



OPEN

DATA DESCRIPTOR

# QMugs, quantum mechanical properties of drug-like molecules

Clemens Isert<sup>1,4</sup>, Kenneth Atz<sup>1,4</sup>, José Jiménez-Luna<sup>1,2</sup> & Gisbert Schneider<sup>1,3</sup>

Machine learning approaches in drug discovery, as well as in other areas of the chemical sciences, benefit from curated datasets of physical molecular properties. However, there currently is a lack of data collections featuring large bioactive molecules alongside first-principle quantum chemical information. The open-access QMugs (Quantum-Mechanical Properties of Drug-like Molecules) dataset fills this void. The QMugs collection comprises quantum mechanical properties of more than 665 k biologically and pharmacologically relevant molecules extracted from the ChEMBL database, totaling ~2 M conformers. QMugs contains optimized molecular geometries and thermodynamic data obtained via the semi-empirical method GFN2-xTB. Atomic and molecular properties are provided on both the GFN2-xTB and on the density-functional levels of theory (DFT,  $\omega$ B97X-D/def2-SVP). QMugs features molecules of significantly larger size than previously-reported collections and comprises their respective quantum mechanical wave functions, including DFT density and orbital matrices. This dataset is intended to facilitate the development of models that learn from molecular data on different levels of theory while also providing insight into the corresponding relationships between molecular structure and biological activity.

## Background & Summary

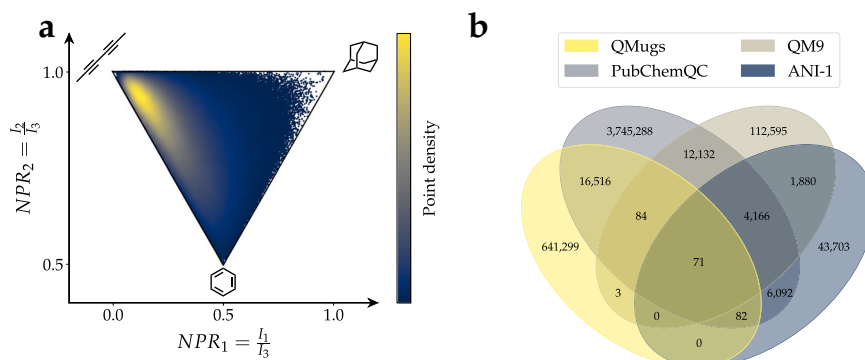
Machine learning methodologies are increasingly becoming well-established tools in many chemistry-related disciplines, such as drug discovery<sup>1</sup>, material science<sup>2</sup>, and physical chemistry<sup>3</sup>. In recent years, significant progress has been made in quantum-based machine learning (QML) methods<sup>4</sup>, which aim to accurately and computationally inexpensively predict the governing properties of atomistic systems, such as energies and forces<sup>5–12</sup>, dipole moments<sup>13</sup>, wave functions<sup>14,15</sup> and electron densities<sup>16,17</sup>. Despite the success and promise surrounding the applicability of such approaches, several challenges remain for QML. Arguably, one of the most important challenges is the increasing need for curated, comprehensive datasets<sup>13</sup>. While several options, such as the QM9<sup>18</sup>, ANI-1<sup>19</sup>, or PubChemQC<sup>20</sup> datasets have paved the way for the development of current-generation QML methods<sup>5–8,21–23</sup>, the computational cost entailed in their generation limits both the scope of the explored chemical space (*e.g.*, molecule size, atom-type diversity), and prospective modeling applicability<sup>13,24</sup>.

There has been a recent surge in interest in the delta-learning ( $\Delta$ -learning) of chemical properties, which aims to use a machine learning model to predict a physically relevant quantity, such as those generated by density-functional theory (DFT) by utilizing information extracted with a computationally cheaper method<sup>22,25</sup> (*e.g.*, semi-empirical approaches such as GFN2-xTB<sup>26–29</sup> and PM6<sup>30</sup>). Datasets that enable this type of learning are scarce and could promote the development of accurate models at potentially a fraction of the computational cost of more precise alternatives<sup>31</sup>. Furthermore, datasets that provide three-dimensional conformational data, for a wide variety of chemical space, at levels of theory higher than classical force fields<sup>32,33</sup>, could boost the performance of machine learning methods in predicting properties from ensembles as well as generative models of conformations. Relevant examples include the PubChemQC-PM6<sup>23</sup> and GEOM<sup>33</sup> datasets, which include molecules with properties computed using different semi-empirical levels of theory. Finally, there is a clear potential to open up new lines of research by combining biological annotations (*e.g.*, from molecular databases such as ChEMBL<sup>34</sup>), and additional QM-derived physical information.

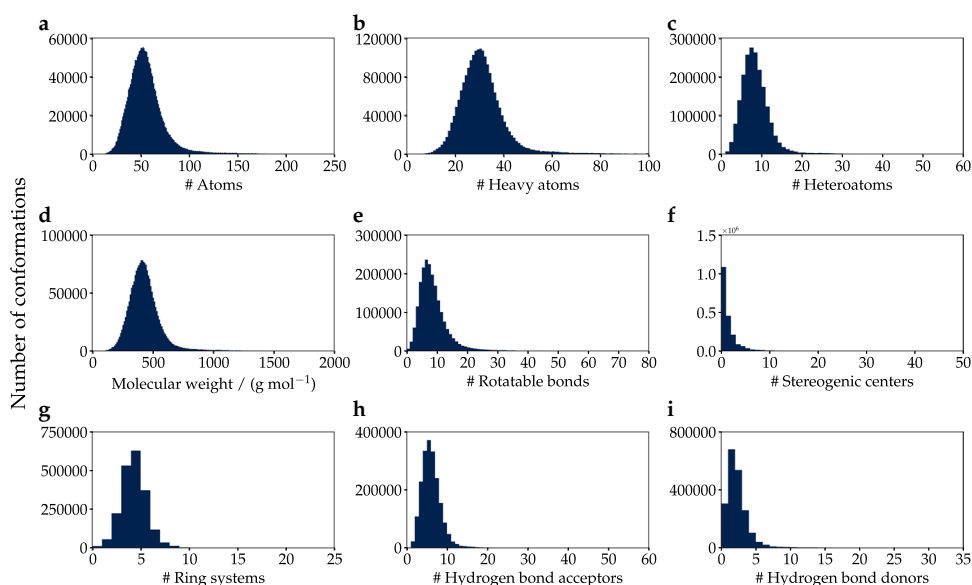
This work introduces QMugs (Quantum-Mechanical Properties of Drug-like Molecules), a data collection of over 665 k curated molecular structures extracted from the ChEMBL database, with accompanying computed quantum

<sup>1</sup>Department of Chemistry and Applied Biosciences, RETHINK, ETH Zurich, 8093, Zurich, Switzerland. <sup>2</sup>Department of Medicinal Chemistry, Boehringer Ingelheim Pharma GmbH & Co. KG, Birkendorfer Straße 65, 88397, Biberach an der Riss, Germany. <sup>3</sup>ETH Singapore SEC Ltd, 1 CREATE Way, #06-01 CREATE Tower, Singapore, 138602, Singapore.

<sup>4</sup>These authors contributed equally: Clemens Isert, Kenneth Atz. ✉e-mail: [jose.jimenez@rethink.ethz.ch](mailto:jose.jimenez@rethink.ethz.ch); [gisbert@ethz.ch](mailto:gisbert@ethz.ch)



**Fig. 1** (a) Principal-moments-of-inertia plot<sup>39</sup> for molecules in the QMugs dataset.  $NPR_x = x$ -th normalized principal moment,  $I_x = x$ -th smallest principal moment of inertia. (b) Venn diagram showing overlap between QMugs and other well-known datasets with DFT-level computed properties: QM9<sup>18</sup>, PubChemQC<sup>20</sup>, and ANI-1<sup>19</sup>. Overlap was computed based on the uniqueness of the InChI representations of the contained molecules. Numbers do not add up to those reported in Table 1 because of InChI strings that occur multiple times.

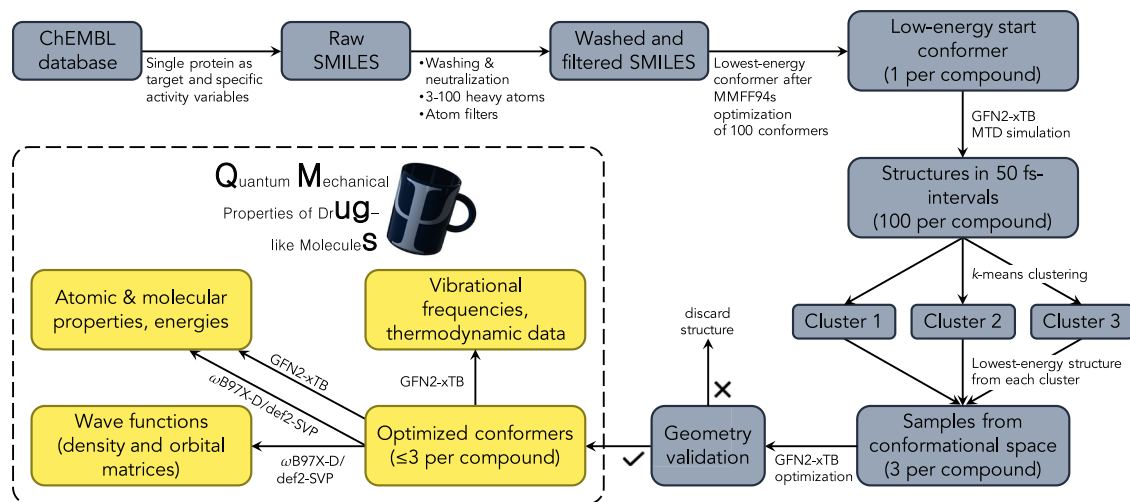


**Fig. 2** Distribution of properties for the molecules contained in the QMugs dataset.

Dataset	Unique compounds	Total conformations	Heavy atoms max (mean)	Method	$\Delta$ -learning possible	Wave functions
QM9	133,885	133,885	9 (8.8)	B3LYP/6-31 G(2df,p)	✗	✗
ANI-1	57,462	22,057,374	8 (7.1)	$\omega$ B97X/6-31 G(d)	✗	✗
PubChemQC	3,982,436	3,982,436	51 (14.1)	B3LYP/6-31 G(d)	✗	✓
QMugs	665,911	1,992,984	100 (30.6)	GFN2-xTB + $\omega$ B97X-D/def2-SVP	✓	✓

**Table 1.** Descriptive statistics of the dataset reported herein in the context of other DFT-level molecular datasets and the information provided by each. The number of molecules for PubChemQC corresponds to that available on the website of the project<sup>57</sup>. Heavy atom averages are weighted by the number of conformations.

mechanical properties. Different levels of theory were combined in these calculations. Per compound, three conformers were generated, and their geometries were optimized using the semi-empirical GFN2-xTB method<sup>26-29</sup>, whereas a comprehensive array of quantum properties was computed at the DFT level of theory using the  $\omega$ B97X-D functional<sup>35</sup> and the def2-SVP Karlsruhe basis set<sup>36</sup>. The data collection presented herein is put in the context of other sets that also feature DFT-level properties. A descriptive evaluation against the QM9<sup>18</sup>, ANI-1<sup>19</sup>, and PubchemQC<sup>20</sup>



**Fig. 3** Overview of the data generation process. Molecules were extracted from the ChEMBL database, standardized, and filtered, and starting conformers were generated using the RDKit software package. Metadynamics (MTD) simulations were performed using the GFN2-xTB semi-empirical method to generate three diverse conformations before final geometry optimization. Molecules that did not pass a series of geometric sanity checks were removed. DFT-level properties ( $\omega$ B97X-D/def2-SVP) were computed using Psi4 software.

datasets is provided in Fig. 1b, as well as in Table 1. With an average of 30.6 and a maximum of 100 heavy atoms per compound (Table 1 & Fig. 2), QMugs features molecular samples that are considerably larger than those provided by other previously-reported datasets. QMugs also provides a larger number of distinct molecules than QM9 and ANI-1. Though the total number of provided molecules in QMugs is lower than that provided in PubChemQC, QMugs provides multiple conformers per molecule and therefore enables the training of QML models which can differentiate between molecular constitution and conformation. QMugs additionally provides a wide range of properties on multiple levels of theory. Furthermore, the vast majority of the included compounds (~641 k, 96.3%) were previously unreported in other DFT data collections, while also providing equivalent information at additional levels of theory, namely GFN2-xTB. QMugs provides quantum mechanical wave functions represented as local bases of atomic orbitals (*i.e.*, DFT density and orbital matrices). Single-point properties as well as wave functions were computed with the Psi4 software suite<sup>37</sup> for all the conformers (~2.0M) present in the database. As previously reported for the ChEMBL database<sup>38</sup>, most of the considered drug-like molecules in this study fall within the rod-disk axis in the principal moments of inertia plot<sup>39</sup> (Fig. 1a).

Overall, the potential utility of the presented dataset is fourfold: (i) it will provide researchers with a dataset containing substantially larger molecules than previously reported, in order to either directly predict quantum chemical properties, or to learn a property mapping between two popular quantum mechanical levels of theory (*i.e.*, GFN2-xTB and  $\omega$ B97X-D/def2-SVP); (ii) it will facilitate the development of novel machine-learning methodologies for the generation of molecular conformations and molecular property predictions via their ensembles; (iii) it will facilitate the development of novel deep learning frameworks for the prediction of the quantum mechanical wave function in a local basis of atomic orbitals; and (iv) it will enable research towards the exploration of quantum featurization in the context of pharmacologically relevant, annotated biological data.

## Methods

Molecules were extracted from the ChEMBL database<sup>34</sup> (version 27). Conformers were generated using RDKit (<http://www.rdkit.org>) and GFN2-xTB<sup>26–29</sup>. DFT ( $\omega$ B97X-D/def2-SVP) calculations were carried out via Psi4<sup>37</sup>. A similar approach was adopted in a previous study on transition-metal complexes<sup>13</sup>. An overview of the data processing pipeline is given in Fig. 3, while individual steps are described in more detail in the following subsections.

In chemical terminology, the term “conformation” refers to any arrangement of atoms in space, whereas “conformer” refers to a conformation that is a local minimum on the potential energy surface of the molecule<sup>40</sup>. In the analyses that follow, the term “conformation” is loosely used to refer to both, unless explicitly mentioned otherwise.

**Data extraction and SMILES processing.** Single-protein targets with assay information for at least 10 compounds with unique internal identifiers were extracted from the ChEMBL database. Several activity and annotation filters were subsequently applied to these compounds (see Supporting Information for a detailed query description). This procedure resulted in 685,917 molecules with unique external identifiers (ChEMBL-IDs), represented by their Simplified Molecular Input Line Entry Specification (SMILES)<sup>41</sup>. Molecules were neutralized, and salts and solvents were removed (“washing”) using the ChEMBL Structure Pipeline package<sup>42</sup>. For compounds consisting of multiple separate fragments, all fragments except the one with the highest number of heavy atoms were discarded after the washing. Additionally, molecules containing fewer than 3 or more than 100

heavy atoms, as well as radical species and molecules with a net charge different from zero after the attempted neutralization, were removed. Atom types included in the QMugs dataset are hydrogen, carbon, nitrogen, oxygen, fluorine, phosphorus, sulfur, chlorine, bromine, and iodine.

**Conformer generation and optimization.** With the procedure described herein, a compromise between efficient molecular conformational search and practical computational expense considerations was sought. Similar to previous studies<sup>13,43</sup>, the semiempirical GFN2-xTB method was used as a surrogate for full DFT geometry optimization, as the latter is associated with challenging computational costs when considering the size and number of molecules in the QMugs dataset.

The RDKit (<http://www.rdkit.org>) implementation of the Experimental-Torsion Knowledge Distance Geometry (ETKDG) method<sup>44</sup> was used to generate up to 100 conformers for each molecule, with a maximum of 1000 embedding attempts and an initial coordinate assignment using distance-matrix eigenvalues and default settings (`boxSizeMult=2.0`, `forceTol=1e-3`). Upon no successful conformer generation, it was re-attempted via random assignment of the starting coordinates. The resulting conformers were further minimized using the Merck molecular force field<sup>45</sup> (MMFF94s) for a maximum of 1000 iterations, with default settings (`nonBondedThresh = 100.0`). The lowest-energy conformer (according to the selected force field) for each structure was then used as a starting point for meta-dynamics (MTD) simulations. Stereocenters that were previously undefined in the SMILES extraction procedure were assigned in this conformer generation process.

For each generated conformer, an MTD simulation was performed with the xTB software package<sup>28</sup> for a duration of 5 ps with time steps of 1 fs, at a temperature of 300 K. The biasing root-mean-square deviation (RMSD) potential used for all MTD simulations is given by  $E_{\text{bias}}^{\text{RMSD}} = \sum_{i=1}^N k_i \exp(\alpha \Delta_i^2)$ , where  $N$  is the number of reference structures,  $k_i$  the pushing strength,  $\Delta_i$  the collective variable (*i.e.*, the RMSD between structure  $i$  and a reference structure), and  $\alpha$  the width of the Gaussian potential used in the RMSD criterion. Simulations were carried out with  $\alpha = 1.2^{-1}$  and  $k_i = 0.2$  mEh with snapshots taken every 50 fs, resulting in 100 conformations stored with their corresponding energies. To obtain conformationally diverse samples, these structures were subsequently clustered into three groups via the  $k$ -means<sup>46</sup> algorithm, as implemented in the scikit-learn<sup>47</sup> (version 0.23.1) Python package using the pairwise RMSD of the aligned structures as molecular features. The conformation with the lowest-energy value from each cluster was then selected for further processing. The three resulting conformations for each molecule were then optimized using the GFN2-xTB<sup>26–29</sup> method using energy and gradient convergence criteria of  $5 \times 10^{-6}$  Eh and  $1 \times 10^{-3}$  Eh  $\alpha^{-1}$ , respectively, and the approximate normal coordinate rational function optimizer (ANCopt). Harmonic frequencies, entropies, enthalpies and heat capacities at 298.15 K were extracted at the end of the geometry optimization process. Structures for which vibrational frequencies with imaginary wave numbers were obtained — indicative of failure to reach energy minima — were subjected to additional optimizations until no significant ones remained, up to a maximum of 100 attempts.

**Quantum mechanical calculations.** Single-point electronic calculations were performed for the optimized geometries using the  $\omega$ B97X-D quantum functional and the def2-SVP basis set as implemented in the open-source quantum-chemistry software suite Psi4<sup>37</sup>. Single-point properties such as formation and orbital energies, dipole moments, rotational constants, partial charges, bond orders, valence numbers, as well as wave functions including  $\alpha$  and  $\beta$  DFT-density matrices, orbital matrices, and the atomic-orbital-to-symmetry-orbital transformer matrix were obtained. For practical reasons, 52 structures whose DFT calculations required computational resources that exceeded empirically determined limits, or for which calculations were unsuccessful, were discarded (see Supporting Information for details).

## Data Records

All computed molecular structures, as well as their corresponding properties and wave functions are accessible through the ETH Library Collection service<sup>48</sup>.

**Format specification.** A `summary.csv` comma-separated file contains computed molecular-level properties and additional annotations. A compressed tarball file (`structures.tar.gz`) of ~7 gigabytes (GB) contains plain MDL structure-data files<sup>49</sup> (SDFs) with embedded atomic and molecular properties, grouped in sub-directories according to their respective ChEMBL identifiers. These SDFs include single-point electronic properties calculated on the GFN2-xTB and  $\omega$ B97X-D/def-SVP levels of theory, as described in Table 2. A second compressed tarball file (`vibspectra.tar.gz`, ~3 GB) contains vibrational spectra.

Wave function files (*i.e.*, DFT density and orbital matrices) as described in Table 3, are split into 100 compressed tarballs (`wfn_xx.tar.gz`) of ~50 GB each for easier management and downloading. These are supplied as NumPy<sup>50</sup> (`.npy`) binary files, which can be read using the Psi4 software package. Molecules (with all conformers grouped together) were assigned at random to the tarballs to enable easy use of subsets of the QMugs dataset without having to download all the files. The assignment of ChEMBL identifiers to tarballs is described in a `tarball_assignment.csv` file.

## Technical Validation

**Optimized geometry sanity checks.** Four consecutive geometry checks were performed to filter out structures for which the geometry optimization procedure converged to unrealistic conformations. To determine suitable thresholds for removing a structure from our dataset, the generated geometries were compared to experimental reference values and to DFT-optimized geometries extracted from the PubChemQC dataset<sup>20</sup>. Specifically, we investigated (i) the deviation of bond lengths from experimental reference values, (ii) isomorphism between the initial molecular graphs and those obtained after geometry optimization, (iii) linearity of triple bonds, and

Property	Symbol	Unit	Key	$\Delta$ -ML
ChEMBL identifier	—	—	CHEMBL_ID	
Conformer identifier	—	—	CONF_ID	
Total energy	$U_{RT}$	$E_h$	GFN2:TOTAL_ENERGY	◆
Internal atomic energy	$E_{Atom}$	$E_h$	GFN2:ATOMIC_ENERGY	
Formation energy	$E_{Form}$	$E_h$	GFN2:FORMATION_ENERGY	◆
Total enthalpy	$H_{RT}$	$E_h$	GFN2:TOTAL_ENTHALPY	
Total free energy	$G_{RT}$	$E_h$	GFN2:TOTAL_FREE_ENERGY	
Dipole (x, y, z, total)	$\mu$	D	GFN2:DIPOLE	◆
Quadrupole (xx, xy, yy, xz, yz, zz)	$Q_{ij}$	D Å	GFN2:QUADRUPOLE	
Rotational constants (A, B, C)	A, B, C	cm <sup>-1</sup>	GFN2:ROT_CONSTANTS	◆
Enthalpy (vib., rot., transl., total)	$\Delta H$	cal mol <sup>-1</sup>	GFN2:ENTHALPY	
Heat capacity (vib., rot., transl., total)	$C_V$	cal K <sup>-1</sup> mol <sup>-1</sup>	GFN2:HEAT_CAPACITY	
Entropy (vib., rot., transl., and total)	$\Delta S$	cal K <sup>-1</sup> mol <sup>-1</sup>	GFN2:ENTROPY	
HOMO energy	$E_{HOMO}$	$E_h$	GFN2:HOMO_ENERGY	◆
LUMO energy	$E_{LUMO}$	$E_h$	GFN2:LUMO_ENERGY	◆
HOMO-LUMO gap	$E_{Gap}$	$E_h$	GFN2:HOMO_LUMO_GAP	◆
Fermi level	$E_{Fermi}$	$E_h$	GFN2:FERMI_LEVEL	
Mulliken partial charges	$\delta_M$	e	GFN2:MULLIKEN_CHARGES	◆
Covalent coordination number	$N_{coord}$	—	GFN2:COVALENT_COORDINATION_NUMBER	
Molecular dispersion coefficient	$C_6$	a.u.	GFN2:DISPERSION_COEFFICIENT_MOLECULAR	
Atomic dispersion coefficients	$C_6$	a.u.	GFN2:DISPERSION_COEFFICIENT_ATOMIC	
Molecular polarizability	$\alpha(0)$	a.u.	GFN2:POLARIZABILITY_MOLECULAR	
Atomic polarizabilities	$\alpha(0)$	a.u.	GFN2:POLARIZABILITY_ATOMIC	
Wiberg bond orders	$M_{AB}$	—	GFN2:WIBERG_BOND_ORDER	◆
Total Wiberg bond orders	$\sum_{A(A \neq B)} M_{AB}$	—	GFN2:TOTAL_WIBERG_BOND_ORDER	◆
Total energy	$U_{RT}$	$E_h$	DFT:TOTAL_ENERGY	◆
Total internal atomic energy	$E_{Atom}$	$E_h$	DFT:ATOMIC_ENERGY	
Formation energy	$E_{Form}$	$E_h$	DFT:FORMATION_ENERGY	◆
Electrostatic potential	$V_{ESP}$	V	DFT:ESP_AT_NUCLEI	
Löwdin partial charges	$\delta_L$	e	DFT:LOWDIN_CHARGES	
Mulliken partial charges	$\delta_M$	e	DFT:MULLIKEN_CHARGES	◆
Rotational constants (A, B, C)	A, B, C	cm <sup>-1</sup>	DFT:ROT_CONSTANTS	◆
Dipole (x, y, z, total)	$\mu$	D	DFT:DIPOLE	
Exchange correlation energy	$E_{XC}$	$E_h$	DFT:XC_ENERGY	
Nuclear repulsion energy	$\hat{V}_{eN}$	$E_h$	DFT:NUCLEAR_REPULSION_ENERGY	
One-electron energy	$\hat{T}_e$	$E_h$	DFT:ONE_ELECTRON_ENERGY	
Two-electron energy	$\hat{V}_{ee}$	$E_h$	DFT:TWO_ELECTRON_ENERGY	
HOMO energy	$E_{HOMO}$	$E_h$	DFT:HOMO_ENERGY	◆
LUMO energy	$E_{LUMO}$	$E_h$	DFT:LUMO_ENERGY	◆
HOMO-LUMO gap	$E_{Gap}$	$E_h$	DFT:HOMO_LUMO_GAP	◆
Mayer bond orders	$M_{AB}$	—	DFT:MAYER_BOND_ORDER	
Wiberg-Löwdin bond orders	$W_{AB}$	—	DFT:WIBERG_LOWDIN_BOND_ORDER	◆
Total Mayer bond orders	$\sum_{A(A \neq B)} M_{AB}$	—	DFT:TOTAL_MAYER_BOND_ORDER	
Total Wiberg-Löwdin bond orders	$\sum_{A(A \neq B)} W_{AB}$	—	DFT:TOTAL_WIBERG_LOWDIN_BOND_ORDER	◆

**Table 2.** Calculated properties as stored in the SDFs of the QMugs data collection. Abbreviations: a.u., atomic units; vib., vibrational; rot., rotational; transl., translational. Properties that enable  $\Delta$  machine learning are labelled with ◆.

(iv) planarity of aromatic rings. Structures were removed from the dataset if they failed any of these tests. In total, 10,986 (0.55%) conformations were discarded from the dataset. Each test is briefly described in the following subsections, with further technical details reported in the Supporting Information.

*Deviation of bond lengths from experimental reference values.* Bond lengths in the optimized structures were compared to average experimental reference values for bonds of the same bond type (single, double, triple, or aromatic) and between the same atoms. Reference values were obtained from the Computational Chemistry

Property	Symbol	Key
Alpha density matrix	$D_\alpha$	matrix, Da
Beta density matrix	$D_\beta$	matrix, Db
Alpha orbitals	$C_\alpha$	matrix, Ca
Beta orbitals	$C_\beta$	matrix, Cb
Atomic-orbital-to-symmetry-orbital transformer	$C_{\text{AOTOSO}}$	matrix, aotoso
Mayer bond orders	$M_{AB}$	MAYER_INDICES
Wiberg-Löwdin bond orders	$W_{AB}$	WIBERG_LOWDIN_INDICES

**Table 3.** Calculated molecular properties stored in the wave function files provided in the QMugs data collection. Mayer and Wiberg-Löwdin bond orders included here represent a superset of the bond orders in the SDFs which additionally comprise bond orders for non-covalent bonds.

Comparison and Benchmark DataBase (CCCBDB)<sup>51</sup>, and the largest absolute bond-length deviation from reference values was recorded per molecule. Bonds for which no reference value was available (0.75%) were omitted. The same analysis was carried out for molecules from the PubChemQC dataset containing the same atom types as QMugs, in order to obtain a comparable set with respect to the present atom types. The PubChemQC set (3,834,382 conformations with reference bond lengths) showed a deviation of  $0.06 \pm 0.04 \text{ \AA}$  (median  $\pm 1$  standard deviation), whereas the QMugs dataset (2,004,003 conformations with reference bond lengths) showed a deviation of  $0.07 \pm 0.03 \text{ \AA}$ . Based on the observed distribution of bond-length deviations from experimental reference values (Figure S1) and manual investigation of example structures, 0.2 was determined to be a suitable threshold for a conformation to be removed from the dataset, which included 6,131 (0.31%) examples.

**Molecular graph isomorphism.** It was investigated whether atom connectivity could be reconstructed after removing bond information from the generated SDFs. To this end, molecular graphs constructed exclusively from atom positions and types were compared to those obtained using the original atom connectivity (see Supporting Information for details). 1,568 (0.08%) conformations for which the resulting molecular graphs were non-isomorphic failed this test.

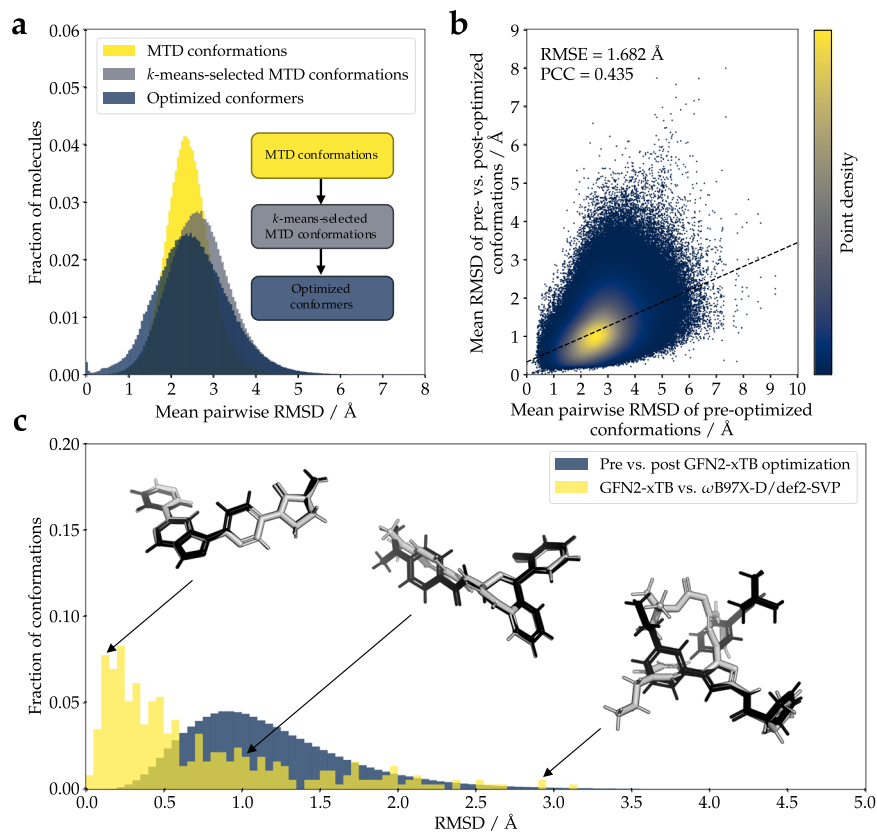
**Deviation of triple bonds from linear geometry.** The deviation of triple bonds from their ideal linear geometry was examined. In this investigation, ring triple bonds were not considered owing to routinely-occurring deviations from linear geometry in systems with high ring strain<sup>52</sup>. The largest deviation from a 180 (linear) bond angle was recorded for each molecule containing at least one non-ring triple bond. The same analysis was performed on the PubChemQC dataset<sup>20</sup>. Triple-bond-containing molecules from PubchemQC and QMugs (273,320 and 165,101 samples, respectively) showed deviations of  $1.38 \pm 1.46^\circ$  (median  $\pm 1$  standard deviation), and  $1.46 \pm 2.13^\circ$ , respectively. Based on the observed distribution of triple bond angles from a linear geometry (Figure S2) and manual inspection of structures, a  $10^\circ$  deviation was identified as a suitable threshold. 1,147 (0.06%) conformations failed this test.

**Deviation of aromatic rings from planar geometry.** The planarity of carbon-containing aromatic rings was also investigated. For each molecule containing aromatic carbon atoms, the largest dihedral angle between the two planes spanned by each aromatic carbon atom and its three neighbors was recorded (see Supporting Information for details). The same analysis was performed on the PubChemQC dataset<sup>20</sup>. Molecules from PubchemQC and QMugs (2,391,589 and 1,950,929 conformations with aromatic carbons, respectively) showed median dihedral angles ( $\pm 1$  standard deviation) of  $1.70 \pm 1.85^\circ$  and  $2.99 \pm 2.20^\circ$ , respectively. Based on the observed distribution of dihedral angles from planar geometries (Figure S3) and manual inspection of structures, 2,769 (0.14%) conformations with aromatic carbon dihedral angles above  $15^\circ$  were discarded.

**Further geometrical assessment.** The changes in the molecular geometries along the applied pipeline were examined in order to evaluate the effects of the applied steps. Figure 4a shows the mean pairwise RMSD of atom positions between the conformations of each molecule at different steps along the pipeline. Conformations sampled during MTD simulations showed a mean pairwise RMSD of  $2.40 \pm 0.52^\circ$  (median  $\pm 1$  standard deviation). The *k*-means clustering procedure accomplished the envisaged task of sampling conformations with higher geometric diversity ( $2.67 \pm 0.74^\circ$ ). During the geometry optimization process, conformational diversity decreased ( $2.48 \pm 0.86^\circ$ ). Unsurprisingly, for some molecules featuring rigid structures, conformations tended to converge toward the same energy minimum (0.09% of molecules show a mean pairwise RMSD  $< 0.01^\circ$  between their optimized conformers).

The degree to which the molecular geometries changed during the final optimization step was further analyzed. Molecules with initially more diverse conformations (higher mean pairwise RMSD of pre-optimized conformations) were shown to undergo a greater change in atom positions (mean RMSD of pre- vs. post-optimized conformations) during optimization with the GFN2-xTB method (Fig. 4b). The observed heteroscedastic behavior of these two properties indicates that while the mean RMSD of pre- vs. post-optimized conformations tends to increase with higher mean pairwise RMSD of pre-optimized conformations, its variance also increases.

Finally, the suitability of GFN2-xTB as a lower-cost surrogate for DFT-level geometry optimization (Fig. 4c) was confirmed. 500 randomly-chosen structures prior to semi-empirical geometry optimization from the QMugs dataset were further subjected to DFT-level geometry optimization ( $\omega$ B97X-D/def2-SVP), discarding structures that could not be converged in 100 iterations or with the computational resources described in the

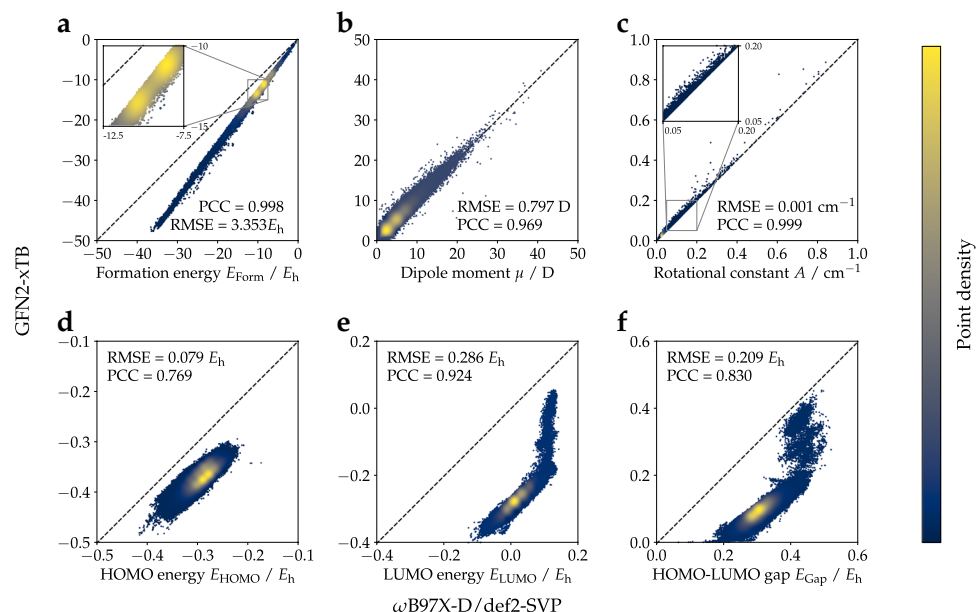


**Fig. 4** (a) Distributions of mean pairwise RMSD of atom positions between conformations of each molecule in the QMugs dataset at different stages along the pipeline. While the *k*-means sampling process selects conformations that are, on average, more geometrically diverse than the average pair of structures generated by MTD simulations, geometry optimization reduces the geometrical diversity between the optimized conformers. (b) Change in atom positions during geometry optimization vs. mean pairwise RMSD of conformations before optimization. Molecules with initially more diverse conformations displayed a greater change in atom positions than those with initially less diverse conformations. (c) Distribution of RMSD of structures prior to and after optimization with the semi-empirical GFN2-xTB method, and of structures optimized with the same approach vs. with  $\omega$ B97X-D/def2-SVP. The structures of three molecules with varying differences between the two methods are shown as illustrative examples (black and gray correspond to GFN2-xTB and  $\omega$ B97X-D/def2-SVP-optimized structures, respectively). For illustrative purposes, the example molecules are aligned on their substructures.

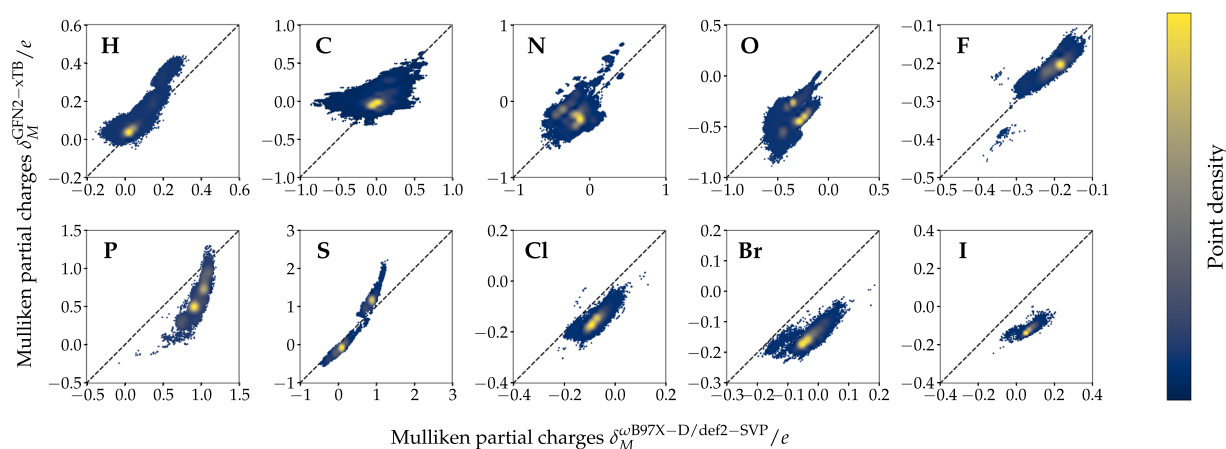
Supporting Information. The RMSDs between the structures independently optimized at both levels of theory were then measured. The pairs of structures showed RMSDs of  $0.47 \pm 0.63^\circ$  (median  $\pm$  1 standard deviation), indicating that the chosen semi-empirical method obtains similar geometries to those obtained with more expensive first-principle calculations. Large RMSDs in some example pairs (Fig. 4c) could be interpreted as indicative of convergence to distinct local minima.

For 2,067 molecules, their individual conformations have different SMILES describing two different (*E*)/(*Z*) isomers. Those structures are either  $\alpha$ - $\beta$ -unsaturated ketones,  $\alpha$ - $\beta$ -unsaturated nitriles, imine, or azo compounds, for which isomerization might be plausible<sup>53,54</sup>. In part due to the applied “washing” procedure, 17,176 molecules can be represented with a SMILES string that is shared with at least one other ChEMBL-ID.

**Validation of single-point properties.** To validate the general agreement between the two methods employed in this work, the correlation between a series of single-point properties computed on both levels of theory was analyzed. Both global molecular (Fig. 5) and local atomic/bond properties (Fig. 6, 7) were considered. All single-point molecular properties showed a high degree of correlation. Formation energies  $E_{\text{Form}}$  (Fig. 5a), which were obtained by subtracting atomic energies  $E_{\text{Atom}}$  from total internal energies  $U_{\text{RT}}$ , show a Pearson correlation coefficient (PCC) of 0.998. Dipole moments  $\mu$  and rotational constants *A* (excl. 22 small structures with very high rotational constants; Fig. 5b,c) displayed PCCs of 0.969 and 0.999, respectively. Orbital energies, namely the energies for highest occupied (HOMO)  $E_{\text{HOMO}}$  and lowest unoccupied molecular orbitals (LUMO)  $E_{\text{LUMO}}$  and HOMO-LUMO gap energies  $E_{\text{Cap}}$  showed PCCs of 0.769, 0.924 and 0.830, respectively (Fig. 5d–f). The observed PCCs for all six single-point molecular properties indicate good agreement between the two methods. Atom-type-specific partial charges for the 10 atom-types in QMugs (Fig. 6, Table S1) as well as the 15 most abundant covalent bond types (Fig. 7, Table S2) also showed high correlations



**Fig. 5** Comparison of molecular properties computed at the two levels of theory considered herein (GFN2-xTB,  $\omega$ B97X-D/def2-SVP) for the molecules contained in QMugs. The molecular formation energy  $E_{\text{Form}}$  in (a) was calculated by subtracting the atomic  $U_{\text{Atom}}$  contributions from the total molecular energies  $U_{\text{RT}}$ . Only the rotational constants  $A$  are shown in (c) as their  $B$  and  $C$  counterparts showed highly similar values. 22 conformations of small molecules show very large rotational constants and are not shown. RMSE and PCC for rotational constant  $A$  are  $845.834 \text{ cm}^{-1}$  and  $0.091$  respectively, if those structures are included. Abbreviations: RMSE, root mean squared error; PCC, Pearson's correlation coefficient.



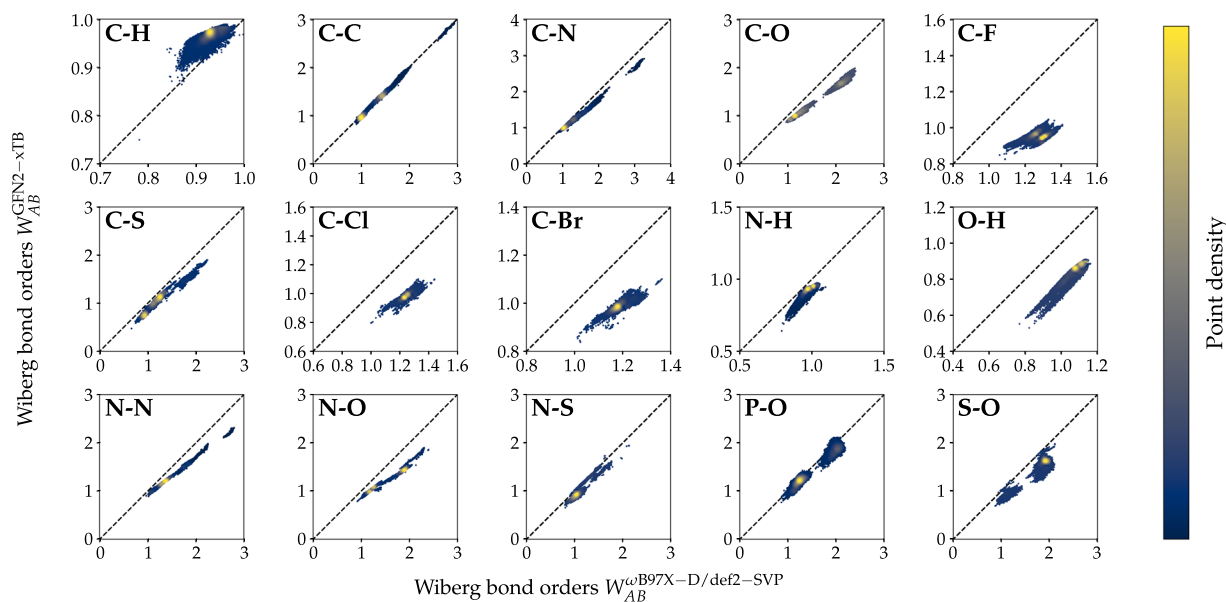
**Fig. 6** Atom-type-specific partial charge correlations (GFN2-xTB,  $\omega$ B97X-D/def2-SVP) for the QMugs dataset (see Table S1 in the Supporting Information for additional metrics).

between the two methods used herein. Regarding partial charges, 7 out of the 10 atom types considered in QMugs were observed to have PCCs  $> 0.8$ , with the remaining carbon, nitrogen, and oxygen atom-types resulting in lower PCCs of 0.574, 0.124, and 0.274, respectively. Regarding bond orders, 10 out of the 15 show PCCs  $> 0.9$  and 14 out of 15 displayed PCCs  $> 0.75$  (see Table S2 for additional metrics). Notably only carbon-fluorine bonds displayed a larger discrepancy between both levels of theory, with an observed PCC of 0.153. The observed correlations in both molecular and atomic single-point properties between GFN2-xTB and  $\omega$ B97X-D/def2-SVP confirm the suitability of the former method as a computationally affordable starting point for  $\Delta$ -learning of DFT-level properties.

### Usage Notes

All data files can be accessed via any modern web browser, and can be programmatically downloaded using the provided instructions in the archive's readme. The provided SDFs can be processed using standard cheminformatics software (for example, RDKit, KNIME<sup>55</sup>), and wave function files using the Psi4<sup>37</sup> software package or directly using Numpy<sup>50</sup>.





**Fig. 7** Comparison of Wiberg bond orders between GFN2-xTB and  $\omega$ B97X-D/def2-SVP for the 15 most frequently occurring bond types in the QMugs dataset. The latter level of theory uses Löwdin-orthogonalization. See Table S2 in the Supporting Information for additional metrics. For bond types which occurred  $> 1$  M times in the dataset, a randomly chosen sample of 1 M bonds is plotted.

### Code availability

All analyses were supported by the Python programming language (version 3.7.7) and its scientific software stack<sup>50</sup>. Molecular conformations were generated using RDKit (<http://www.rdkit.org>, version 2020.03.3) and GFN2-xTB<sup>26–29</sup> (version 6.3.1). All quantum mechanical calculations were carried out with Psi4<sup>37</sup> (version 1.3.2). Molecular structure visualizations were created using PyMol<sup>56</sup> (version 2.3.5) and ChemDraw (version 19.1.1.32). The rclone (<https://rclone.org>, version 1.54.0) WebDAV client was used for all data uploading purposes.

Received: 15 August 2021; Accepted: 17 May 2022;

Published online: 07 June 2022

### References

- Gawehn, E., Hiss, J. A. & Schneider, G. Deep learning in drug discovery. *Mol. Inform.* **35**, 3–14 (2016).
- Schmidt, J., Marques, M. R., Botti, S. & Marques, M. A. Recent advances and applications of machine learning in solid-state materials science. *Npj Comput. Mater.* **5**, 83 (2019).
- von Lilienfeld, O. A. Quantum machine learning in chemical compound space. *Angew. Chem. Int. Ed.* **57**, 4164–4169 (2018).
- von Lilienfeld, O. A., Müller, K.-R. & Tkatchenko, A. Exploring chemical compound space with quantum-based machine learning. *Nat. Rev. Chem.* **4**, 347–358 (2020).
- Satorras, V. G., Hoogeboom, E. & Welling, M. E(n) equivariant graph neural networks. In *International Conference on Machine Learning*, 9323–9332 (PMLR, 2021).
- Schütt, K., Unke, O. & Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International Conference on Machine Learning*, 9377–9388 (PMLR, 2021).
- Huang, B. & von Lilienfeld, O. A. Quantum machine learning using atom-in-molecule-based fragments selected on the fly. *Nat. Chem.* **12**, 945–951 (2020).
- Christensen, A. S., Bratholm, L. A. & Faber, F. A. & Anatole von Lilienfeld, O. FCHL revisited: Faster and more accurate quantum machine learning. *J. Chem. Phys.* **152**, 044107 (2020).
- Heinen, S., von Rudorff, G. F. & von Lilienfeld, O. A. Toward the design of chemical reactions: Machine learning barriers of competing mechanisms in reactant space. *J. Chem. Phys.* **155**, 064105 (2021).
- Heinen, S., Schwilk, M., von Rudorff, G. F. & von Lilienfeld, O. A. Machine learning the computational cost of quantum chemistry. *Mach. Learn.: Sci. Technol.* **1**, 025002 (2020).
- 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 (2019).
- Faber, F. A., Christensen, A. S. & Huang, B. & Von Lilienfeld, O. A. Alchemical and structural distribution based representation for universal quantum machine learning. *J. Chem. Phys.* **148**, 241717 (2018).
- Balcells, D. & Skjelstad, B. B. tmQM dataset-quantum geometries and properties of 86k transition metal complexes. *J. Chem. Inf. Model.* **60**, 6135–6146 (2020).
- Unke, O. *et al.* SE(3)-equivariant prediction of molecular wavefunctions and electronic densities. *Advances in Neural Information Processing Systems* **34** (2021).
- Schütt, K., Gastegger, M., Tkatchenko, A., Müller, K.-R. & Maurer, R. J. Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nat. Commun.* **10** (2019).
- Grisafi, A. *et al.* Transferable machine-learning model of the electron density. *ACS Cent. Sci.* **5**, 57–64 (2018).
- Fabrizio, A., Grisafi, A., Meyer, B., Ceriotti, M. & Corminboeuf, C. Electron density learning of non-covalent systems. *Chem. Sci.* **10**, 9424–9432 (2019).

18. Ramakrishnan, R., Dral, P. O. & Rupp, M. & Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **1**, 140022 (2014).
19. 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 (2017).
20. Nakata, M. & Shimazaki, T. PubChemQC project: A large-scale first-principles electronic structure database for data-driven chemistry. *J. Chem. Inf. Model.* **57**, 1300–1308 (2017).
21. 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 (2017).
22. 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**, 134 (2020).
23. Nakata, M., Shimazaki, T., Hashimoto, M. & Maeda, T. PubChemQC PM6: Data sets of 221 million molecules with optimized molecular geometries and electronic properties. *J. Chem. Inf. Model.* **60**, 5891–5899 (2020).
24. Glavatskikh, M., Leguy, J., Hunault, G., Cauchy, T. & Da Mota, B. Dataset's chemical diversity limits the generalizability of machine learning predictions. *J. Cheminformatics* **11**, 1–15 (2019).
25. Qiao, Z., Welborn, M., Anandkumar, A., Manby, F. R. & Miller, T. F. III Orbnet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features. *J. Chem. Phys.* **153**, 124111 (2020).
26. Grimme, S., Bannwarth, C. & Shushkov, P. A robust and accurate tight-binding quantum chemical method for structures, vibrational frequencies, and noncovalent interactions of large molecular systems parametrized for all spd-block elements ( $Z = 1-86$ ). *J. Chem. Theory Comput.* **13**, 1989–2009 (2017).
27. 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. *J. Chem. Theory Comput.* **15**, 1652–1671 (2019).
28. Grimme, S. Exploration of chemical compound, conformer, and reaction space with meta-dynamics simulations based on tight-binding quantum chemical calculations. *J. Chem. Theory Comput.* **15**, 2847–2862 (2019).
29. Bannwarth, C. *et al.* Extended tight-binding quantum chemistry methods. *WIREs Comput. Mol. Sci.* **11**, e1493 (2021).
30. Rezac, J., Fanfrik, J., Salahub, D. & Hobza, P. Semiempirical quantum chemical PM6 method augmented by dispersion and H-bonding correction terms reliably describes various types of noncovalent complexes. *J. Chem. Theory Comput.* **5**, 1749–1760 (2009).
31. Folmsbee, D. & Hutchison, G. Assessing conformer energies using electronic structure and machine learning methods. *Int. J. Quantum Chem.* **121**, e26381 (2021).
32. Bolton, E. E., Kim, S. & Bryant, S. H. PubChem3D: Conformer generation. *J. Cheminformatics* **3**, 4 (2011).
33. Axelrod, S. & Gomez-Bombarelli, R. GEOM: Energy-annotated molecular conformations for property prediction and molecular generation. *arXiv preprint arXiv:2006.05531* (2020).
34. Mendez, D. *et al.* ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Res.* **47**, D930–D940 (2019).
35. Chai, J.-D. & Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Phys. Chem. Chem. Phys.* **10**, 6615–6620 (2008).
36. Weigend, F. & Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **7**, 3297–3305 (2005).
37. Smith, D. G. *et al.* Psi4 1.4: Open-source software for high-throughput quantum chemistry. *J. Chem. Phys.* **152**, 184108 (2020).
38. Meyers, J., Carter, M., Mok, N. Y. & Brown, N. On the origins of three-dimensionality in drug-like molecules. *Future Med. Chem.* **8**, 1753–1767 (2016).
39. Sauer, W. H. & Schwarz, M. K. Molecular shape diversity of combinatorial libraries: A prerequisite for broad bioactivity. *J. Chem. Inf. Comput. Sci.* **43**, 987–1003 (2003).
40. Moss, G. *et al.* Basic terminology of stereochemistry (IUPAC recommendations 1996). *Pure Appl. Chem.* **68**, 2193–2222 (1996).
41. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
42. Bento, A. P. *et al.* An open source chemical structure curation pipeline using rdkit. *J. Cheminformatics* **12**, 1–16 (2020).
43. Christensen, A. S. *et al.* Orbnet Denali: A machine learning potential for biological and organic chemistry with semi-empirical cost and DFT accuracy. *J. Chem. Phys.* **155**, 204103 (2021).
44. Riniker, S. & Landrum, G. A. Better informed distance geometry: Using what we know to improve conformation generation. *J. Chem. Inf. Model.* **55**, 2562–2574 (2015).
45. Tosco, P., Stiefl, N. & Landrum, G. Bringing the MMFF force field to the RDKit: Implementation and validation. *J. Cheminformatics* **6**, 37 (2014).
46. Lloyd, S. Least squares quantization in PCM. *IEEE Transactions on Information Theory* **28**, 129–137 (1982).
47. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
48. Isert, C., Atz, K., Jiménez-Luna, J. & Schneider, G. QMugs: Quantum Mechanical Properties of Drug-like Molecules., *ETH Zurich*, <https://doi.org/10.3929/ethz-b-000482129> (2021).
49. Dalby, A. *et al.* Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inform. Comput. Sci.* **32**, 244–255 (1992).
50. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).
51. NIST Standard Reference Database 101. Computational Chemistry Comparison and Benchmark DataBase, Release 21. <https://cccbdb.nist.gov/expbondlengths1.asp> August 2020.
52. Bach, R. D. Ring strain energy in the cyclooctyl system. the effect of strain energy on [3 + 2] cycloaddition reactions with azides. *J. Am. Chem. Soc.* **131**, 5233–5243 (2009).
53. Goulet-Hanssens, A. *et al.* Electrocatalytic Z/E isomerization of azobenzenes. *J. Am. Chem. Soc.* **139**, 335–341 (2017).
54. Roca-Lopez, D., Tejero, T. & Merino, P. DFT investigation of the mechanism of E/Z isomerization of nitrones. *J. Org. Chem.* **79**, 8358–8365 (2014).
55. Berthold, M. R. *et al.* KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)* (Springer, 2007).
56. Schrödinger, L. L. C. The PyMOL Molecular Graphics System, Version 2.3.5.
57. Nakata, M., Maeda, T., Shimazaki, T., Hashimoto, M. The PubChemQC Project. <http://pubchemqc.riken.jp/> Accessed Sept. 2020.

## Acknowledgements

We thank Dr. Jan A. Hiss, Dr. Petra Schneider, and Dr. Agnieszka Ilnicka for their valuable insights and discussions. Special thanks to Andreas la Roi from the ETH Research Collection team and to the ETH cluster team for their technical support throughout this project. This work was financially supported by the ETH RETHINK initiative, the Swiss National Science Foundation (grant no. 205321\_182176), and Boehringer Ingelheim Pharma GmbH & Co. KG. C.I. acknowledges support from the Scholarship Fund of the Swiss Chemical Industry.

### Author contributions

Clemens Isert: Methodology, Formal Analyses, Writing. Kenneth Atz: Methodology, Formal Analyses, Writing. José Jiménez-Luna: Conceptualization, Methodology, Writing. Gisbert Schneider: Supervision, Writing.

### Competing interests

G.S. is a cofounder of inSili.com LLC, Zurich, and a consultant to the pharmaceutical industry.

### Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41597-022-01390-7>.

**Correspondence** and requests for materials should be addressed to J.J.-L. or G.S.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

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



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022