Tkatchenko[1✉]

We introduce QM7-X, a comprehensive dataset of 42 physicochemical properties for $\approx$4.2 million equilibrium and non-equilibrium structures of small organic molecules with up to seven non-hydrogen (C, N, O, S, Cl) atoms. To span this fundamentally important region of chemical compound space (CCS), QM7-X includes an exhaustive sampling of (meta-)stable equilibrium structures—comprised of constitutional/structural isomers and stereoisomers, e.g., enantiomers and diastereomers (including *cis-/trans-* and conformational isomers)—as well as 100 non-equilibrium structural variations thereof to reach a total of $\approx$4.2 million molecular structures. Computed at the tightly converged quantum-mechanical PBE0+MBD level of theory, QM7-X contains global (molecular) and local (atom-in-a-molecule) properties ranging from ground state quantities (such as atomization energies and dipole moments) to response quantities (such as polarizability tensors and dispersion coefficients). By providing a systematic, extensive, and tightly-converged dataset of quantum-mechanically computed physicochemical properties, we expect that QM7-X will play a critical role in the development of next-generation machine-learning based models for exploring greater swaths of CCS and performing in silico design of molecules with targeted properties.

Background & Summary

A crucial aspect of drug discovery[1] and molecular materials design[2] is an extensive exploration and understanding of chemical compound space (CCS)—the extremely high-dimensional space containing all feasible molecular compositions and conformations. Recently, the combination of quantum mechanical (QM) calculations with machine learning (ML) has led to considerable insight into CCS[3–9]. However, progress along these lines can only happen with the availability of extensive and comprehensive QM-based datasets that adequately describe the complex structure–property relationships in molecules across CCS. In this regard, one challenge that is encountered during the generation of such datasets is the fact that their dimension scales exponentially with molecule size, thereby making it difficult to explore increasingly large swaths of CCS. The second challenge is the steep computational cost of tightly converged QM calculations, which are critical for obtaining an accurate and reliable description of the structure and physicochemical properties of each molecule.

To begin such an extensive exploration of CCS, the GDB datasets[1,10,11] have enumerated up to 166 B organic molecules containing up to 17 heavy (non-hydrogen) atoms. Encoded as canonical SMILES (simplified molecular-input line-entry system) strings, the GDB datasets only provide the molecular formula and chemical connectivity, and do not contain any structural or molecular property information. As such, QM calculations on small subsets of the GDB datasets have subsequently been used to generate meta-stable conformations for each molecular composition. This has led to seminal QM-based datasets like QM7[10,12,13] and QM9[11,14], which are comprised of a single meta-stable molecular structure per SMILES string with up to seven and nine heavy (non-hydrogen) atoms, respectively. The QM7 and QM9 datasets have been widely used for benchmarking ML

[1]Department of Physics and Materials Science, University of Luxembourg, L-1511, Luxembourg City, Luxembourg. [2]Institute of Chemistry, University of Graz, 8010, Graz, Austria. [3]Department of Chemistry and Chemical Biology, Cornell University, Ithaca, NY, 14853, USA. [4]Computational Science Division, Argonne National Laboratory, Lemont, IL, 60439, USA. ✉e-mail: distasio@cornell.edu; alexandre.tkatchenko@uni.lu
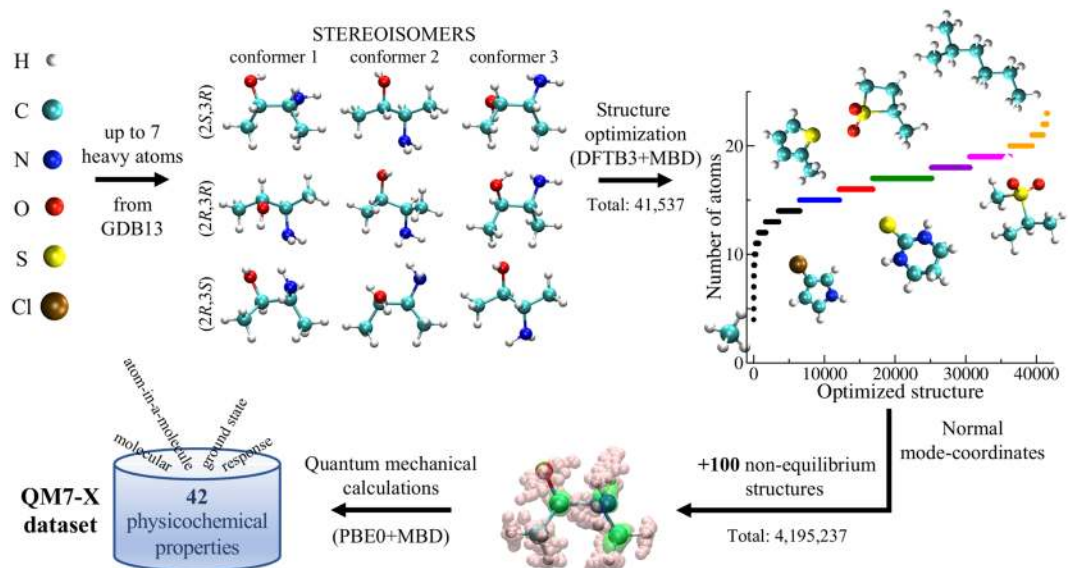
Fig. 1 Schematic representation of the QM7-X dataset. The "building blocks" for the QM7-X dataset are the set of ≈7 k molecular formulae which contain up to seven heavy (non-hydrogen) atoms from the GDB13 database. The first step in the construction of QM7-X is an extensive sampling of the structural/constitutional isomers and stereoisomers, *e.g.*, enantiomers and diastereomers (including *cis-/trans-* and conformational isomers) for each of the initial ≈7 k molecular formulae using the MMFF94 force field (schematically depicted above for some of the stereoisomers of 3-Amino-2-butanol). Each of these molecules was optimized at the DFTB3+MBD level of theory for a total of ≈40 k (meta-)stable equilibrium molecular structures. For each of these equilibrium structures, 100 non-equilibrium structures were generated by displacing the atoms according to linear combinations of the DFTB3+MBD normal modes for a total of ≈4.2 M molecular structures (with the range of atomic displacements (hydrogen/non-hydrogen atoms in pink/green) depicted above for a single 3-Amino-2-butanol conformer). We then performed QM calculations at the PBE0+MBD level to obtain 42 physicochemical properties for each of the ≈4.2 M molecular structures.

approaches and exploring molecular structure–property correlations[3]. In addition, molecular dynamics (MD) based datasets have become available for a few selected molecules and solids, and contain both equilibrium and non-equilibrium structures[15–21]; such datasets are becoming increasingly more useful for constructing advanced interatomic potentials[16,22,23].

A more substantial coverage of CCS for small organic molecules was provided by Smith *et al.*[24,25] with the generation of the ANI-1 dataset, which consists of more than 20 M equilibrium and non-equilibrium conformations of molecules containing up to eight heavy (C, N, O) atoms from GDB-11[26,27]. More recently, the ANI-1x dataset[28] was introduced, which also contains 20 physicochemical properties for about 5 M structures computed using the semi-empirical ωB97X[29] density functional. In addition, Smith *et al.*[28] also provided local CCSD(T) energies for a smaller subset of 0.5 M structures. To date, the ANI-1 datasets contain the largest available collection of GDB-based QM calculations of molecular structures and properties. However, four challenges still remain to enable a systematic exploration of CCS for small organic molecules: (i) providing a systematic sampling of CCS in terms of constitutional/structural isomers and stereoisomers (*e.g.*, enantiomers and diastereomers, including *cis-/trans-* and meta-stable conformational isomers), (ii) assessing the accuracy and reliability of QM structures and properties with respect to the employed density-functional approximation (especially for non-equilibrium conformations), (iii) offering a large set of local (atom-in-a-molecule) and global (molecular) physicochemical properties that would enable a comprehensive exploration of structure–property relationships throughout CCS, and (iv) providing accurate and reliable QM data that will enable the construction of models for describing covalent and non-covalent van der Waals (vdW) interactions.

In order to convincingly address these four challenges in this work, we present QM7-X, which aims to provide a systematic, extensive, and tightly converged dataset of QM-based physical and chemical properties for a fundamentally important region of CCS covering small organic molecules (see Fig. 1). To do so, we performed a systematic and exhaustive sampling of the (meta-)stable equilibrium structures of all molecules with up to seven heavy (C, N, O, S, Cl) atoms in the GDB13 database[10] using a density-functional tight binding (DFTB) approach; this includes constitutional/structural isomers and stereoisomers, *e.g.*, enantiomers and diastereomers (including *cis-/trans-* and conformational isomers). This was followed by the generation of 100 non-equilibrium structures (*via* DFTB normal-mode displacements of each equilibrium structure) for a total of ≈4.2 M molecular structures. For each of these equilibrium and non-equilibrium molecular structures, QM7-X also includes an extensive number of QM-obtained physicochemical properties, most of which were computed using non-empirical hybrid density-functional theory (DFT) with a many-body treatment of vdW dispersion interactions (*i.e.*, PBE0+MBD) in conjunction with tightly-converged numeric atom-centered orbitals[30]. In total, QM7-X contains 42 molecular (global) and atom-in-a-molecule (local) properties, which range from ground state quantities (such as total and

atomization energies, atomic forces, HOMO-LUMO gaps, dipole/quadrupole moments, Hirshfeld quantities, etc.) to response quantities (such as polarizability tensors and dispersion coefficients)—all of which could be utilized for the construction of next-generation intra- and inter-molecular force fields. As such, we expect that QM7-X will be useful for the development of accurate and reliable ML-based techniques that will provide new insight into the complex structure–property relationships in molecules, and ultimately allow for more extensive exploration of CCS and the rational design of molecules with tailored properties.

Methods

Generation of equilibrium structures. As a basis for the QM7-X dataset, we considered all molecules containing up to seven heavy (non-hydrogen) atoms in the GDB13 database[10], which provides an enumeration of the CCS spanned by small organic molecules comprised of H, C, N, O, S, and Cl atoms. These molecules range in size from $N = 4$–23 atoms (see Fig. 1). For each molecular formula, the GDB13 database only includes chemical connectivity information (*i.e.*, some structural/constitutional isomers) encoded as SMILES strings. To generate *all* of the corresponding structural/constitutional isomers and stereoisomers, *e.g.*, enantiomers and diastereomers (including *cis-/trans-* isomers) for each molecular formula, we created canonical SMILES strings for each possible molecular structure. Based on these SMILES strings, initial 3D structures were obtained using the MMFF94 force field[31–35] *via* the gen3d option in Open Babel[36]. For each of these structures, we also generated a set of sufficiently different (meta-)stable conformational isomers, *i.e.*, isomers which can be interconverted by rotations around single bonds. To do this, we performed a conformational isomer search with the Confab tool[37] in conjunction with the MMFF94 force field. At the MMFF94 level, we retained the set of conformers that were within 50 kcal/mol of the most stable structure and differed by a root-mean-square deviation (RMSD) of 0.5 Å.

All molecular structures were subsequently re-optimized with third-order self-consistent charge density functional tight binding (DFTB3)[38–40] supplemented with a treatment of many-body dispersion (MBD) interactions[41–44], using the 3ob parameters[45,46]. All DFTB3+MBD calculations were performed using in-house versions of the DFTB+ code[47] and the Atomic Simulation Environment (ASE)[48]. The lowest-energy structure obtained at the DFTB3+MBD level was considered the first conformer, and additional structures were accepted as separate conformers if the RMSD between the respective structure and the first conformer (as well as all subsequently accepted conformers) was larger than 1.0 Å. While most of these molecular structures correspond to local minima at the DFTB3+MBD level, we note in passing that some correspond to saddle points on the DFTB3+MBD potential energy surface (PES).

Generation of non-equilibrium structures. In order to sample some of the non-equilibrium sectors of CCS, we generated 100 non-equilibrium structures for each of the (meta-)stable equilibrium structures described above. This was achieved by displacing each molecular structure along a linear combination of normal mode coordinates computed at the DFTB3+MBD level within the harmonic approximation (for consistency with the level of theory used to optimize the corresponding molecular structures). In order to generate comparable displacements for all equilibrium molecular structures (despite differing molecular sizes and relative stabilities), we generated a set of displaced structures which has an *average* energy difference (with respect to the corresponding (meta-)stable equilibrium structure) of

$$\langle \Delta E \rangle = \frac{3}{2} N k_B T, \tag{1}$$

in analogy to the equipartition theorem from classical statistical mechanics, in which each of the $3N$ degrees of freedom in a monatomic ideal gas contributes $\frac{1}{2} k_B T$ to the internal energy (with k_B being the Boltzmann constant and T the temperature). To sufficiently sample the PES for each (meta-)stable equilibrium structure, we set $T = 1500$ K. In addition to the $<\Delta E>$ convention defined in Eq. (1), the 100 non-equilibrium molecular structures were also required to follow the corresponding Boltzmann distribution:

$$P(\Delta E) = exp\left[-\frac{\Delta E}{k_B T}\right]. \tag{2}$$

To generate a set of *candidate* non-equilibrium molecular structures, we randomly chose an energy difference ($\Delta \varepsilon$) for each of the $3N$–6 (or $3N$–5 in the case of a diatomic and/or linear polyatomic molecule) vibrational/normal modes in a given molecule. For the i-th vibrational/normal mode (with frequency ν_i), the corresponding displacement amplitude was computed *via* $A_i = \frac{\sqrt{2\,\Delta\varepsilon_i}}{\nu_i}$, and the corresponding displacement vector (along this mode) was obtained by multiplying the eigenvectors with A_i and a randomized sign. A candidate non-equilibrium molecular structure was then obtained by simultaneously perturbing the equilibrium molecular structure along every displacement vector. Since the generation of candidate non-equilibrium molecular structures *via* a linear combination of normal-mode displacements assumes complete independence of the normal modes as well as the validity of the harmonic approximation, the actual ΔE corresponding to each candidate was explicitly calculated with DFTB3+MBD and the required Boltzmann distribution in Eq. (2) was strictly enforced. To do so, we pre-defined a histogram that accounts for 100 structures and covers a range of ΔE values from 0.0 to $4.5 \times \frac{3}{2} N k_B T$, as shown in Fig. S1 of the Supplementary Information (SI). To generate a final set of 100 non-equilibrium molecular structures which meets the aforementioned criteria, the corresponding distribution of ΔE values had to fit this histogram. More specifically, we accepted candidate structures (with an RMSD $>0.0075N$ Å to all previously accepted structures) until all of the histogram bins were filled.

Heavy Atoms	Molecules from GDB13	Stereoisomers	Equilibrium Structures	Total Structures (Equilibrium + Non-Equilibrium)
1	1	1	1	101
2	2	2	2	202
3	10	10	10	1,010
4	42	45	58	5,858
5	149	193	351	35,451
6	901	1,400	3,677	371,377
7	5,845	11,627	37,438	3,781,238
Total	6,950	13,278	41,537	4,195,237

Table 1. Content and size of the QM7-X dataset. The number of molecules from GDB13, stereoisomers (*e.g.*, enantiomers and *cis-/trans-* diastereomers), (meta-)stable equilibrium structures (including conformational isomers), and the total number of equilibrium and non-equilibrium structures are listed for different numbers of heavy atoms and the entire QM7-X dataset.

The approximately 4.2 M equilibrium and non-equilibrium structures generated using these procedures form the QM7-X dataset (see Table 1). Since we were unable to obtain a complete energy distribution (as described above) for 261 of the 7,211 molecular formulae taken from the GDB13 dataset, molecules with these molecular formulae were not included in QM7-X. As such, QM7-X covers 96.4% of all molecules containing up to seven heavy atoms from GDB13 (and 98.8% of the corresponding (meta-)stable stereoisomers), which should provide a sufficiently accurate and representative sample of the small organic molecules contained in CCS. We note in passing that some of the generated equilibrium structures are not entirely unique, *e.g.*, some of the stereoisomers of molecules with multiple chiral centers were identical as well as some of the (rotational) conformational isomers; permutations of identical atoms were not initially considered when computing RMSD values utilizing the RMSD minimization tool available in ASE[49]. Therefore, approximately 4.6% of the equilibrium structures constitute duplicates (see file `DupMols.dat` in the ZENODO repository[50]), which were identified by an additional *a posteriori* approach that considered similarities between potential duplicate structures in the following six physicochemical properties: PBE0 energy, MBD energy, HOMO-LUMO gap, HOMO energy, molecular polarizability, and total dipole moment. The threshold used to check for similarity was set to $10^{-3} \times$ the corresponding unit for each property. While the presence of some duplicate equilibrium structures in the QM7-X dataset might influence the performance of ML models developed using this data, the non-equilibrium structures associated with these duplicate equilibrium structures are **not identical** and probe different regions of the molecular PES. As such, their inclusion in QM7-X covers a larger swath of molecular property space, and can contribute to the development of more robust ML models for predicting the physicochemical properties in the QM7-X dataset. Thus, we offer users the option to keep or exclude these duplicate equilibrium structures (as well as their corresponding non-equilibrium structures) in the QM7-X dataset *via* the `createDB.py` script uploaded to the ZENODO repository[50].

Calculation of physicochemical properties. These ≈4.2 M DFTB3+MBD structures were now utilized for more accurate QM single-point calculations using dispersion-inclusive hybrid DFT. Energies, forces, and several other physicochemical properties (see Table 2) were calculated at the PBE0+MBD[41,51,52] level using the `FHI-aims` code[53,54] (version 180218). For all calculations, "tight" settings were applied for basis functions and integration grids. Energies were converged to 10^{-6} eV and the accuracy of the forces was set to 10^{-4} eV/Å. The convergence criteria used during self-consistent field (SCF) optimizations were 10^{-3} eV for the sum of eigenvalues and 10^{-6} electrons/Å^3 for the charge density.

Here, the MBD energies, MBD atomic forces, atomic C_6 coefficients, and isotropic atomic polarizabilities were computed using the range-separated self-consistent screening (rsSCS) approach[42], while the molecular C_6 coefficients and polarizabilities (both isotropic and tensor) were obtained *via* the SCS approach[41]. Hirshfeld ratios correspond to the Hirshfeld volumes divided by the free atom volumes. The TS dispersion energy refers to the pairwise Tkatchenko-Scheffler (TS) dispersion energy in conjunction with the PBE0 functional[55]. The vdW radii were also obtained using the SCS approach *via* $R_{vdW} = (\alpha^{SCS}/\alpha^{TS})^{1/3} R_{vdW}^{TS}$, where α^{TS} and R_{vdW}^{TS} are the atomic polarizability and vdW radius computed according to the TS scheme, respectively. Atomization energies were obtained by subtracting the atomic PBE0 energies from the PBE0 energy of each molecular conformation (see Table S2). The exact exchange energy is the amount of exact (or Hartree-Fock) exchange that has been admixed into the exchange-correlation energy.

Data Records

The QM7-X dataset is provided in eight HDF5 based files in a ZENODO.ORG data repository[50]. One can also find there a README file with technical usage details and examples of how to access the information stored in QM7-X (with and without considering duplicates, see `createDB.py` file).

HDF5 file format. The information for each molecular structure is stored in a Python dictionary (`dict`) type containing all relevant properties and recorded in *groups* in HDF5 file format. HDF5 keys to access the atomic numbers, atomic positions (coordinates), and physicochemical properties in each dictionary are provided in Table 2. The dimension of each array depends on the number of atoms and the required property, *e.g.*, for a

#	Symbol	Property	Unit	Dimension	Type	Level	HDF5 keys
1	Z	Atomic numbers	—	N	S	—	'atNUM'
2	R	Atomic positions (coordinates)	Å	$3N$	S	TB	'atXYZ'
3	ΔR	RMSD to optimized structure	Å	1	S	—	'sRMSD'
4	I	Moment of inertia tensor	amu·Å^2	9	S	—	'sMIT'
5	E_{tot}	Total PBE0+MBD energy	eV	1	M,G	P0M	'ePBE0+MBD'
6	E_{TB}	Total DFTB3+MBD energy	eV	1	M,G	TB	'eDFTB+MBD'
7	E_{at}	Atomization energy	eV	1	M,G	P0	'eAT'
8	E_{PBE0}	PBE0 energy	eV	1	M,G	P0	'ePBE0'
9	E_{MBD}	MBD energy	eV	1	M,G	P0M	'eMBD'
10	E_{TS}	TS dispersion energy	eV	1	M,G	P0	'eTS'
11	E_{nn}	Nuclear-nuclear repulsion energy	eV	1	M,G	—	'eNN'
12	E_{kin}	Kinetic energy	eV	1	M,G	P0	'eKIN'
13	E_{ne}	Nuclear-electron attraction	eV	1	M,G	P0	'eNE'
14	E_{coul}	Classical coulomb energy (el-el)	eV	1	M,G	P0	'eEE'
15	E_{xc}	Exchange-correlation energy	eV	1	M,G	P0	'eXC'
16	E_x	Exchange energy	eV	1	M,G	P0	'eX'
17	E_c	Correlation energy	eV	1	M,G	P0	'eC'
18	E_{xx}	Exact exchange energy	eV	1	M,G	P0	'eXX'
19	E_{KS}	Sum of Kohn-Sham eigenvalues	eV	1	M,G	P0	'eKSE'
20	ε	Kohn-Sham eigenvalues	eV	*	M,G	P0	'KSE'
21	E_{HOMO}	HOMO energy	eV	1	M,G	P0	'eH'
22	E_{LUMO}	LUMO energy	eV	1	M,G	P0	'eL'
23	E_{gap}	HOMO-LUMO gap	eV	1	M,G	P0	'HLgap'
24	D_s	Scalar dipole moment	e·Å	1	M,G	P0	'DIP'
25	D	Dipole moment	e·Å	3	M,G	P0	'vDIP'
26	Q_{tot}	Total quadrupole moment	e·Å^2	3	M,G	P0	'vTQ'
27	Q_{ion}	Ionic quadrupole moment	e·Å^2	3	M,G	P0	'vIQ'
28	Q_{elec}	Electronic quadrupole moment	e·Å^2	3	M,G	P0	'vEQ'
29	C_6	Molecular C_6 coefficient	$E_h \cdot a_0^3$	1	M,R	P0M	'mC6'
30	α_s	Molecular polarizability (isotropic)	a_0^3	1	M,R	P0M	'mPOL'
31	α	Molecular polarizability tensor	a_0^3	9	M,R	P0M	'mTPOL'
32	F_{tot}	Total PBE0+MBD atomic forces	eV/Å	$3N$	A,G	P0M	'totFOR'
33	F_{PBE0}	PBE0 atomic forces	eV/Å	$3N$	A,G	P0	'pbe0FOR'
34	F_{MBD}	MBD atomic forces	eV/Å	$3N$	A,G	P0M	'vdwFOR'
35	V_H	Hirshfeld volumes	a_0^3	N	A,G	P0	'hVOL'
36	V_{ratio}	Hirshfeld ratios	—	N	A,G	P0	'hRAT'
37	q_H	Hirshfeld charges	e	N	A,G	P0	'hCHG'
38	$D_{H,s}$	Scalar Hirshfeld dipole moments	e·a_0	N	A,G	P0	'hDIP'
39	D_H	Hirshfeld dipole moments	e·a_0	$3N$	A,G	P0	'hVDIP'
40	$\widetilde{C}_6$	Atomic C_6 coefficients	$E_h \cdot a_0^6$	N	A,R	P0M	'atC6'
41	$\widetilde{\alpha}_s$	Atomic polarizabilities (isotropic)	a_0^3	N	A,R	P0M	'atPOL'
42	R_{vdW}	vdW radii	a_0	N	A,R	P0M	'vdwR'

Table 2. List of physicochemical properties in the QM7-X dataset. Each property is represented by a symbol (with units and dimension) and can be found in the HDF5 files using the corresponding HDF5 keys. Different property types are distinguished as follows: structural (S), molecular (M), atom-in-a-molecule (A), ground-state (G), and response (R). Different levels of theory are indicated as follows: DFTB3+MBD (TB), PBE0 (P0), and PBE0+MBD (P0M). The P0M label indicates which properties explicitly include dispersion interactions. E_h and a_0 refer to the atomic units of energy (Hartree) and length (Bohr radius), respectively. *The number of Kohn-Sham eigenvalues varies for each molecule.

methane (CH_4) molecule 'atNUM' is a 1D array of $N = 5$ elements ([6, 1, 1, 1, 1]) while 'atXYZ' is a 2D array comprised of $N = 5$ rows and three columns (x, y, z coordinates). All structures are labeled as Geom-mr-is-ct-u, where r enumerates the SMILES strings, s the stereoisomers (excluding conformational isomers), t the considered (meta-)stable conformational isomers, and u the optimized/displaced structures (u = opt indicates the DFTB3+MBD optimized structures and u = 1, …, 100 indicates the displaced non-equilibrium structures). We note in passing that the indices used in the QM7-X dataset reflect the order in which a given structure was generated and do not correspond to sorted DFTB3+MBD (or PBE0+MBD) energies.
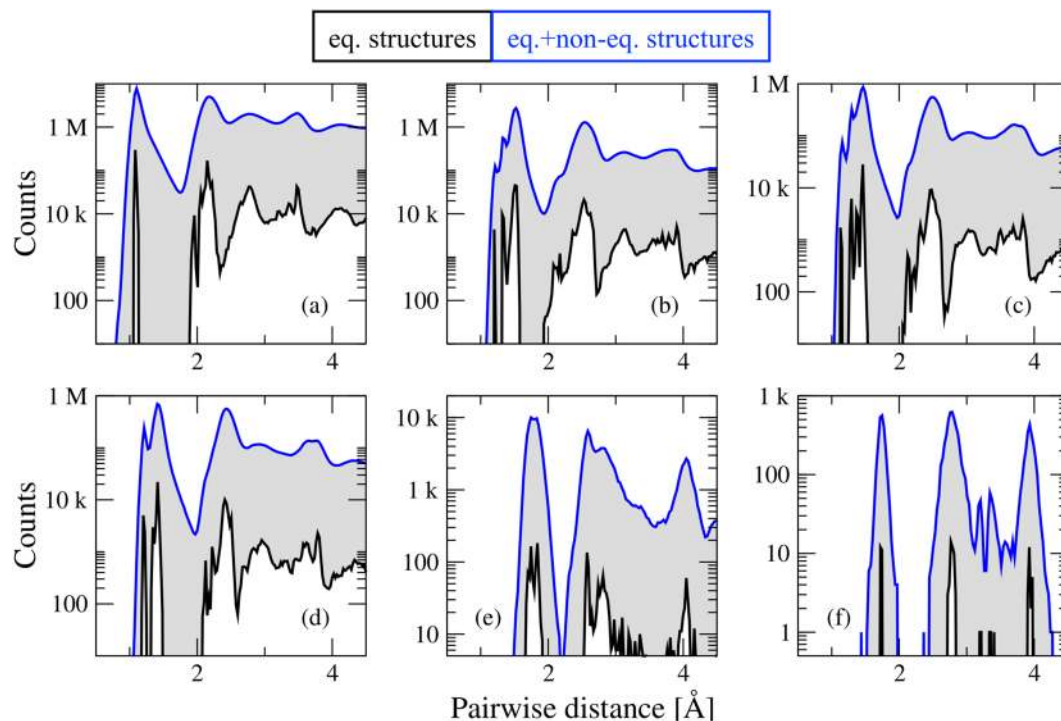
Fig. 2 Distribution of atom-pairwise distances in the QM7-X dataset. Corresponding atom-pair distance distributions for: (**a**) C-H, (**b**) C-C, (**c**) C-N, (**d**) C-O, (**e**) C-S, and (**f**) C-Cl in the QM7-X dataset. For each distribution, we show the results corresponding to the equilibrium structures only (black lines) as well as the entire QM7-X dataset (blue lines). Contributions from non-equilibrium structures are highlighted in grey.

Technical Validation

In contrast to many other QM-based datasets like QM7 and QM9, which contain a single structure per molecule, the QM7-X dataset not only includes constitutional/structural isomers, but also stereoisomers (*e.g.*, enantiomers and diastereomers, including *cis-/trans-* and conformational isomers) as well as non-equilibrium structural variations thereof. While many of the considered molecules do overlap with those in the ANI-1 dataset[24], QM7-X also contains a more extensive sampling of equilibrium stereoisomers and a significantly different sampling of non-equilibrium structures (*vide infra*). Moreover, the number of reported physicochemical properties and the employed level of electronic structure theory in QM7-X is substantially more advanced compared to prior work. Therefore, the QM7-X dataset contains significantly more data for flexible molecules with stereocenters than for rigid molecules without stereocenters. Since stereochemistry can play an important role in drug design, we consider that the data provided in this work will enable ML models to capture the subtle physicochemical differences existing between stereoisomers.

Another important target of the QM7-X dataset is efficient PES sampling for a large number of small organic molecules. Since the most crucial regions of the PES are reasonably close to the relevant (meta-)stable equilibrium structures, QM7-X contains 100 distorted/non-equilibrium structures for every such conformer. While most of these structures are in the vicinity of the optimized/equilibrium structures, some correspond to more highly distorted structures on the PES. Since it was only computationally feasible to consider 100 non-equilibrium structures at the PBE0+MBD level, these structures were chosen to cover as much of the PES surrounding each (meta-)stable equilibrium structure as possible. As such, we created these non-equilibrium structures by displacing the atoms in a given molecule along linear combinations of normal modes. Although a similar approach was utilized during the construction of the ANI-1 dataset[24], we have eliminated all issues related to the inaccuracy of the harmonic approximation by recomputing the energy of every candidate structure and only accepting those structures with a Boltzmann energy distribution (see Methods). In doing so, we ensure that the relevant regions of each PES are well-sampled and only a small (and pre-determined) number of high-energy non-equilibrium structures are included in our dataset. The inclusion of non-equilibrium structures for each (meta-)stable equilibrium structure also provides better coverage of the conformational space than initially provided by the equilibrium structures, as seen in the pairwise distance distribution plots shown in Fig. 2.

In this work, the structure generation process was performed using DFTB3+MBD, while the subsequent energy, force, and property calculations were performed at the more robust PBE0+MBD level of theory. Since the focus of this work is the physicochemical properties of non-equilibrium structures surrounding (meta-)stable equilibrium points on each PES, the use of these two methods does not introduce any significant complications and/or errors. Although the PBE0+MBD (relative) energies will not strictly follow the same Boltzmann distribution as that used to generate the molecular structures at the DFTB3+MBD level, these two distributions are often very similar and lead to virtually the same $<\Delta E>$. An average over all histograms is depicted in Fig. S1, where
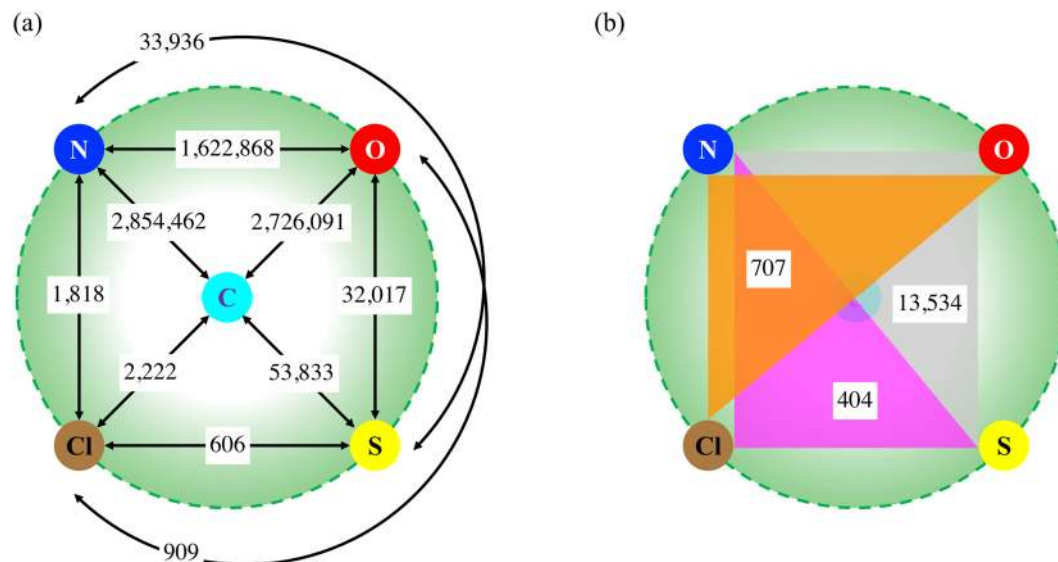
Fig. 3 Statistical analysis of the molecular composition of the QM7-X dataset. Schematic illustrations representing the number of structures containing at least (**a**) two or (**b**) three different heavy (non-hydrogen) elements. Note that all molecules in QM7-X contain C and H atoms. Triple combinations including C atoms are not shown in (**b**) since that information can be extracted from (**a**).

one can see some minor variations in the heights of the first few bins and only an insignificant amount of structures located slightly outside of the initially envisioned window of ΔE values. In addition, we have also quantitatively analyzed the structural agreement between conformers optimized at the DFTB3+MBD and PBE0+MBD levels. To do this, the DFTB3+MBD-optimized structures of 10 flexible molecules with at least five (meta-)stable conformers (see SI for more details) were randomly selected and re-optimized with PBE0+MBD, followed by a (harmonic) vibrational frequency analysis at the same level of theory. In doing so, it was found that all 63 so obtained PBE0+MBD geometries constitute local or global minima (*i.e.*, with no imaginary frequencies), and the average RMSD between the DFTB3+MBD and PBE0+MBD structures amounts to only 0.1 Å. These structural differences are visualized for several conformers in Figs. S3–S5 in the SI. Even for the conformer with the largest observed RMSD of 0.66 Å, we note that the PBE0+MBD structure did not move to any of the other considered conformers, and resulted from modifications to the backbone structure. From this analysis, we would argue that the DFTB3+MBD-optimized structures are in excellent agreement with those optimized at the PBE0+MBD level, and hence provide a high-fidelity representation of the critical points on the PBE0+MBD PES.

In QM7-X, molecules made of seven heavy atoms are the most abundant with 37,438 equilibrium structures (including conformational isomers) and 3,743,800 non-equilibrium structures, followed by molecules containing six heavy atoms (see Table 1). $C_5H_{11}NO$ turned out to be the most plentiful molecular formula with 3,200 (meta-) stable equilibrium structures (see Table S1). To examine the molecular composition of the QM7-X dataset, we have performed a statistical analysis by counting the number of structures containing at least two or three different heavy (non-hydrogen) elements (see Fig. 3).

Here, it was found that 68.0% and 65.0% of the QM7-X dataset is comprised of molecular structures which contain only [H,C,N] and [H,C,O] atoms, respectively, while structures containing [H,N,O] atoms are found in 38.7% of the dataset. As such, we consider that QM7-X provides a representative sample of CCS formed by small [C,H,O,N]-based molecules. Regarding S-containing molecules, we still have a considerable number of structures with [H,C,S] (53,833), [H,N,S] (33,936), and [H,O,S] (32,017) combinations (even considering the 13,534 [H,N,O,S]-based molecules in Fig. 3(b)) which might help to describe a wider swath of CCS. On the other hand, only 0.06% of QM7-X includes Cl-containing molecules.

Over 40 different physicochemical properties were computed to describe the structure–property and property–property relationships in QM7-X. In Table 2, we showcase a number of properties obtained from the output of QM calculations performed at the PBE0+MBD level. This QM level of theory has proven to be accurate and reliable for describing intramolecular degrees of freedom in addition to intermolecular interactions in organic molecular dimers, supramolecular complexes, and molecular crystals[41,51,52,56–59]. We therefore consider that these calculations are suitable to validate the quality of future work using this dataset.

To provide an example of the significant information that can be obtained from QM7-X, we have plotted the distribution of several physicochemical properties in Fig. 4 according to the classification scheme defined in Table 2, *i.e.*, division into global (molecular) and local (atom-in-a-molecule) properties, as well as ground-state and response properties. For a cursory look at how some of these properties vary with molecular size, see Fig. S2. Here, the influence of including non-equilibrium structures on a given property is highlighted by comparing the property distributions corresponding to the equilibrium structures only (black lines) with that of the entire QM7-X dataset (blue lines). Overall, one can see many interesting trends in this data. Generally speaking, structural distortions produce significant fluctuations around the values of each property for the equilibrium
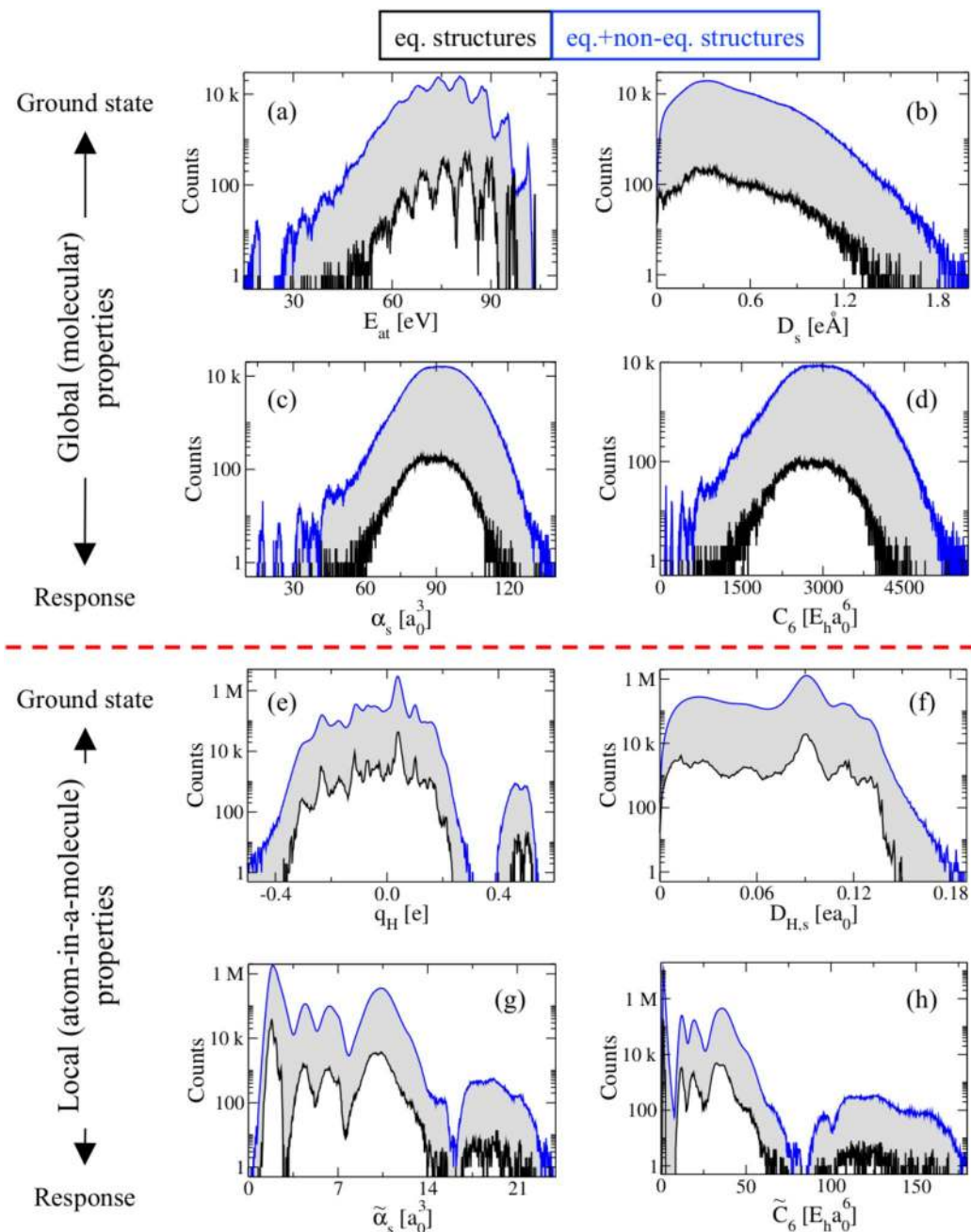
Fig. 4 Distribution of physicochemical properties considered in the QM7-X dataset. Global (molecular) properties are considered in the top panel, while local (atom-in-a-molecule) properties are considered in the bottom panel. For each case, we have selected two different examples ground-state and response properties: (**a**) atomization energy, (**b**) scalar dipole moment, (**c**) molecular polarizability (isotropic), (**d**) molecular C_6 coefficient, (**e**) Hirshfeld charges, (**f**) scalar Hirshfeld dipole moment, (**g**) atomic polarizability (isotropic), and (**h**) atomic C_6 coefficient. For each property, we show the results corresponding to the equilibrium structures only (black lines) as well as the entire QM7-X dataset (blue lines). Contributions from non-equilibrium structures are highlighted in grey.

structures, and therefore improve the exploration and description of molecular property space. In the examples provided here, we find that global properties such as molecular polarizabilities (α_s) and dispersion coefficients (C_6) show similar distributions due to the strong correlation existing between them *via* the Casimir-Polder integral[60] (see Fig. 4(c,d)). Whereas, their local analogs, *i.e.*, the atomic polarizabilities ($\widetilde{\alpha}_s$) and dispersion coefficients ($\widetilde{C}_6$), display characteristic peaks corresponding to the specific local atomic environments found in the equilibrium and non-equilibrium molecular structures (see Fig. 4(g,h)). We also find that intensive properties (*e.g.*, HOMO-LUMO gaps and dipole moments) are more sensitive to structural distortions as compared to extensive properties (*e.g.*, atomization energies), see Fig. S2. Accordingly, QM7-X offers us the possibility to explore a great

diversity of physicochemical properties and to search for unknown correlations among components of the CCS for small molecules. It also opens up a new route for rational design and precise control over the physicochemical properties of small drug-like organic molecules.

Regarding ML applications, we are confident that the data provided by QM7-X will enable continued development of ML approaches, identification of novel QM-based molecular descriptors, as well as determination of robust training sample selection methods that will improve the accuracy and reliability of molecular property predictions. For instance, Stöhr et al.[61] have recently employed different methods for selecting the molecular structures from QM7-X to define training sets for the prediction of DFTB repulsive potentials. In this work, it was found that randomly selected training samples have a better performance compared to more refined training sets (e.g., consisting of equilibrium structures together with a given number (X) of non-equilibrium structures) when developing accurate many-body repulsive potentials for the DFTB method via deep tensor neural networks. Hence, QM7-X can be used to improve the accuracy of electronic structure methods using ML approaches such as Δ-learning. Further improvements of ML models may also be accomplished by considering physics-/chemistry-based approaches for selecting training samples (e.g., based on the local chemical environment surrounding each atom).

Code availability

The initial structure generation was carried out by using `Open Babel`[36]. Further structure optimization and the creation of non-equilibrium structures was performed by utilizing an in-house version of `DFTB+`[47] together with `ASE`[48]. Note that all necessary features regarding the utilized DFTB3+MBD approach are available in the current `DFTB+` version[62]. All DFT calculations were carried out using `FHI-aims`[53] (version 180218).

References

1. Reymond, J.-L. & Awale, M. Exploring chemical space for drug discovery using the chemical universe database. *ACS Chem. Neurosci.* **3**, 649–657, https://doi.org/10.1021/cn3000422 (2012).
2. Gómez-Bombarelli, R. *et al.* Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **15**, 1120–1127, https://doi.org/10.1038/nmat4717 (2016).
3. von Lilienfeld, O. A., Müller, K.-R. & Tkatchenko, A. Exploring chemical compound space with quantum-based machine learning. *Nat. Rev. Chem., in press*, https://arxiv.org/abs/1911.10084 (2020).
4. von Lilienfeld, O. A. Quantum machine learning in chemical compound space. *Angew. Chem. Int. Ed.* **57**, 4164–4169, https://doi.org/10.1002/anie.201709686 (2018).
5. Hansen, K. *et al.* Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *J. Phys. Chem. Lett.* **6**, 2326–2331, https://doi.org/10.1021/acs.jpclett.5b00831 (2015).
6. Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R. & Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **8**, 13890, https://doi.org/10.1038/ncomms13890 (2017).
7. Christensen, A. S., Faber, F. A. & von Lilienfeld, O. A. Operators in quantum machine learning: Response properties in chemical space. *J. Chem. Phys.* **150**, 064105, https://doi.org/10.1063/1.5053562 (2019).
8. De, S., Bartók, A. P., Csányi, G. & Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **18**, 13754–13769, https://doi.org/10.1039/c6cp00415f (2016).
9. Bartók, A. P. *et al.* Machine learning unifies the modeling of materials and molecules. *Sci. Adv.* **3**, e1701816, https://doi.org/10.1126/sciadv.1701816 (2017).
10. Blum, L. C. & Reymond, J.-L. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.* **131**, 8732–8733, https://doi.org/10.1021/ja902302h (2009).
11. Ruddigkeit, L., van Deursen, R., Blum, L. C. & Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* **52**, 2864–2875, https://doi.org/10.1021/ci300415d (2012).
12. Montavon, G. *et al.* Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* **15**, 095003, https://doi.org/10.1088/1367-2630/15/9/095003 (2013).
13. Yang, Y. *et al.* Quantum mechanical static dipole polarizabilities in the QM7b and AlphaML showcase databases. *Sci. Data* **6**, 1–10, https://doi.org/10.1038/s41597-019-0157-8 (2019).
14. Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **1**, 140022, https://doi.org/10.1038/sdata.2014.22 (2014).
15. Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A. & Müller, K.-R. SchNet – a deep learning architecture for molecules and materials. *J. Chem. Phys.* **148**, 241722, https://doi.org/10.1063/1.5019779 (2018).
16. Chmiela, S., Sauceda, H. E., Müller, K.-R. & Tkatchenko, A. Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. Commun.* **9**, 3887, https://doi.org/10.1038/s41467-018-06169-2 (2018).
17. Behler, J. Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations. *Phys. Chem. Chem. Phys.* **13**, 17930, https://doi.org/10.1039/c1cp21668f (2011).
18. Behler, J. Perspective: Machine learning potentials for atomistic simulations. *J. Chem. Phys.* **145**, 170901, https://doi.org/10.1063/1.4966192 (2016).
19. Dral, P. O., Owens, A., Yurchenko, S. N. & Thiel, W. Structure-based sampling and self-correcting machine learning for accurate calculations of potential energy surfaces and vibrational levels. *J. Chem. Phys.* **146**, 244108, https://doi.org/10.1063/1.4989536 (2017).
20. Gastegger, M., Behler, J. & Marquetand, P. Machine learning molecular dynamics for the simulation of infrared spectra. *Chem. Sci.* **8**, 6924–6935, https://doi.org/10.1039/c7sc02267k (2017).
21. Glielmo, A., Zeni, C. & Vita, A. D. Efficient nonparametric n-body force fields from machine learning. *Phy. Rev. B* **97**, 184307, https://doi.org/10.1103/physrevb.97.184307 (2018).
22. Bereau, T., DiStasio, R. A. Jr., Tkatchenko, A. & von Lilienfeld, O. A. Non-covalent interactions across organic and biological subsets of chemical space: Physics-based potentials parametrized from machine learning. *J. Chem. Phys.* **148**, 241706, https://doi.org/10.1063/1.5009502 (2018).
23. Metcalf, D. P. *et al.* Approaches for machine learning intermolecular interaction energies and application to energy components from symmetry adapted perturbation theory. *J. Chem. Phys.* **152**, 074103, https://doi.org/10.1063/1.5142636 (2020).
24. Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1, A data set of 20 million calculated off-equilibrium conformations for organic molecules. *Sci. Data* **4**, 170193, https://doi.org/10.1038/sdata.2017.193 (2017).

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

26. Fink, T., Bruggesser, H. & Reymond, J.-L. Virtual exploration of the small-molecule chemical universe below 160 Daltons. *Angew. Chem. Int. Ed.* **44**, 1504–1508, https://doi.org/10.1002/anie.200462457 (2005).

27. Fink, T. & Reymond, J.-L. Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *J. Chem. Inf. Model.* **47**, 342–353, https://doi.org/10.1021/ci600423u (2007).

28. Smith, J. S. *et al.* The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules. *Sci. Data* **7**, https://doi.org/10.1038/s41597-020-0473-z (2020).

29. Chai, J.-D. & Head-Gordon, M. Systematic optimization of long-range corrected hybrid density functionals. *J. Chem. Phys.* **128**, 084106, https://doi.org/10.1063/1.2834918 (2008).

30. Havu, V., Blum, V., Havu, P. & Scheffler, M. Efficient O(N) integration for all-electron electronic structure calculation using numeric basis functions. *J. Comput. Phys.* **228**, 8367–8379, https://doi.org/10.1016/j.jcp.2009.08.008 (2009).

31. Halgren, T. A. Merck molecular force field. i. basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **17**, 490–519, https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6<490::AID-JCC1>3.0.CO;2-P (1996).

32. Halgren, T. A. Merck molecular force field. ii. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. *J. Comput. Chem.* **17**, 520–552, https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6<520::AID-JCC2>3.0.CO;2-W (1996).

33. Halgren, T. A. Merck molecular force field. iii. molecular geometries and vibrational frequencies for MMFF94. *J. Comput. Chem.* **17**, 553–586, https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6<553::AID-JCC3>3.0.CO;2-T (1996).

34. Halgren, T. A. & Nachbar, R. B. Merck molecular force field. iv. conformational energies and geometries for MMFF94. *J. Comput. Chem.* **17**, 587–615, https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6<587::AID-JCC4>3.0.CO;2-Q (1996).

35. Halgren, T. A. Merck molecular force field. v. extension of MMFF94 using experimental data, additional computational data, and empirical rules. *J. Comput. Chem.* **17**, 616–641, https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6<616::AID-JCC5>3.0.CO;2-X (1996).

36. O'Boyle, N. M. *et al.* Open babel: An open chemical toolbox. *J. Cheminformatics* **3**, 33, https://doi.org/10.1186/1758-2946-3-33 (2011).

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

38. Seifert, G., Porezag, D. & Frauenheim, T. Calculations of molecules, clusters, and solids with a simplified LCAO-DFTLDA scheme. *Int. J. Quantum Chem.* **58**, 185–192, https://doi.org/10.1002/(SICI)1097-461X(1996)58:2<185::AID-QUA7>3.0.CO;2-U (1996).

39. Elstner, M. *et al.* Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Phys. Rev. B* **58**, 7260–7268, https://doi.org/10.1103/PhysRevB.58.7260 (1998).

40. Gaus, M., Cui, Q. & Elstner, M. DFTB3: Extension of the self-consistent-charge density-functional tight-binding method (SCC-DFTB). *J. Chem. Theory Comput.* **7**, 931–948, https://doi.org/10.1021/ct100684s (2011).

41. Tkatchenko, A., DiStasio, R. A. Jr., Car, R. & Scheffler, M. Accurate and efficient method for many-body van der Waals interactions. *Phys. Rev. Lett.* **108**, 236402, https://doi.org/10.1103/PhysRevLett.108.236402 (2012).

42. Ambrosetti, A., Reilly, A. M., DiStasio, R. A. Jr. & Tkatchenko, A. Long-range correlation energy calculated from coupled atomic response functions. *J. Chem. Phys.* **140**, 18A508, https://doi.org/10.1063/1.4865104 (2014).

43. Stöhr, M., Michelitsch, G. S., Tully, J. C., Reuter, K. & Maurer, R. J. Communication: Charge-population based dispersion interactions for molecules and materials. *J. Chem. Phys.* **144**, 151101, https://doi.org/10.1063/1.4947214 (2016).

44. Mortazavi, M., Brandenburg, J. G., Maurer, R. J. & Tkatchenko, A. Structure and stability of molecular crystals with manybody dispersion-inclusive density functional tight binding. *J. Phys. Chem. Lett.* **9**, 399–405, https://doi.org/10.1021/acs.jpclett.7b03234 (2018).

45. Gaus, M., Goez, A. & Elstner, M. Parametrization and benchmark of DFTB3 for organic molecules. *J. Chem. Theory Comput.* **9**, 338–354, https://doi.org/10.1021/ct300849w (2013).

46. Gaus, M., Lu, X., Elstner, M. & Cui, Q. Parameterization of DFTB3/3OB for sulfur and phosphorus for chemical and biological applications. *J. Chem. Theory Comput.* **10**, 1518–1537, https://doi.org/10.1021/ct401002w (2014).

47. Aradi, B., Hourahine, B. & Frauenheim, T. DFTB+, a sparse matrix-based implementation of the DFTB method. *J. Phys. Chem. A* **111**, 5678–5684, https://doi.org/10.1021/jp070186p (2007).

48. Larsen, A. H. *et al.* The atomic simulation environment—a python library for working with atoms. *J. Phys. Condens. Matter* **29**, 273002, https://doi.org/10.1088/1361-648x/aa680e (2017).

49. Melander, M., Laasonen, K. & Jónsson, H. Removing external degrees of freedom from transition-state search methods using quaternions. *J. Chem. Theory Comput.* **11**, 1055–1062, https://doi.org/10.1021/ct501155k (2015).

50. Hoja, J. *et al.* QM7-X: a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules (version 2.0). *ZENODO* https://doi.org/10.5281/zenodo.4288677 (2020).

51. Perdew, J. P., Ernzerhof, M. & Burke, K. Rationale for mixing exact exchange with density functional approximations. *J. Chem. Phys.* **105**, 9982–9985, https://doi.org/10.1063/1.472933 (1996).

52. Adamo, C. & Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **110**, 6158–6170, https://doi.org/10.1063/1.478522 (1999).

53. Blum, V. *et al.* Ab initio molecular simulations with numeric atom-centered orbitals. *Comp. Phys. Commun.* **180**, 2175–2196, https://doi.org/10.1016/j.cpc.2009.06.022 (2009).

54. Ren, X. *et al.* Resolution-of-identity approach to Hartree–Fock, hybrid density functionals, RPA, MP2 and GW with numeric atom-centered orbital basis functions. *New J. Phys.* **14**, 053020, https://doi.org/10.1088/1367-2630/14/5/053020 (2012).

55. Tkatchenko, A. & Scheffler, M. Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data. *Phy. Rev. Lett.* **102**, 073005, https://doi.org/10.1103/physrevlett.102.073005 (2009).

56. Ernzerhof, M. & Scuseria, G. E. Assessment of the Perdew–Burke–Ernzerhof exchange-correlation functional. *J. Chem. Phys.* **110**, 5029–5036, https://doi.org/10.1063/1.478401 (1999).

57. Lynch, B. J. & Truhlar, D. G. Robust and affordable multicoefficient methods for thermochemistry and thermochemical kinetics: the MCCM/3 suite and SAC/3. *J. Phys. Chem. A* **107**, 3898–3906, https://doi.org/10.1021/jp0221993 (2003).

58. Reilly, A. M. & Tkatchenko, A. Understanding the role of vibrations, exact exchange, and many-body van der Waals interactions in the cohesive properties of molecular crystals. *J. Chem. Phys.* **139**, 024705, https://doi.org/10.1063/1.4812819 (2013).

59. Hoja, J. *et al.* Reliable and practical computational description of molecular crystal polymorphs. *Sci. Adv.* **5**, eaau3338, https://doi.org/10.1126/sciadv.aau3338 (2019).

60. Stone, A. *The Theory of Intermolecular Forces, Second Edition*. (Oxford Press, Oxford, 2013).

61. Stöhr, M., Medrano Sandonas, L. & Tkatchenko, A. Accurate many-body repulsive potentials for density-functional tight binding from deep tensor neural networks. *J. Phys. Chem. Lett.* **11**, 6835–6843, https://doi.org/10.1021/acs.jpclett.0c01307 (2020).

62. Hourahine, B. *et al.* DFTB+, a software package for efficient approximate density functional theory based atomistic simulations. *J. Chem. Phys.* **152**, 124101, https://doi.org/10.1063/1.5143190 (2020).

Author contributions

J.H. generated the 3D molecular structures with DFTB3+MBD using the HPC facilities of the University of Luxembourg. B.G.E., A.V.M. and R.A.D. performed the PBE0+MBD calculations for all structures using the HPC facilities at the Argonne Leadership Computing Facility. L.M.S. and J.H. designed and wrote the manuscript. R.A.D. and A.T. supervised and revised all stages of the work. All authors discussed the results and contributed to the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-021-00812-2.

Correspondence and requests for materials should be addressed to R.A.D.J. or A.T.

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.