scientific data

OPEN

DATA DESCRIPTOR

Transition1x - a dataset for building generalizable reactive machine learning potentials

Mathias Schreiner[1] ✉, Arghya Bhowmik[2], Tejs Vegge[2], Jonas Busk[2] & Ole Winther[1,3,4]

Machine Learning (ML) models have, in contrast to their usefulness in molecular dynamics studies, had limited success as surrogate potentials for reaction barrier search. This is primarily because available datasets for training ML models on small molecular systems almost exclusively contain configurations at or near equilibrium. In this work, we present the dataset Transition1x containing 9.6 million Density Functional Theory (DFT) calculations of forces and energies of molecular configurations on and around reaction pathways at the ωB97x/6–31 G(d) level of theory. The data was generated by running Nudged Elastic Band (NEB) with DFT on 10k organic reactions of various types while saving intermediate calculations. We train equivariant graph message-passing neural network models on Transition1x and cross-validate on the popular ANI1x and QM9 datasets. We show that ML models cannot learn features in transition state regions solely by training on hitherto popular benchmark datasets. Transition1x is a new challenging benchmark that will provide an important step towards developing next-generation ML force fields that also work far away from equilibrium configurations and reactive systems.

Background & Summary

ML models for molecular systems have accuracy comparable to quantum mechanical (QM) methods but the computational cost of classical interatomic potentials[1–5]. The development of such data-driven models has ushered in a new age in computational chemistry over the last few years[6–9]. ML potentials have been used for a variety of tasks such as structural optimization[10] or the study of finite-temperature dynamical properties through molecular dynamics[11]. ML potentials are especially suited for screening through large numbers of molecules or simulating systems that are too large for traditional QM methods due to a complexity scaling that is orders of magnitudes lower. The applicability of these models depends on the sampling of training data of chemical and structural space[12]. Fitting ML models to the entire potential energy surface (PES) requires lots of carefully selected data as the underlying electronic interaction between atoms is of a complex, quantum mechanical nature. Thus the focus remains on an efficient sampling strategy of the useful parts of the PES that are relevant to the application at hand. For example, models for optimization tasks should be trained on datasets including small perturbations to equilibrium geometries, and models for molecular dynamics (MD) simulations and reactive systems should be trained on datasets with high energy geometries and states that represent the making and breaking of bonds.

ML potentials that allow accurate modeling of general reaction barriers are challenging to train and only limited demonstrations have been shown to date. Acceptable accuracy has been achieved by focusing on single or few types of reactions involving small molecules with tractable dataset size[13–16] or by studying simple molecular dissociation[17]. ML models that can accurately predict PESs for unseen chemical reactions must be incredibly expressive and have access to training data that extensively samples structures from reactive and high-energy regions (compared to near-equilibrium geometries) of chemical space. Recently, the development of Neural Network (NN) architectures that learn representation and energy/force mapping[6] has tackled the problem of expressive models, but creating datasets with millions of data points sampled around reactions of various types, allowing such models to generalize across a large number of reactions, has remained an open challenge. Thus, ML potentials have not yet proved capable of accurate and general prediction of reaction barriers and transition states.

[1]DTU Compute, Technical University of Denmark (DTU), 2800, Lyngby, Denmark. [2]DTU Energy, Technical University of Denmark, 2800, Lyngby, Denmark. [3]Department of Biology, University of Copenhagen (UCph), 2700, Copenhagen N, Denmark. [4]Genomic Medicine, Copenhagen University Hospital, Rigshospitalet, 2100, Copenhagen Ø, Denmark. ✉e-mail: matschreiner@gmail.com

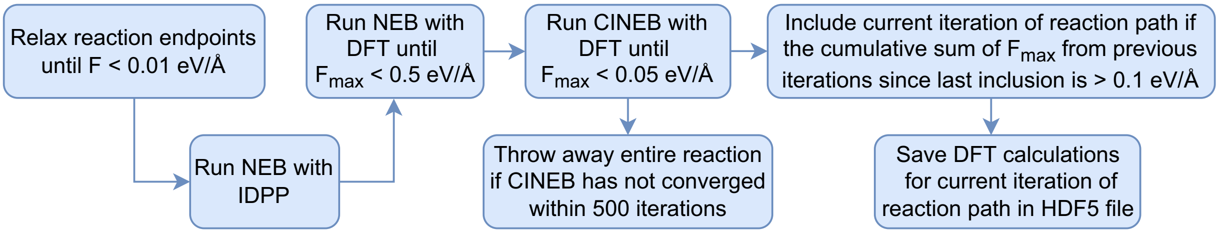
Relax reaction endpoints until F < 0.01 eV/Å
Run NEB with IDPP
Run NEB with DFT until F_max < 0.5 eV/Å
Run CINEB with DFT until F_max < 0.05 eV/Å
Throw away entire reaction if CINEB has not converged within 500 iterations
Include current iteration of reaction path if the cumulative sum of F_max from previous iterations since last inclusion is > 0.1 eV/Å
Save DFT calculations for current iteration of reaction path in HDF5 file

Fig. 1 Overview of the data generation workflow. First, reactant and product are relaxed before generating an initial MEP guess with IDPP[20]. Next NEB[18] and CINEB[21] is run on the initial path until convergence. If the MEP does not converge within 500 iterations we discard the reaction, as unphysical configurations may have been encountered. If the reaction converges, all intermediate paths are saved in the dataset, as long as they are sufficiently different from previously saved paths.

Sampling of rare transition events is efficiently done with the NEB method[18]. Here we propose a new dataset for building ML models capable of generalizing across a large variety of reaction PESs. We base our work on a dataset of reaction-product pairs from Grambow *et al.*[19]. The original dataset contains a wide range of organic reactions representing bond changes between all possible combinations of H, C, N, and O atoms. We leverage NEB-based PES exploration as an efficient data collection tool and prove its superiority compared to MD-based dataset preparation on reactive molecular configurations by testing the accuracy of ML models built from both types of data. Moreover, Transition1x is compatible with ANI1x in the level of DFT such that ML models can be trained on the two datasets in conjunction to leverage both of their strengths.

Ultra-fast prediction of chemical reaction kinetics, especially for computational modeling of complex reaction networks, is groundbreaking for the entire field of chemical and molecular sciences. We believe that the Transition1x dataset will expedite the development and testing of universal reactive ML potentials that help the community achieve that goal.

Methods

Starting from a set of 11961 reactions[19] with reactants, transition states, and products, NEB is used to explore millions of molecular configurations in transition state regions, using DFT to evaluate forces and energies. The resulting DFT calculations are available in the Transition1x dataset. Figure 1 presents an overview of the workflow. Reactant and product are relaxed for any particular reaction before generating an initial path using Image Dependent Pair Potential (IDPP)[20]. Next, the minimal energy path (MEP) is optimized with NEB[18] and consecutively cineb[21] until convergence. If the path converges, we save the DFT calculations from the iterations for which the current reaction path moved significantly.

Initial data. The data generating procedure starts by taking an exhaustive database[19] of product-reactant pairs based on the GDB7 dataset[22]. Each reaction consists of up to seven heavy atoms including C, N, and O. The authors of this data used the Growing String Method (GSM)[23] with the ωB97X-D3[24]/def2-TZVP level of theory to generate reactants, products, and transition states for 11961 reactions using Qchem[25].

Density functional theory. For compatibility with ANI1x[26], the ωB97x[24] and 6–31 G(d)[27] basis set is applied to perform all calculations in ORCA 5.0.2[28].

Optimizer. The BFGS optimizer[29] implemented in Atomic Simulation Environment (ASE)[30] with $\alpha = 70$ and a maximal step size of 0.03 Å is used for all optimization tasks, including relaxing endpoints and running both NEB and CINEB.

Initial path generation. Product and reactant geometries are relaxed in the potential before running NEB. The configuration is considered relaxed once the norm of the forces in configurational space is less than a threshold of 0.01 eVÅ^{-1}. After relaxing the endpoints an initial path is proposed, built from two segments - one interpolated from the reactant to the transition state from the original data, and another interpolated from the transition state to the product. Next, the initial path is minimised with NEB using IDPP[20], a potential specifically designed to generate physically realistic reaction paths for NEB at a low computational cost. Finally, the path is proposed as the initial MEP in the DFT potential.

Nudged elastic band. NEB[18] is a double ended search method for finding MEPs connecting reactant and product states. It works by iteratively improving an initial guess for the MEP by using information about the PES as calculated by some potential. NEB represents the path as a series of configurations called images connected with an artificial spring force. The energy of the path is minimized by iteratively nudging it in the direction of the force perpendicular to it until convergence. After the path has converged, there is no guarantee that the maximal energy image represents the correct transition state as the maximal energy image may not correspond with the true maximum along the path. CINEB[21] is an improvement to the NEB algorithm as it imposes, as an additional condition to the convergence, that the maximal energy image has to lie at a maximum. It does so by, in each iteration, letting the image with the maximal energy climb freely along the reaction path. Running CINEB from the beginning, however, can interfere with the optimizer and result in slow (or wrong) convergence of the

MEP as the climbing image can pull the current path off the target MEP if the paths are not close. Therefore, first, the path is relaxed with regular NEB until the maximal perpendicular force to the MEP is below a threshold of 0.5 eVÅ^{-1}. At this point, NEB has usually found the qualitatively correct energy valley, and further optimization only nudges the path slightly while finding the bottom. At this point, CINEB is turned on to let the highest energy image climb along the path until it finds an energy maximum. CINEB is run until the path has been relaxed fully and the maximal perpendicular force on the path does not exceed 0.05 eVÅ^{-1}. This threshold was chosen as a compromise between having accurate reaction paths in the dataset and limiting redundant DFT calculations. No further refinement of the transition states was done at this point as the goal is to generate a dataset of molecular configurations close to reaction pathways rather than finding accurate transition states. Ten images were used to represent all reaction paths and the spring constant between them was 0.1 eVÅ^{-2}.

Data selection. When running NEB, unphysical configurations are often encountered in reactions that do not converge. Such images in the data will interfere with model when training, and therefore those reactions are discarded entirely. There are 10073 converged reactions in Transition1x. In the final steps of NEB, the molecular geometries of images are similar between each iteration as the images are nudged only slightly close to convergence. Data points should be spread out so that models do not overfit to specific regions of the data. Updated paths are only included in the dataset if they are significantly different from previous ones. The maximal perpendicular force, F_{max}, to the path is a proxy for how much the path moves between iterations. Once the cumulative sum of F_{max} from previous iterations, since the last included path, exceeds 0.1 eVÅ^{-1} the current path is included. This means that often in the first iterations of NEB every path is included, but as we move towards convergence new data points are included at a lower frequency.

Model and training. To validate the dataset, we train and evaluate PaiNN[31] models on Transition1x, QM9x and ANI1x and compare the their performances. PaiNN is an equivariant Message Passing Neural Network (MPNN)[32] model specifically designed to predict properties of molecules and materials. Forces are calculated as the negative gradient of the energy wrt. the Cartesian coordinates of the atoms rather than as a direct output from the model. This ensures consistent forces. We used a cut-off distance of 5 Å to generate the molecular graph neighborhood, three message-passing steps, and 256 neurons in each hidden layer of the model. The model was trained using the ADAM[33] optimizer and an initial learning rate of 10^{-4}. During training, the learning rate was decreased by 20% if no improvement was seen on training data for 10^4 batches. The loss is a combination of a squared error loss on force and energy. The force error is the Euclidian distance between the predicted and the true force vector divided by the number of atoms in the molecule, as otherwise, the force term would contribute more to the loss on bigger molecules. All datasets are stratified by molecular formula such that no two configurations that come from different data splits are constituted of the same atoms. Test and validation data each consist of 5% of the total data and are chosen such that configurations contain all heavy atoms (C, N, O). Potentially, models can learn fundamental features faster from simpler molecules, therefore, all molecules with less than three heavy atom types are kept in the training data. The models are trained on the training data with early stopping on the validation data, and we report the mean and standard deviation of Root Mean Square Error (RMSE) and Mean Average Error (MAE) from the evaluation of test data.

QM9 and QM9x. QM9[34] consists of DFT calculations of various properties for 135k small organic molecules in equilibrium configurations. All molecules in the dataset contain up to 9 heavy atoms, including C, N, O, and F. QM9 is ubiquitous as a benchmark for new QM methods, and to enable direct comparison with Transition1x, all geometries from the QM9 dataset is recalculated with the appropriate level of DFT. Since configurations in the original QM9[34] are not relaxed in our potential, there will be forces on some configurations. All recalculated geometries are saved in a dataset that we shall refer to as QM9x.

ANI1x. ANI1x[26] is a dataset of off-equilibrium molecular configurations generated by perturbing equilibrium configurations using pseudo molecular dynamics. Data is included or rejected from the dataset based on the Query by Comittee (QbC) algorithm. In QbC an ensemble (or committee) of models is trained on the dataset, and the relevance of new proposed data is assessed through the variance of the ensemble's predictions without having to perform expensive calculations on the data. It is assumed that data points will contribute new information to the dataset if the committee disagrees. It is cheaper to evaluate the committee on data than running DFT calculations, so it is possible to screen many candidate configurations before calculating force and energy with more expensive methods. The dataset is generated by alternating between training models and expanding the dataset. The procedure resulted in force and energy calculations for approximately 5 million configurations containing C, O, N, and H.

Data Records

Data records for Transition1x are available in a single Hierarchical Data Format (HDF5)[35] file; Transition1x.h5, hosted by figshare[36]. It can be downloaded at https://doi.org/10.6084/m9.figshare.19614657.v4 or through the repository https://gitlab.com/matschreiner/Transition1x. The HDF5 file structure is as shown in Fig. 2. The parent file has four groups, one group contains all data and the three other groups contain symbolic links to the train, test, and validation data - these are the data splits used in this paper. In each data split group, there is a group for each chemical formula under which there is a subgroup for each reaction with the corresponding atoms. Each reaction has four datasets; the atomic numbers of the reaction, the energies of the configurations, the forces acting on the individual atoms, and the positions of atoms. For a reaction with m atoms where we have saved n images, the *atomic_numbers* dataset will have dimensions (m,), one for each atom. The energy dataset will have dimensions (n,), one energy per configuration. The force and position datasets will have dimensions (n, m, 3) as we need three components of position and force for each of m atoms in n configurations. Under each

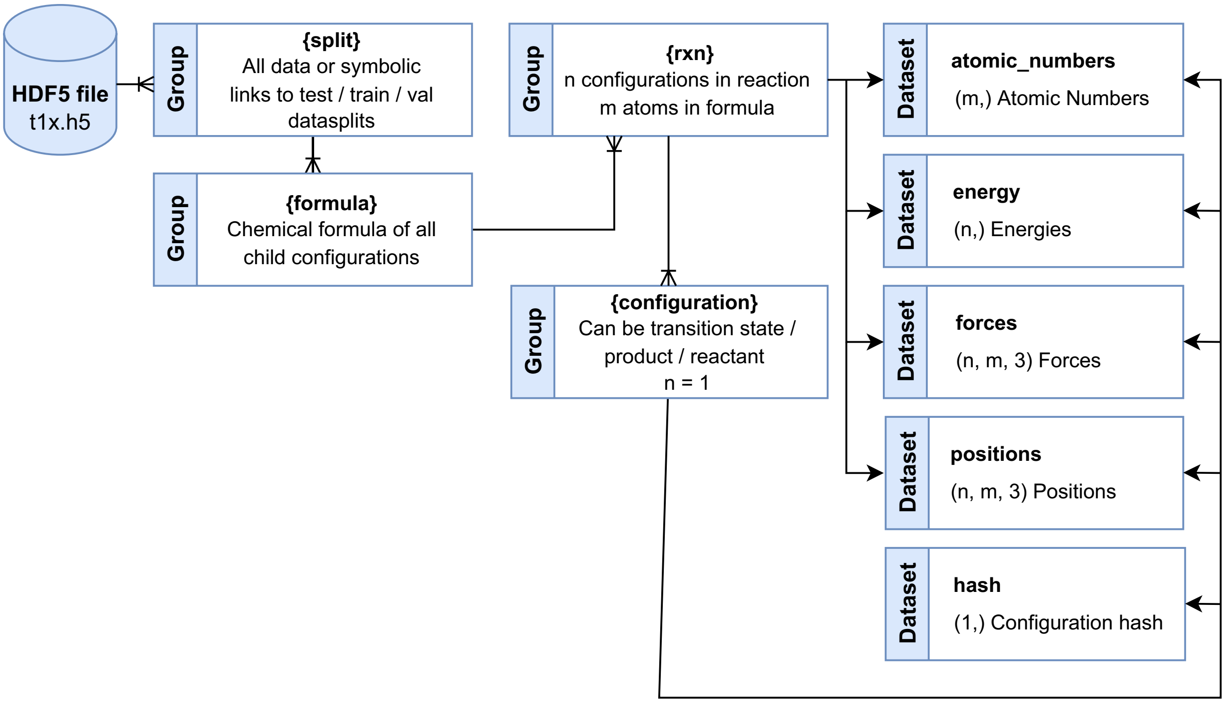

Fig. 2 Structure of Transition1x HDF5 file. Parent groups are data/train/val/test. The data group contains all configurations in the set, and the train/val/test groups contain symlinks to the suggested data splits used in this paper. Each split has a set of chemical formulas unique to that split, and each formula contains all reactions with the given atoms. Finally, energy and force calculations can be accessed from the reaction groups for all intermediate configurations, including transition state, product, and reactant.

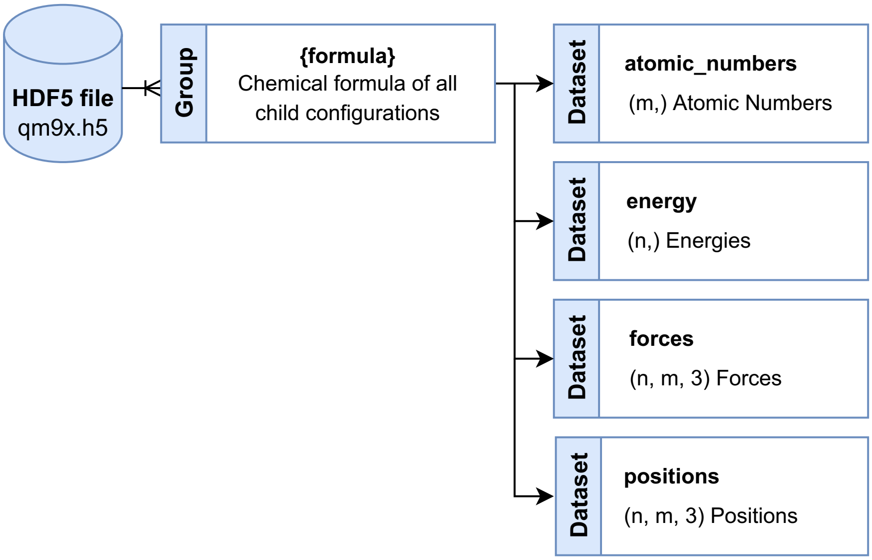

Fig. 3 Structure of QM9x HDF5 file. Energy and force calculations for all configurations in the QM9x dataset consisting of a certain combination of atoms can be accessed as datasets through the formula group.

reaction group, there is a child group for reactant, transition state, and product that follow the same structure as described above with $n = 1$. Products from some reactions are reactants for the next, and they can be linked with a hash value available for each product, transition state, and reactant in the hash dataset. The data has been uploaded to figshare, and there is a git repository with data loaders that can turn the HDF5 file into an ASE database or save the configurations as.xyz files.

Data records for QM9x are also available in a HDF5 file; QM9x.h5, hosted by figshare[37]. It can be downloaded at https://doi.org/10.6084/m9.figshare.20449701.v2 or through the repository https://gitlab.com/matschreiner/QM9x. The HDF5 file structure is shown in Fig. 3. Energy and force calculations for all configurations in the

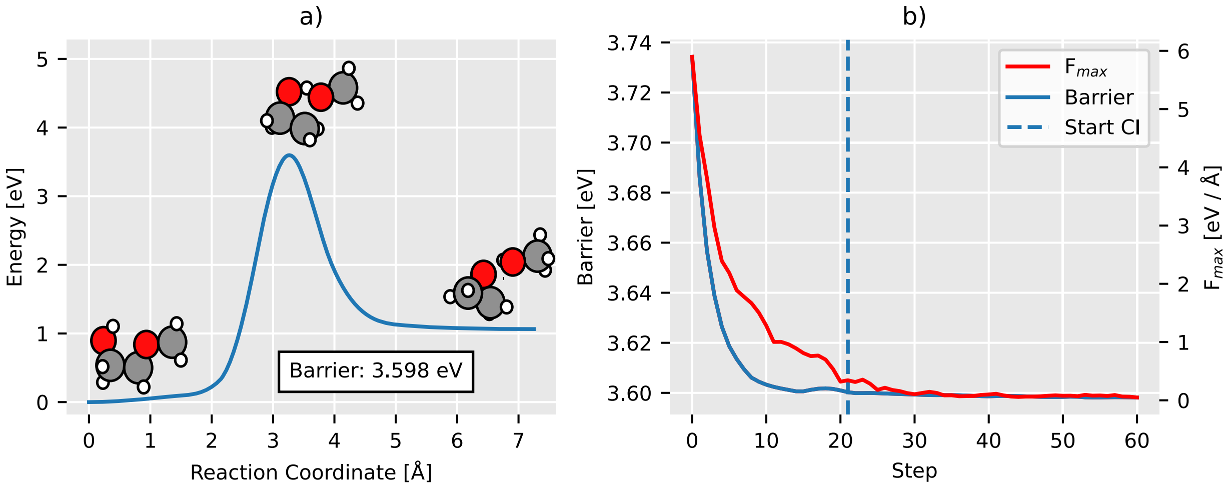

Fig. 4 Plot of NEB convergence on example reaction. Panel (**a**) displays the final MEP with reactant, transition state, and product plotted on top with H, C, and O in white, black, and red, respectively. On the x-axis; the reaction coordinate - distance along the reaction path in configurational space, measured in Å. On the y-axis; the difference in potential energy between reactant and current configuration. Panel (**b**) displays the convergence of NEB. On the x-axis; iterations of NEB. On the y-axis; force in eVÅ^{-1} and energy barrier in eV at the current step. F_{max}, shown in red, is the maximal perpendicular force acting on any geometry along the path, and Barrier, shown in blue, is the height of the energy barrier found at the current step. Moving right in the plot both F_{max} converges towards zero as NEB finds the saddle point, and the Barrier converges towards the final value of 3.6 eV that can be seen in panel a.

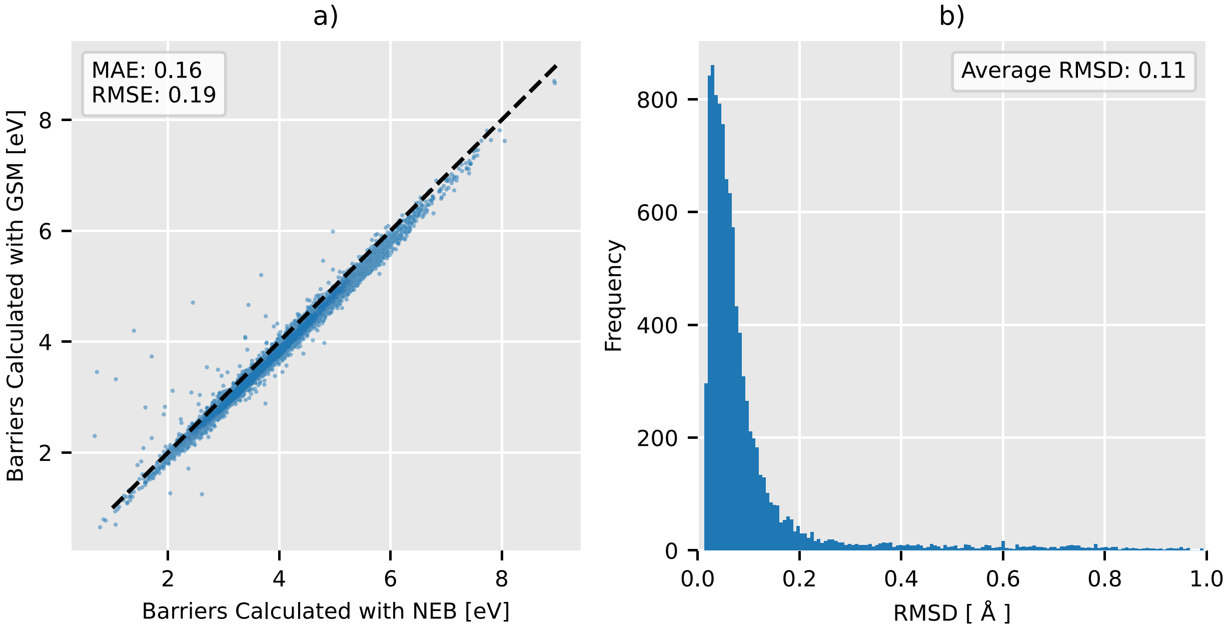

Fig. 5 Comparison of transition states and barriers found in this work with neb and the 6–31 G(d) basis set, and in the original work with gsm and the def2-mSVP basis set. Panel (**a**) displays energies in eV for all transition states calculated using neb on the x-axis and gsm on the y-axis. Panel (**b**) displays a histogram of Root Mean Square Deviation (RMSD) between transition states found.

QM9x dataset consisting of a certain combination of atoms can be accessed as datasets through the formula group. The dimensions and structure of these datasets follow the same logic as described above.

Technical Validation

In Fig. 4 we show the MEP for a reaction involving C3H7O2 and the convergence of NEB for it. Often the barrier grows initially after turning on the climbing image because we start maximizing the energy in a new degree of freedom. NEB converged on 10073 out of 11961 reactions, and 89 percent of the converged reactions did so

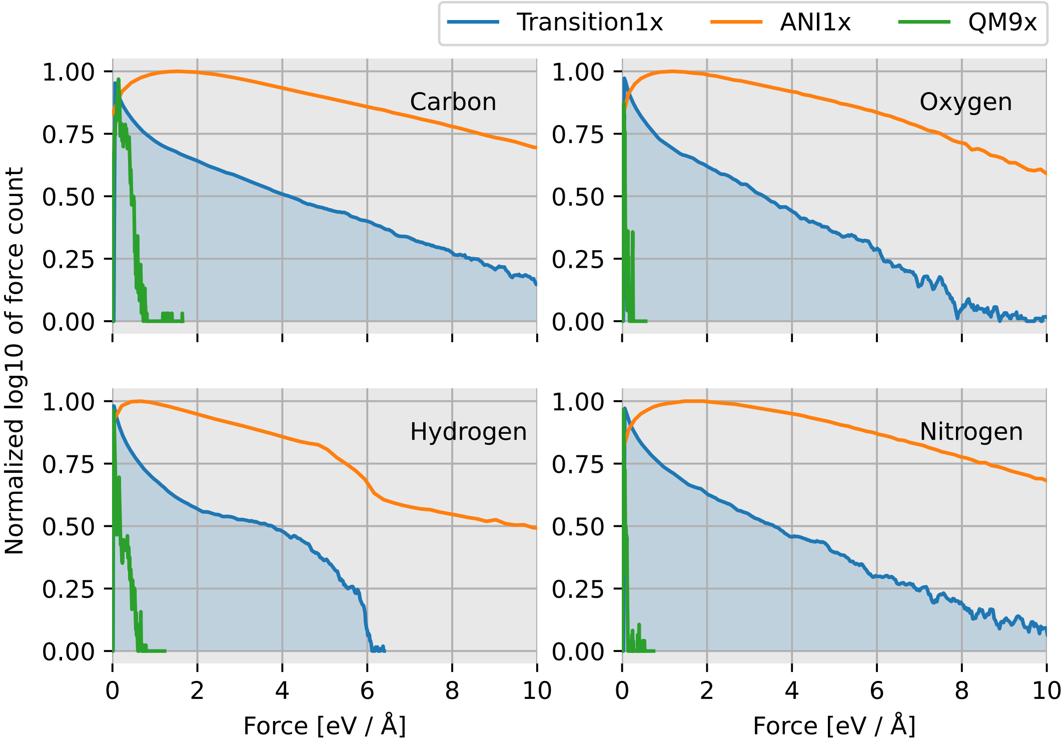

Fig. 6 Distribution of forces acting on atom-types in each dataset. The x-axis is the force measured in eV/Å. The y-axis is the base 10 logarithm of the count of forces in each bin, normalized over the full domain so that all sets can be compared. In blue; Transition1x. In yellow; ANI1x. In green QM9x.

within the first 200 iterations. To ensure the cleanliness of the data, we choose to discard all reactions that do not converge - these reactions often contain unphysical structures that do more harm than good as training data.

The dataset includes a wide range of organic reactions. All reactions contain up to 7 heavy atoms including C, N, and O, and up to six bond changes where bonds are breaking and forming between all combinations of heavy atoms. Detailed analysis of the number of bond changes per reaction, number of bond changes involving specific pairs of atoms, the spread of activation energies, and SMARTS strings describing reactive centers of the reactions, is included in the original work Grambow *et al.*[19].

The perpendicular force drops off rapidly when running NEB and so does the variation in data between iterations as the path is nudged less between iterations towards convergence of the algorithm. F_{max} is used as a proxy for how much the path moves between iterations and once the cumulative F_{max} since the last included path exceeds a threshold of 0.1 eV/Å, the path is included in the dataset.

The transition states found with NEB correspond to the transition states from the original GSM data with a MAE of 0.16 eV, RMSE of 0.19 eV, and an average Root Mean Square Deviation (RMSD) of 0.11 Å. See Fig. 5 in the Appendix for details. Barrier energies match, but the NEB energies tend to be shifted higher. Generally, it is easier to describe electron clouds around relaxed configurations than around transition states where bonds are breaking and other complex interactions take place. Therefore, the more expressive basis set enables us to relax configurations around transition states further than around equilibrium states which results in lower barrier heights. There are more outliers above the $x = y$ line than below it which indicates that GSM was caught in sub-optimal reaction pathways more often than NEB.

Figure 6 displays the distribution of forces on each type of atom in Transition1x compared with ANI1x and QM9. Interestingly, even though geometries in Transition1x are further away from equilibrium than in ANI1x (regions between equilibria are actively sought out in Transition1x), the distribution of forces on ANI1x has flatter, wider tails signifying higher variance in forces. Moreover, Transition1x cusps at zero whereas ANI1x maxima lie further out. Large forces are not necessarily involved when dealing with reactive systems. When reaction pathways are minimized, forces are minimized too in all but one degree of freedom. The transition states contribute as much to the force distribution as equilibrium configurations, as the transition state is a saddle point with no net force on any atoms. ANI1x has no inherent bias towards low forces on geometries as it explores configurational space with pseudo-MD and therefore, even though the configurations are closer to equilibrium we see a higher variance in forces. On the heavy atoms, the tails are qualitatively equal between ANI1x and Transition1x, trailing off exponentially, but it is different for hydrogen. In ANI1x, forces on Hydrogen trail off exponentially as with the other atoms, but for Transition1x there is a sudden drop of the distribution. Hydrogen atoms are often at the outskirts of the molecules and are relatively free to move compared to heavier atoms on the backbone. In the case of the Transition1x data generation procedure, energy and forces are minimized, and therefore Hydrogen atoms do not experience large forces as they have lots of freedom to relax in the geometry. In ANI1x, configurations are generated by perturbing the geometries randomly, and hydrogen atoms might end

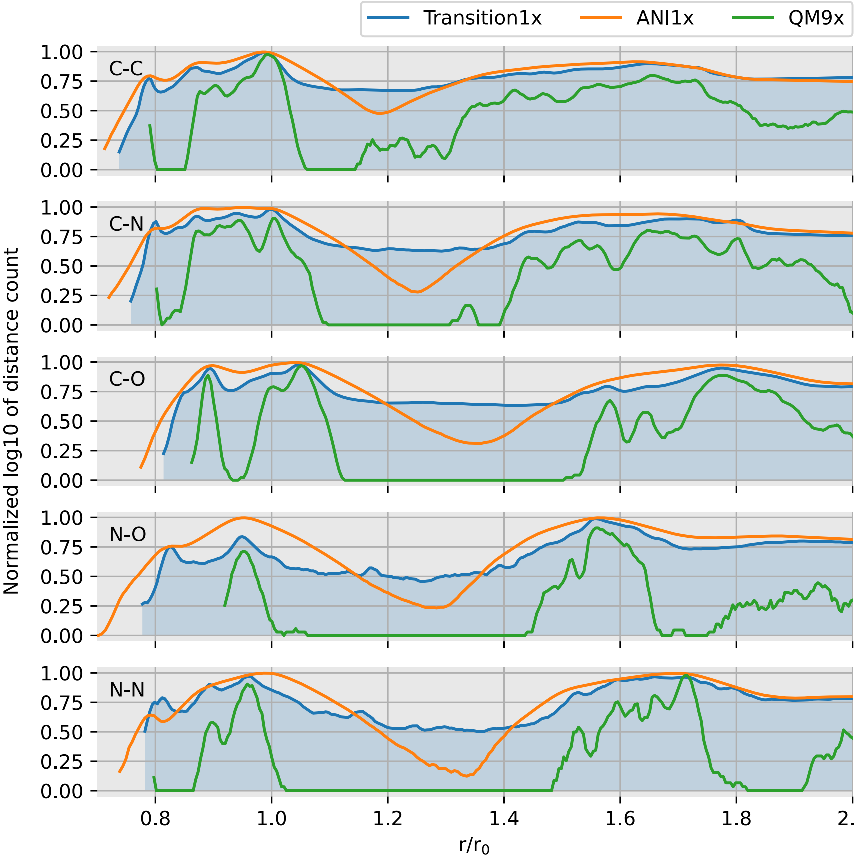

Fig. 7 Distribution of interatomic distances between heavy atoms in each dataset. A configuration with n heavy atoms contributes with $n(n-1)/2$ distances in the count. On the y-axis; the log frequency of interatomic distance, normalized between 0 and 1 for comparison as datasets vary in size. On the x-axis; distance given in units of r_0 where r_0 is the equilibrium bond length for a single bond between the smallest possible stable molecule that can be made with the atoms in question. In blue; Transition1x. In yellow; ANI1x. In green; QM9x, recalculated using our level of theory.

up with unrealistically large forces on them. This might be a general problem with ANI1x and also a reason why ANI1x is not a proper dataset to learn reaction mechanisms.

Even though ANI1x has a wider distribution of forces, the inter-atomic distances between pairs of heavy atoms are less varied than in Transition1x. Figure 7 displays the distribution of distances between pairs of heavy atoms for Transition1x, ANI1x, and QM9x. For QM9x, a dataset of only equilibrium configurations, some inter-atomic distances are not present at all. Distances are measured in units of r_0, the single bond equilibrium distance between the atoms in the smallest possible molecule constructed out of the two. For example, in the case of “C, C” we measure in units of the distance between carbon atoms in ethane. Many of the more extreme inter-atomic distances in Transition1x are difficult to produce by the normal mode sampling technique of ANI1x as many atoms would randomly have to move such that the whole molecule moves along a low-energy valley. However, because NEB samples low-energy valleys by design, we discover likely molecules with inter-atomic distances that are otherwise energetically unfavorable.

We test all resulting models against the test data from each dataset and Transition States from the test reactions. Table 1 displays the results. It is clear from their evaluation of Transition1x and transition states, that models trained on ANI1x do not have sufficient data in transition state regions to properly learn the complex interactions present here. ANI1x has a broad variety of chemical structures, but many of the fundamental interactions found in ANI1x are present in Transition1x, which is why models trained on Transition1x perform better on ANI1x than vice versa. In general, the PES of a set of atoms is an incredibly complex function of quantum mechanical nature. Models trained on QM9x do not perform well on either Transition1x or ANI1x. This is as expected as QM9x contains only equilibrium (or very close to equilibrium in the new potential) structures, so the models trained on it have not seen any of the out-of-equilibrium interactions that are present in the more challenging datasets of ANI1x and Transition1x.

Transition state data is required if we want to replace DFT with cheap ML potentials in algorithms such as NEB[38] or GSM, or train molecular dynamics models to work in transition state regions. NNs are phenomenal function approximators, given sufficient training examples, but they do not extrapolate well. Training examples spanning the whole energy surface are needed to train reliable and general-purpose ML models. Transition1x is

		Energy [eV]		Forces [eV/Å]	
Trained on	**Tested on**	**RMSE**	**MAE**	**RMSE**	**MAE**
ANI1x	Transition States	0.629 (11)	0.495 (10)	0.71(2)	0.53(1)
Transition1x		**0.112 (3)**	**0.075 (1)**	**0.211(1)**	**0.111(1)**
QM9x		3.132 (23)	2.957 (25)	0.71(2)	0.316(5)
ANI1x	ANI1x	**0.044(5)**	**0.023(1)**	**0.062(1)**	**0.039(1)**
Transition1x		0.365(17)	0.226(8)	0.43(3)	0.179(1)
QM9x		3.042(13)	2.313(11)	1.9(1)	1.29(1)
ANI1x	Transition1x	0.628(63)	0.289(13)	0.65(1)	0.20(8)
Transition1x		**0.102(2)**	**0.048(1)**	**0.136(1)**	**0.058(1)**
QMx		2.613(18)	1.421(11)	0.495(3)	0.241(1)
ANI1x	QM9x	0.134(1)	0.124(1)	0.061(1)	0.038(2)
Transition1x		0.111(2)	0.074(3)	0.082(1)	0.048(1)
QM9x		**0.04(2)**	**0.015(1)**	**0.016(0)**	**0.007(0)**

Table 1. Test results of PaiNN models trained on ANI1x, QM9x, and Transition1x. We report RMSE and MAE on energy and forces. Force error is the component-wise error between the predicted and true force vector. The test sets have been constructed such that all configurations contain C, N, O, and H, and such that no formula has been seen previously in the training data. We show the best performing model in bold in each test-setup.

a new type of dataset that explores different regions of chemical space than other popular datasets and it is highly relevant as it expands on the completeness of available data in the literature.

Code availability

There are download scripts and data loaders available in the repositories https://gitlab.com/matschreiner/Transition1x and https://gitlab.com/matschreiner/QM9x. See the repositories and README for examples and explanations of how to use the scripts and datasets.

All electronic structure calculations were computed with the ORCA electronic structure packages, version 5.0.2. All NEB calculations were computed with ASE version 3.22.1. Scripts for calculating, gathering, and filtering data can be found in the Transition1x repository. `scripts/neb.py` takes reactant, product, output directory, and various optional arguments and runs NEB on the reaction while saving all intermediate calculations in an ASE database in the specified output directory. `scripts/combine_dbs.py` takes an output path for the HDF5 file and a JSON list of all output directories produced by running the previous script and combines them in the HDF5 file as described in the paper. See the repository for how to install, specific commands, options, and further documentation.

Received: 23 September 2022; Accepted: 16 November 2022;
Published online: 24 December 2022

References

1. Faber, F. A. *et al*. Prediction errors of molecular machine learning models lower than hybrid dft error. *Journal of Chemical Theory and Computation* **13**, 5255–5264, https://doi.org/10.1021/ACS.JCTC.7B00577/SUPPL_FILE/CT7B00577_SI_001.PDF (2017).
2. Westermayr, J., Gastegger, M., Schütt, K. T. & Maurer, R. J. Perspective on integrating machine learning into computational chemistry and materials science. *The Journal of Chemical Physics* **154**, 230903, https://doi.org/10.1063/5.0047760 (2021).
3. Campbell, S. I., Allan, D. B. & Barbour, A. M. Machine learning for the solution of the schrödinger equation. *Machine Learning: Science and Technology* **1**, 013002, https://doi.org/10.1088/2632-2153/AB7D30 (2020).
4. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical review letters* **98**, 146401 (2007).
5. Westermayr, J. & Marquetand, P. Machine learning for electronically excited states of molecules. *Chemical Reviews* **121**, 9873–9926, https://doi.org/10.1021/ACS.CHEMREV.0C00749 (2021).
6. Unke, O. T. *et al*. Machine learning force fields. *Chemical Reviews* **121**, 10142–10186 (2021).
7. Behler, J. Four generations of high-dimensional neural network potentials. *Chemical Reviews* **121**, 10037–10072 (2021).
8. Huang, B. & von Lilienfeld, O. A. Ab initio machine learning in chemical compound space. *Chemical reviews* **121**, 10001–10036 (2021).
9. Deringer, V. L. *et al*. Gaussian process regression for materials and molecules. *Chemical Reviews* **121**, 10073–10141 (2021).
10. Kaappa, S., Larsen, C. & Jacobsen, K. W. Atomic structure optimization with machine-learning enabled interpolation between chemical elements. *Physical Review Letters* **127**, https://doi.org/10.1103/PhysRevLett.127.166001 (2021).
11. Wang, J., Shin, S. & Lee, S. Interatomic potential model development: Finite-temperature dynamics machine learning. *Advanced Theory and Simulations* **3**, 1900210, https://doi.org/10.1002/ADTS.201900210 (2020).
12. von Lilienfeld, O. A., Müller, K. R. & Tkatchenko, A. Exploring chemical compound space with quantum-based machine learning. *Nature Reviews Chemistry 2020 4:7* **4**, 347–358, https://doi.org/10.1038/s41570-020-0189-9 (2020).
13. Lu, X., Meng, Q., Wang, X., Fu, B. & Zhang, D. H. Rate coefficients of the h + h2o2→h2 + ho2 reaction on an accurate fundamental invariant-neural network potential energy surface. *The Journal of chemical physics* **149**, 174303 (2018).
14. Young, T. A., Johnston-Wood, T., Deringer, V. L. & Duarte, F. A transferable active-learning strategy for reactive molecular force fields. *Chemical science* **12**, 10944–10955 (2021).
15. Manzhos, S. & Carrington, T. Jr. Neural network potential energy surfaces for small molecules and reactions. *Chemical Reviews* **121**, 10187–10217 (2020).
16. von Rudorff, G. F., Heinen, S. N., Bragato, M. & von Lilienfeld, O. A. Thousands of reactants and transition states for competing e2 and s2 reactions. *Machine Learning: Science and Technology* **1**, 045026, https://doi.org/10.1088/2632-2153/ABA822 (2020).

17. Malshe, M. *et al*. Theoretical investigation of the dissociation dynamics of vibrationally excited vinyl bromide on an ab initio potential-energy surface obtained using modified novelty sampling and feedforward neural networks. ii. numerical application of the method. *The Journal of chemical physics* **127**, 134105 (2007).
18. Sheppard, D., Terrell, R. & Henkelman, G. Optimization methods for finding minimum energy paths. *The Journal of Chemical Physics* **128**, 134106, https://doi.org/10.1063/1.2841941 (2008).
19. Grambow, C. A., Pattanaik, L. & Green, W. H. Reactants, products, and transition states of elementary chemical reactions based on quantum chemistry. *Scientific Data* **7**, https://doi.org/10.1038/s41597-020-0460-4 (2020).
20. Smidstrup, S., Pedersen, A., Stokbro, K. & Jónsson, H. Improved initial guess for minimum energy path calculations. *The Journal of Chemical Physics* **140**, 214106, https://doi.org/10.1063/1.4878664 (2014).
21. Henkelman, G., Uberuaga, B. P. & Jónsson, H. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *The Journal of Chemical Physics* **113**, 9901, https://doi.org/10.1063/1.1329672 (2000).
22. 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. *Journal of chemical information and modeling* **52**, 2864–2875 (2012).
23. Zimmerman, P. M. Single-ended transition state finding with the growing string method. *Journal of Computational Chemistry* **36**, 601–611, https://doi.org/10.1002/JCC.23833 (2015).
24. Chai, J. D. & Head-Gordon, M. Systematic optimization of long-range corrected hybrid density functionals. *The Journal of Chemical Physics* **128**, 084106, https://doi.org/10.1063/1.2834918 (2008).
25. Epifanovsky, E. *et al*. Software for the frontiers of quantum chemistry: An overview of developments in the q-chem 5 package. *The Journal of Chemical Physics* **155**, 084801, https://doi.org/10.1063/5.0055522 (2021).
26. Smith, J. S., Nebgen, B., Lubbers, N., Isayev, O. & Roitberg, A. E. Less is more: Sampling chemical space with active learning. *The Journal of Chemical Physics* **148**, 241733, https://doi.org/10.1063/1.5023802 (2018).
27. Ditchfield, R., Hehre, W. J. & Pople, J. A. Self-consistent molecular-orbital methods. ix. an extended gaussian-type basis for molecular-orbital studies of organic molecules. *The Journal of Chemical Physics* **54**, 724, https://doi.org/10.1063/1.1674902 (2003).
28. Neese, F., Wennmohs, F., Becker, U. & Riplinger, C. The orca quantum chemistry program package. *The Journal of Chemical Physics* **152**, 224108, https://doi.org/10.1063/5.0004608 (2020).
29. Broyden, C. G. The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics* **6**, 76–90, https://doi.org/10.1093/IMAMAT/6.1.76 (1970).
30. Larsen, A. H. *et al*. The atomic simulation environment–a python library for working with atoms. *Journal of Physics: Condensed Matter* **29**, 273002, https://doi.org/10.1088/1361-648X/AA680E (2017).
31. Schütt, K. T., Schütt, S., Unke, O. T. & Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. *Proceedings of Machine Learning Research* 9377–9388 (2021).
32. Bacciu, D., Errica, F., Micheli, A. & Podda, M. A gentle introduction to deep learning for graphs. *Neural Networks* **129**, 203–221, https://doi.org/10.1016/j.neunet.2020.06.006 (2019).
33. Kingma, D. P. & Ba, J. L. Adam: A method for stochastic optimization.
34. Ramakrishnan, R., Dral, P. O., Rupp, M. & Lilienfeld, O. A. V. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data 2014 1:1* **1**, 1–7, https://doi.org/10.1038/sdata.2014.22 (2014).
35. The HDF Group. Hierarchical data format version 5 (2000-2010).
36. Schreiner, M. Transition1x. *Figshare.* https://doi.org/10.6084/m9.figshare.19614657.v4 (2022).
37. Schreiner, M. QM9x. *Figshare.* https://doi.org/10.6084/m9.figshare.20449701.v2 (2022).
38. Schreiner, M., Bhowmik, A., Vegge, T., Jørgensen, P. B. & Winther, O. Neuralneb - neural networks can find reaction paths fast. *Machine Learning: Science and Technology* https://doi.org/10.1088/2632-2153/ACA23E (2022).

Acknowledgements

The authors acknowledge support from the Novo Nordisk Foundation (SURE, NNF19OC0057822) and the European Union's Horizon 2020 research and innovation program under Grant Agreement No. 957189 (BIG-MAP) and No. 957213 (BATTERY2030PLUS). Ole Winther also receives support from Novo Nordisk Foundation through the Center for Basic Machine Learning Research in Life Science (NNF20OC0062606) and the Pioneer Centre for AI, DNRF grant number P1.

Author contributions

M.S., A.B., T.V. and O.W. conceived the study. M.S. wrote the code, conducted the experiments, and wrote the majority of the article. A.B. and M.S. wrote background and summary, A.B. and O.W. provided supervision and reviewed the article, and J.B. reviewed the article and provided helpful discussions.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.S.

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.