



OPEN

DATA DESCRIPTOR

# MultiXC-QM9: Large dataset of molecular and reaction energies from multi-level quantum chemical methods

Surajit Nandi, Tejs Vegge &amp; Arghya Bhowmik

Well curated extensive datasets have helped spur intense molecular machine learning (ML) method development activities over the last few years, encouraging nonchemists to be part of the effort as well. QM9 dataset is one of the benchmark databases for small molecules with molecular energies based on B3LYP functional. G4MP2 based energies of these molecules were published later. To enable a wide variety of ML tasks like transfer learning, delta learning, multitask learning, etc. with QM9 molecules, in this article, we introduce a new dataset with QM9 molecule energies estimated with 76 different DFT functionals and three different basis sets (228 energy numbers for each molecule). We additionally enumerated all possible  $A \leftrightarrow B$  monomolecular interconversions within the QM9 dataset and provided the reaction energies based on these 76 functionals, and basis sets. Lastly, we also provide the bond changes for all the 162 million reactions with the dataset to enable structure- and bond-based reaction energy prediction tools based on ML.

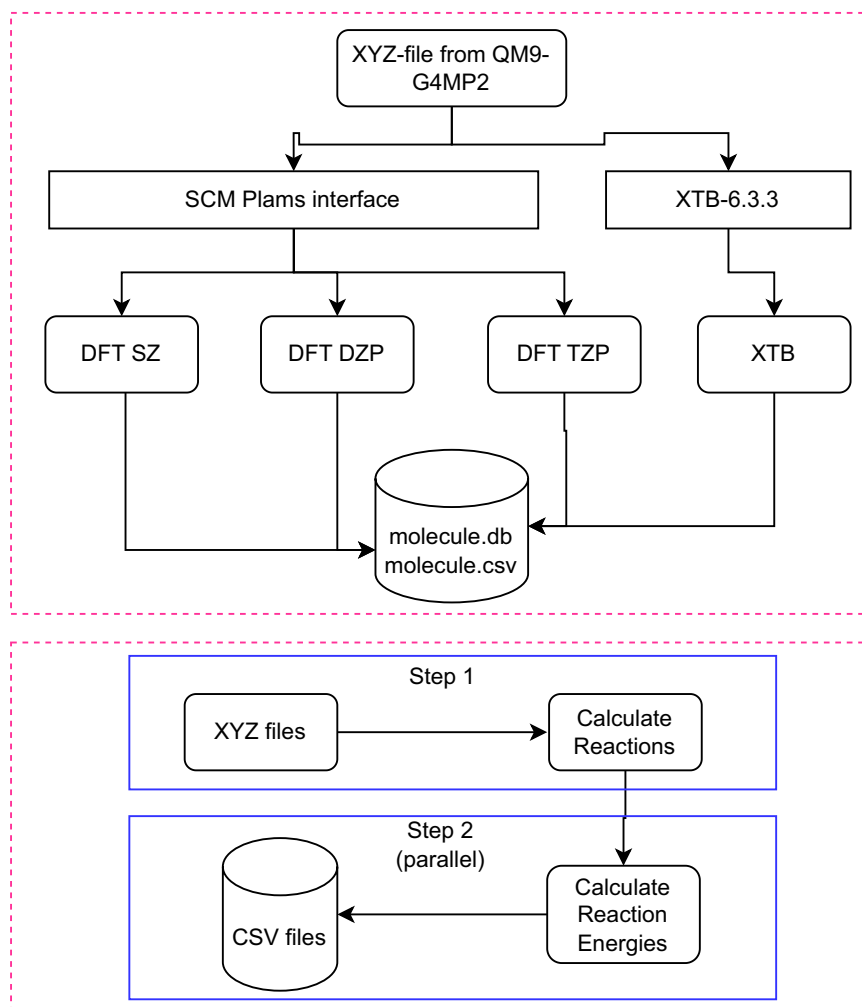
## Background & Summary

The application of machine learning (ML) to predict molecular properties is well established now<sup>1</sup>. The effectiveness and proliferation of molecular ML models rest on large high-quality data sets. Datasets that are easy to access and machine actionable<sup>2–4</sup> have enabled the broader community of machine learning researchers to participate in the building of novel molecular ML models<sup>5–7</sup>. The construction of small molecule properties databases that can be used to prototype and benchmark new ML architectures has been the basis of many ground-breaking machine learning developments in the context of predicting molecular properties<sup>8–10</sup>. Among many such datasets, QM9 dataset<sup>11</sup> has been the gold standard for testing the latest models. In the original QM9 dataset the molecular electronic energies were reported with calculations based on density functional theory at the B3LYP/6–31 G (2df, p)<sup>12–15</sup> level of theory. Recently, those were estimated with high accuracy composite quantum chemistry method (G4MP2)<sup>16–18</sup>. The availability of molecular energies at multiple levels of theory allows interesting ML tasks to be explored, such as the delta learning approach<sup>19</sup>, transfer learning<sup>20</sup>, and multi-task learning<sup>21</sup>.

Beyond the molecule property prediction task, ML can have a tremendous impact on chemical sciences by accelerating the prediction of reaction networks<sup>22</sup>. Building a reaction network requires very fast energy prediction as a large number of energy computation is necessary and thus DFT can be computationally too expensive for exploring large reaction networks. The use of low-level DFT method can reduce computational cost, but the low fidelity of predicted energies could lead to an erroneous analysis of the reaction network. Typically, semiempirical level theories, such as XTb are generally used, but those also have large errors. One way to mitigate this problem is to use energy-correcting methods to achieve higher accuracy via correcting energies from lower-level methods. Building ML-based delta correction models<sup>23</sup> or simple statistical correction schemes<sup>24</sup> that are widely applicable requires large datasets or reaction energies from both cheap/low-fidelity and expensive/high-fidelity methods.

With an aim of providing large datasets with energy targets from a wide variety of methods to help build both these novel molecular ML tasks and ML for reaction networks, we provide multilevel energies of the QM9

Department of Energy Conversion and Storage, Technical University of Denmark, Anker Engelunds Vej 301, 2800 Kongens Lyngby, Copenhagen, Denmark. ✉e-mail: [arbh@dtu.dk](mailto:arbh@dtu.dk)



**Fig. 1** Semi-automated workflow diagram of the database preparation.

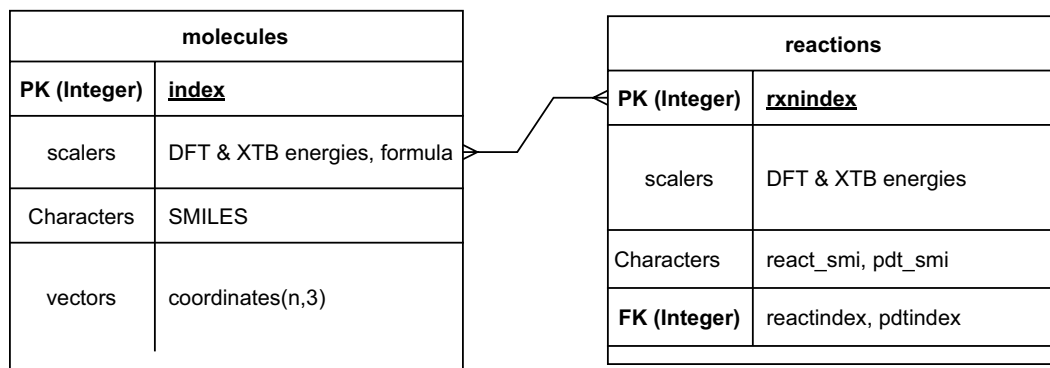
molecules and reactions derived from there. Molecular quantum chemistry community utilizes a wide variety of exchange correlation functional<sup>25</sup> and thus any new ML method needs to be tested for predictions across many different XC to prove generalization and universality. Therefore, providing data for QM9 molecules with a number of DFT functionals and basis sets will provide new challenges to the ML community to build robust ML methods that can be applied to a variety of methods. Furthermore, we provide the energies with the GFN2-XTB method for the QM9 molecules and the reactions, keeping in mind that reaction network exploration methods often use this semi-empirical method.

## Methods

**Dataset generation.** A semi-automated flowchart of data set generation is shown in Fig. 1. The entire process can be divided into approximately three steps. First, energy calculations were performed on the molecules using ADF(SCM) and XTB. Then, the energies were saved in the CSV and SQLite format using Python scripts. Atomistic simulation environment (ASE) was used to process the data and for energy calculation.

In the second step, all possible isodesmic reactions among the QM9 molecules were calculated. Two molecules are considered convertible if their chemical formula is the same. This means that the reactions in the dataset are of type  $A \leftrightarrow B$ . Once all the indices for the reactions are collected and saved in a csv file (“reactions.csv”), we computed the reaction energies and saved the results in multiple CSV files. Therefore, the creation of the datasets requires two steps: one, to compute the indices of the reactants and products, and then, to compute the reaction energies.

The xyz geometries of the QM9 molecules were extracted from the logfiles available in the Figshare repository<sup>26</sup>. We used those geometries to calculate the energies using the post-SCF method as implemented in the SCM software package. For all the post-SCF functionals, the GGA level energy was computed using the PBE method. We run three separate calculations with the SZ, DZP, and TZP basis set for the 133 K molecules. We also computed the single-point energy of the same geometries using the GFN2-XTB semi-empirical method as implemented in the xtb package. The SCF convergence criteria for the DFT calculation was the default one of



**Fig. 2** Entity relationship diagram of the molecule and reaction databases. The “index” and “rxnindex” represent the primary keys of the molecules and reactions data respectively. The “reactindex” and the “pdtindex” represents the foreign key and thus indicates the index of the molecules database.

ADF (1e-6 Hartree), and for the XTB calculation, the SCF convergence criteria were set to the default value of 1e-6 Hartree.

### Data Records

There are two types of data in the dataset - one contains information on molecules and the other contains information on reactions. The information of the molecules is stored in CSV and SQLite formats, while the information on the reactions is provided only in CSV format. The molecules are same as QM9 except that all the molecules with charge separation are excluded from the dataset. The energies are calculated as single point energies of the B3LYP/6–31 G(2df,p) optimized geometries from a previous dataset of the same molecules<sup>16</sup>. The geometries taken for the single point energy calculations were originally optimized at the B3LYP/6–31 G(2df,p) level<sup>11</sup>. It should be noted here that since the theoretical level at which the minima were computed is different from the one used in this article, the energies reported here may not correspond to the minimum energy geometries at the corresponding DFT level.

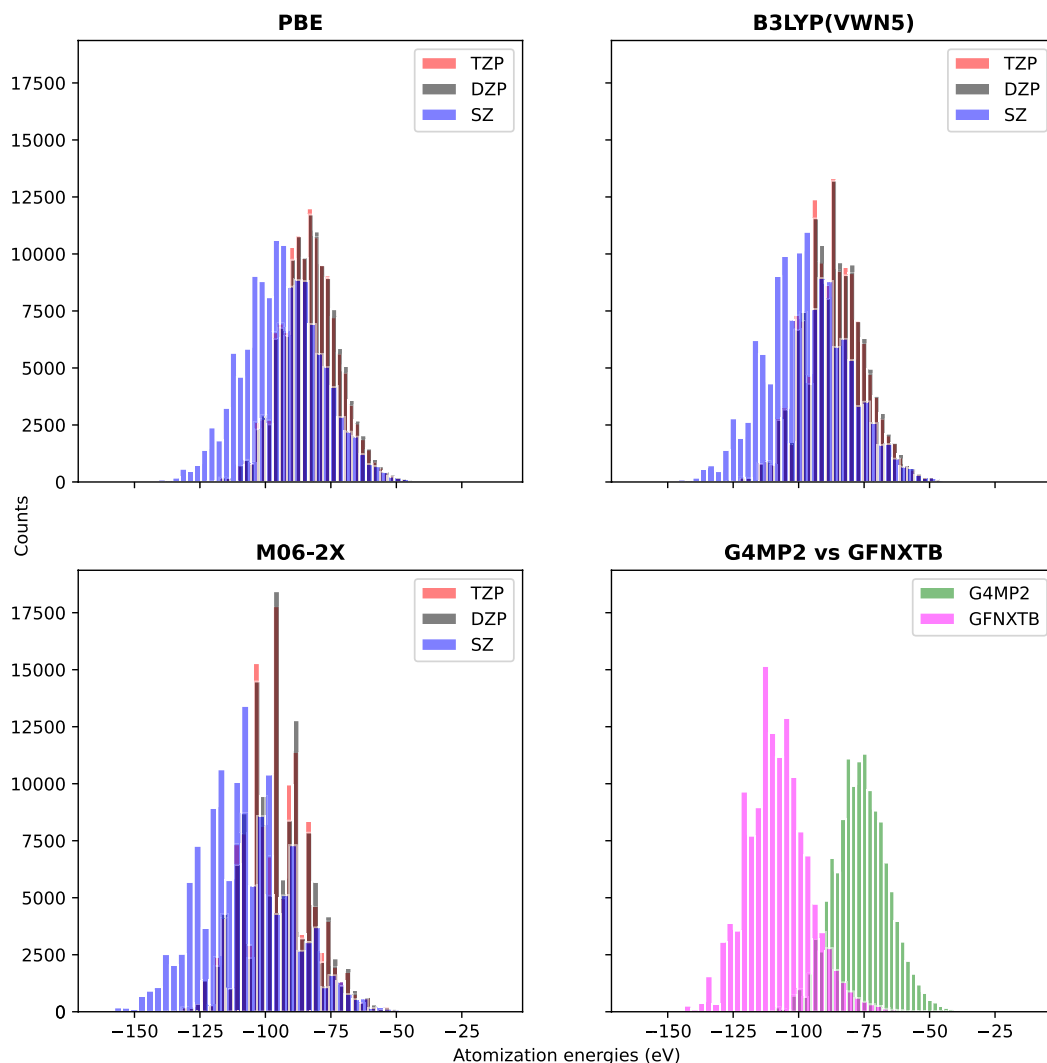
The CSV format database contains energies from different DFT and semi-empirical methods, SMILES strings derived from the xyz files, and chemical formula of each of the species. Each molecule has a unique index number named “index”. The SQLite format molecular dataset contains xyz coordinates, energies, and other relevant properties which were derived automatically by ASE. The reaction database contains reaction energies, indices of the reactants and products, and SMILES of reactants and products. The reactant and product indices are the primary key (“index”) of the molecular database. Each reaction is identified with a key called “rxnindex”. The relationship between the molecular and reactions database is shown in Fig. 2. Each of the reactions are unique and even though the molecule to reaction cardinality is shown as many to many because, each reaction is related to two molecules (more than one).

The log files obtained from the energy calculations are publicly available from DTU Data<sup>27</sup> (version 3 was accessed). All scripts to reproduce the reaction energies, databases, are available from GitHub under the MIT public license (<https://github.com/chemsurajit/largeDFTdata>). The log files are provided into four separate zip files; TZP.zip, DZP.zip, SZ.zip, XTB.zip. It should be noted that in the dataset by Kim *et al.*<sup>16</sup>, the author used the spin multiplicity of N as 2 when calculating the energy of a single atom. Since the spin multiplicity of the ground state of N is 4, we recalculated the atomization energies with the G4MP2 method after replacing the N.log file from the dataset with our own N.log file. We provide the N.log file for the calculation using Gaussian. The N.log file with the energy calculation using G4MP2 is provided in the dataset.

### Technical Validation

We presented the atomization energy distribution in two different ways. Due to the fundamental difference in calculating the atomization energies, we made a comparison first, among the basis sets of three different density functionals, second, between the G4MP2 and the GFN2-xTB methods. The atomization energy distribution plots are given in Fig. 3. In case of the DFTs, the atomization energies that correspond to the TZP and DZP basis sets are quite similar and the atomization energies from the SZ basis are relatively different. This is expected as the SZ basis is the minimalistic representation of the orbitals. It is also evident that the characteristic of the energy distribution varies among the functionals. Between the G4MP2 and GFN2-xTB energies, there are significant differences in the energy distribution, as expected.

Next, we show the distribution of the error in the reaction energies with respect to the G4MP2 reaction energies of different methods in Fig. 4. It is interesting, albeit expected, that the errors of the DFT functionals with the TZP and the DZP basis sets are similar and significantly small. The distribution of error in XTB reaction energies is similar to that of the DFT level with SZ basis set. Also, the error in energy distribution is mostly symmetric and of Gaussian type. This indicates that in any situation even though there can be large errors in atomization energies, the reaction energies mostly follow a normal distribution with the majority of reactions on a relatively small error scale. This type of behavior is known and expected since for a reaction cancellation of errors occurs<sup>28</sup>. Therefore, the errors in reaction energies become lower and sink to a scale that is much smaller



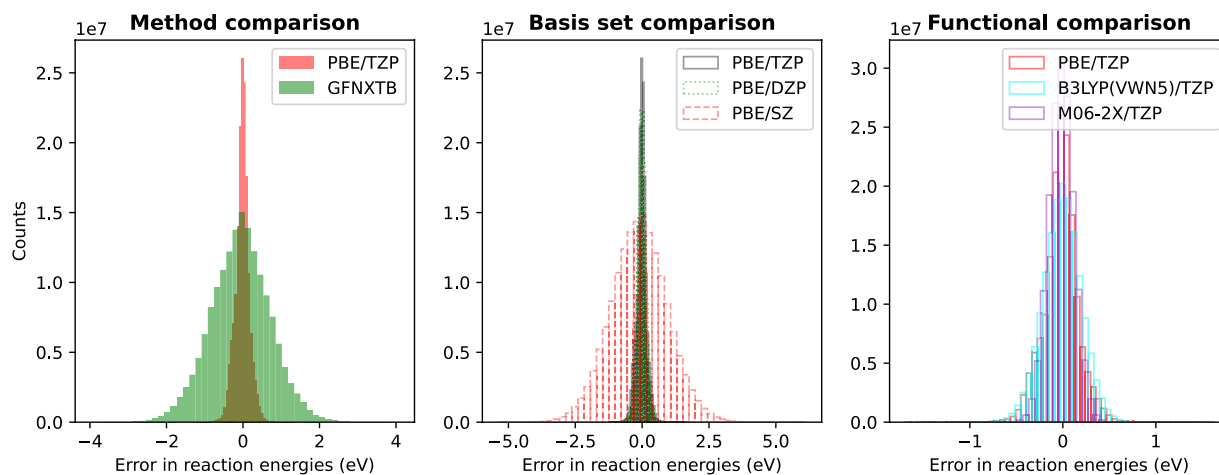
**Fig. 3** Atomization energy distribution of three density functionals: PBE (GGA type), B3LYP(VWN5) (hybrid functional) and M06-2X (highly parametrized meta-hybrid functional) in three different basis sets (SZ, DZP, and TZP). The last plot shows the atomization energy distribution difference between GFN2-xTB and G4MP2 method.

than the atomization energies. It has to be noted here that the energies calculated by using the post-SCF method with the electron density calculated at the PBE level. Therefore, the accuracy of the energy values is expected to be less than the SCF calculations from each of the theoretical levels.

### Usage Notes

All the data are publicly available from DTU data<sup>27</sup> (version 3 was accessed). Log files related to energy calculations are available from the data repository. Scripts for downloading and making the molecular and reaction database are available from the GitHub repository under the MIT license condition. The functional names are used as they are (even if they contain special characters) for the CSV file. For SQLite format, special characters in functional names are replaced by underscore(\_). The changed functional names and the original name of the SCM output file are provided in Table S1 in the supporting information. The CSV file contains SMILES strings, index of the molecules, and the energy values. The SQLite format database contains the molecular coordinates, smile strings, and energy values. Since the SQLite format database is created with ASE, some other information of the molecules (e.g., chemical formula) is also available from this file.

The data on the reaction energies are provided as multiple CSV files. The CSV files contain reaction energies, reactant index (reactindex), product index (pdtindex), and a unique index (rxnindex) for each reaction. The index of the species is the same as the index in the database for molecules. Hence, if a model needs reactant and product coordinates, it can be read from the SQLite format database file of the molecule. Additionally, a README.md file is available in the GitHub repository explaining how to run the scripts.



**Fig. 4** Distribution of the error in reaction energies (eV) with respect to the G4MP2 methods for different functionals.

### Code availability

The energy calculations with the 76 different post-SCF functionals were performed using the SCM software package<sup>29</sup>. The GFN2-xTB energies were computed using the xTB version 6.3.3 software package<sup>30–32</sup>. The G4MP2 energies were obtained from a previous paper by Kim *et al.*<sup>16,26</sup>. The workflow of the calculations and collection of data are build using the Python3.7.10 and BASH scripts. The atomistic simulation environment (ASE)<sup>33</sup> was used to create the database file in SQLite3 format. The csv files were created using pandas. The plots were generated using the matplotlib library. All scripts are available on GitHub under the MIT license agreement (<https://github.com/chemsurajit/largeDFTdata>).

Received: 27 June 2023; Accepted: 25 October 2023;

Published online: 08 November 2023

### References

- Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
- Iserl, C., Atz, K., Jiménez-Luna, J. & Schneider, G. Qmugs, quantum mechanical properties of drug-like molecules. *Sci. Data* **9**, 273 (2022).
- Schreiner, M., Bhowmik, A., Vegge, T., Busk, J. & Winther, O. Transition1x-a dataset for building generalizable reactive machine learning potentials. *Sci. Data* **9**, 779 (2022).
- Liang, J., Xu, Y., Liu, R. & Zhu, X. Qm-sym, a symmetrized quantum chemistry database of 135 kilo molecules. *Sci. Data* **6**, 213 (2019).
- Zubatyuk, R., Smith, J. S., Leszczynski, J. & Isayev, O. Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network. *Sci. Adv.* **5**, eaav6490 (2019).
- Kocer, E., Ko, T. W. & Behler, J. Neural network potentials: A concise overview of methods. *Annu. Rev. Phys. Chem.* **73**, 163–186 (2022).
- Schreiner, M., Bhowmik, A., Vegge, T., Jørgensen, P. B. & Winther, O. Neuralneb—neural networks can find reaction paths fast. *Mach. Learn.: Sci. Technol.* **3**, 045022 (2022).
- Huang, B. & von Lilienfeld, O. A. Ab initio machine learning in chemical compound space. *Chem. Rev.* **121**, 10001–10036 (2021).
- Keith, J. A. *et al.* Combining machine learning and computational chemistry for predictive insights into chemical systems. *Chem. Rev.* **121**, 9816–9872 (2021).
- Musil, F. *et al.* Physics-inspired structural representations for molecules and materials. *Chem. Rev.* **121**, 9759–9815 (2021).
- Ramakrishnan, R., Dral, P. O., Rupp, M. & Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **1**, 1–7 (2014).
- Becke, A. D. Density-functional thermochemistry. iii. the role of exact exchange. *J. Chem. Phys.* **98**, 5648–5652 (1993).
- Lee, C., Yang, W. & Parr, R. G. Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **37**, 785–789 (1988).
- Vosko, S. H., Wilk, L. & Nusair, M. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Can. J. Phys.* **58**, 1200–1211 (1980).
- Stephens, P. J., Devlin, F. J., Chabalowski, C. F. & Frisch, M. J. Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *J. Phys. Chem.* **98**, 11623–11627 (1994).
- Kim, H., Park, J. Y. & Choi, S. Energy refinement and analysis of structures in the qm9 database via a highly accurate quantum chemical method. *Sci. Data* **6**, 109 (2019).
- Curtiss, L. A., Redfern, P. C. & Raghavachari, K. Gaussian-4 theory using reduced order perturbation theory. *J. Chem. Phys.* **127**, 124105 (2007).
- Curtiss, L. A., Redfern, P. C. & Raghavachari, K. Gaussian-4 theory. *J. Chem. Phys.* **126**, 084108 (2007).
- Ramakrishnan, R., Dral, P. O., Rupp, M. & Von Lilienfeld, O. A. Big data meets quantum chemistry approximations: the  $\delta$ -machine learning approach. *J. Chem. Theory Comput.* **11**, 2087–2096 (2015).
- Smith, J. S. *et al.* Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat. Commun.* **10**, 2903 (2019).
- Liu, Z. *et al.* Transferable multilevel attention neural network for accurate prediction of quantum chemistry properties via multitask learning. *J. Chem. Inf. Model.* **61**, 1066–1082 (2021).

22. Kim, Y., Kim, J. W., Kim, Z. & Kim, W. Y. Efficient prediction of reaction paths through molecular graph and reaction network analysis. *Chem. Sci.* **9**, 825–835 (2018).
23. Zhu, J. *et al.* Artificial neural network correction for density-functional tight-binding molecular dynamics simulations. *MRS Commun.* **9**, 867–873 (2019).
24. Nandi, S., Busk, J., Jørgensen, P. B., Vegge, T. & Bhowmik, A. Cheap turns superior: A linear regression-based correction method to reaction energy from the dft. *J. Chem. Inf. Model.* **62**, 4727–4735 (2022).
25. Teale, A. M. *et al.* Dft exchange: Sharing perspectives on the workhorse of quantum chemistry and materials science. *Phys. Chem. Chem. Phys.* (2022).
26. Kim, H., Park, JY. & Choi, S. Highly accurate G4(MP2) benchmark on qm9 database: Energy refinement and analysis of structures, *figshare*, <https://doi.org/10.6084/m9.figshare.c.4351631.v1> (2019).
27. Nandi, S., Vegge, T. & Bhowmik, A. MultiXC-QM9. *DTU Data* <https://doi.org/10.11583/DTU.c.6185986.v3> (2023).
28. Medford, A. J. *et al.* Assessing the reliability of calculated catalytic ammonia synthesis rates. *Science* **345**, 197–200 (2014).
29. te Velde, G. *et al.* Chemistry with adf. *J. Comput. Chem.* **22**, 931–967 (2001).
30. 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).
31. 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).
32. Pracht, P., Caldeweyher, E., Ehlert, S. & Grimme, S. A robust non-self-consistent tight-binding quantum chemistry method for large molecules. *ChemRxiv* (2019).
33. Larsen, A. H. *et al.* The atomic simulation environment—a python library for working with atoms. *J. Phys. Condens. Matter* **29**, 273002 (2017).

## Acknowledgements

The authors acknowledge support from the European Union's Horizon 2020 research and innovation initiative under grants agreement No 957189 (BIG-MAP) and No 957213 (BATTERY 2030+) and Novo Nordisk Foundation (SURE, NNF19OC0057822).

## Author contributions

A.B. conceived the dataset plan, S.N. conducted the simulations, curated the data and created the accompanying codebase, wrote the first draft of the manuscript. All authors analysed the results and reviewed the 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/s41597-023-02690-2>.

**Correspondence** and requests for materials should be addressed to A.B.

**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 licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence 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 licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023