Machine Learning for Chemical Reactions
A Dance of Datasets and Models

Mathias Schreiner[1] **Arghya Bhowmik**[2] **Tejs Vegge**[2]
Jonas Busk[2] **Peter B. Jørgensen**[2] **Ole Winther**[1,3,4]
[1]DTU Compute, Technical University of Denmark (DTU), 2800 Lyngby, Denmark
[2]DTU Energy, Technical University of Denmark (DTU), 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

Abstract

Machine Learning (ML) models have proved to be excellent emulators of Density Functional Theory (DFT) calculations for predicting features of small molecular systems. The activation energy is a defining feature of a chemical reaction, but, despite the success of ML in computational chemistry, an accurate, fast, and general ML-calculator for Minimal Energy Paths (MEPs) has not yet been proposed. Here, we summarize contributions from two of our recent papers, where we apply Graph Neural Network (GNN) based models, trained on various datasets, as potentials for the Nudged Elastic Band (NEB) algorithm to speed up MEP-search. We show that relevant data from reactive regions of the Potential Energy Surface (PES) in training data is paramount to success. Hitherto popular benchmark datasets primarily contain configurations in, or close to, equilibrium, and are not adequate for the task. We propose a new dataset, Transition1x, that contains force and energy calculations for 10 million molecular configurations from on and around MEPs of 10.000 organic reactions of various types. By training GNNs on Transition1x and applying the models as PES-evaluators for NEB, we achieve a Mean Average Error (MAE) of 0.23 eV on predicted activation energies of unseen reactions, compared to DFT, while running the algorithm 1200 times faster. Transition1x is a challenging dataset containing a new type of data that may serve as a benchmark for future methods for transition-state search.

Introduction

The activation energy of a chemical reaction is the difference between the reactant and transition state energies. It describes the energetic barrier of the reaction, and it dominates the reaction-rate exponentially through the Arrhenius Equation. To build a virtual laboratory, where reaction mechanisms for synthesis of drugs and materials can be studied, it is crucially important to be able to quickly and accurately predict activation energies and Minimal Energy Paths (MEPs) between equilibrium configurations. Nudged Elastic Band (NEB)[1] is an effective algorithm, designed to find MEPs, and activation energies in chemical space. It does so by iteratively nudging an initial guess for the reaction path in the direction of the force perpendicular to the path until convergence. The NEB algorithm requires a subordinate algorithm for calculating gradients and energies on the surrounding Potential Energy Surface (PES), and Density Functional Theory (DFT) is a popular choice for this. However, a single DFT calculation can take from a minute and up to several hours, depending on the size of the molecule and level of theory. Even for small molecular systems of 6-7 heavy atoms, NEB has to evaluate thousands of configurations to converge, making DFT inappropriate for large scale exploration of reaction-networks.

Machine Learning and the Physical Sciences workshop, NeurIPS 2022.

Machine Learning (ML) models, and Graph Neural Networks (GNNs)[2,3] in particular, have proved to be capable surrogate DFT potentials that can accurately evaluate molecular properties fast[4–13]. In this paper we train Polarizable Atom interaction Neural Network (PaiNN)[14] models, a GNN architecture, on various datasets and apply them as surrogate potentials for DFT to alleviate the NEB-bottleneck.

Initially, the models were trained on existing datasets in the literature. However, available datasets of significant volume that are interesting for training Neural Network (NN) models, either contain molecular configurations exclusively in equilibrium, or are generated through Molecular Dynamics (MD). Transitions between equilibriums are rare events, and datasets generated through MD-simulations do not contain sufficient samples of data around reaction paths to enable training models with accurate representations of these regions[15,16]. We created a new dataset, Transition1x, to address this problem. Transition1x leverages NEB to sample relevant configurations on and around reaction pathways for thousands of organic reactions. All intermediate configurations encountered while running the algorithm were calculated using the same level of DFT as the popular ANI1x[17,18] dataset, such that models trained on the two datasets can be compared in a meaningful way.

Method

Nudged Elastic Band

NEB[1] is a method for finding transition-states and MEPs given products and reactants of chemical reactions. It does so by iteratively nudging an initial guess for the MEP directed by the force perpendicular to the path. The path is represented by an array of molecular configurations called images connected by an artificial spring force which ensures that the images stay evenly separated and do not fall into the minimas at the reaction endpoints. Eventually, as the perpendicular force on the path converges to zero, the MEP will relax at the bottom of the local low-energy valley. At this point, NEB returns the maximal-energy configuration along the path as the transition-state. However, the true transition-state of the reaction may lie between two images representing the path. Climbing Image Nudged Elastic Band (CINEB)[19] addresses this problem by, between iterations, choosing a transition-state candidate and further maximize its energy by following the gradient on the PES, parallel to the current path. The CINEB algorithm terminates once both the maximal force perpendicular to the path, and climbing force on the transition-state candidate has converged. At this point, the maximal energy-configuration representing the path corresponds to the transition state.

Datasets

Transition1x was generated by running NEB, applying DFT as potential, on all reactions in the dataset released by Grambow et al. (2020)[20]. The original data contains products, reactants, and transition states for 11.000 organic reactions of various types. To ensure compatibility with ANI1x and QM9x, all intermediate calculations were made using the 6-31G(d)[21] basis set and ωB97x functional[22] . NEB typically converges within 200 iterations given the elements and sizes of molecules in the dataset. Calculations become gradually more similar towards convergence of NEB, as the path is nudged less between iterations. In order to reduce redundancy in data, paths are excluded from the dataset unless the cumulative maximal force, perpendicular to the path, from previous iterations exceeds 0.1 eV/Å. The data is limited to molecules with up to 7 heavy atoms, including C, N, and O. In order to train truly generalizable models we need to include all elements and sizes of molecules.

ANI1x[18,17] is a dataset based on active-learning and MD. The data generation procedure alternates between proposing new configurations using various forms of MD and pseudo-MD, and accepting or rejecting data based on the query by committee algorithm[23]. The dataset contains force and energy calculations for 6 million molecular configurations.

QM9 (QM9x) is a ubiquitous benchmark dataset for Quantum Mechanics (QM) methods that contains a multitude of QM features for 135.000 molecular equilibrium configurations. In order to make QM9 compatible with ANI1x and Transition1x we recalculated all configurations with the appropriate level of DFT, and we refer to it as QM9x.

PaiNN

PaiNN[14] is a message-passing GNN architecture designed for predicting molecular properties of atomic systems represented as graphs. The molecular graph is generated by a neighborhood function

	Barrier			NEB Convergence		
	MAE [eV]	RMSE [eV]	RMSD [Å]	Rate	Avg. CPU Time	Avg. Iterations
ANI1x	0.51(1)	1.67(3)	0.28(2)	69.3%	37s	149
T1x	**0.23(3)**	**0.52(1)**	**0.21(1)**	80.3%	33s	135
QM9x	3.4(1)	3.59(8)	0.59(2)	35.0%	**28s**	111
DFTB	0.70	0.85	0.22	65.7%	82s	114
DFT	-	-	-	**84.1%**	12h14m43s	**101**

Table 1: Comparison of potentials for Nudged Elastic Band (NEB). ANI1x, Transition1x and QM9x indicate PaiNN models trained on the respective dataset. The barrier column shows Mean Average Error (MAE) and Root Mean Squared Error (RMSE) of activation energies, and the Root Mean Square Deviation of atomic positions (RMSD) between transition states found with DFT and surrogate potentials. The Convergence column displays the convergence rate, average CPU time to compute a reaction, and average iterations before convergence.

that assigns edges between atoms if they are sufficiently close. The network calculates molecular features by letting neighboring atoms exchange messages calculated from their internal representations. Conservative forces can be calculated by the back-propagation algorithm as the negative gradient of energy with respect to positions.

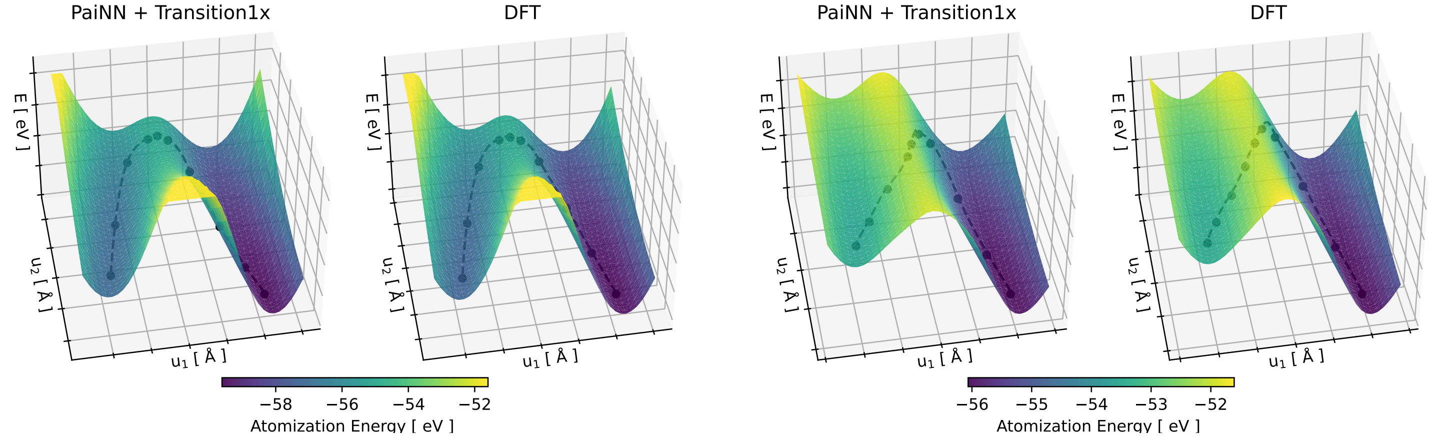
Five hundred reactions are set aside from Transition1x for evaluating various models' capabilities to find reaction pathways. Five equivalent PaiNN models are trained on each of ANI1x, QM9x, and the remaining data from Transition1x. The models have three hidden layers with 256 neurons in each. The molecular graph is generated with a cutoff radius of 5 Å. The models are trained with the ADAM optimizer, a batch size of 75, an initial learning rate of 10^{-3}, and a scheduler that scales the learning rate with a factor of 0.8 if the model has not improved for 5000 training steps.

Results

We trained five PaiNN models on each of Transition1x, ANI1x, and QM9x, and evaluated their capability to predict transition states on test reactions set aside from Transition1x. In Table 1 we report convergence rate, timings, Mean Average Error (MAE), Root Mean Squared Error (RMSE) and Root Mean Square Deviation of atomic positions (RMSD) of the various models' predictions. Datasets in the leftmost column represent PaiNN models trained on the respective dataset. Density Functional based Tight Binding (DFTB), is a fast and cheap approximation to DFT, used as a benchmark. We do not report standard deviation of MAE and RMSE of DFTB predictions as the algorithm is deterministic, whereas the performance of the PaiNN models depends on the training seed. The predicted activation energy is the difference between energies of reactant and transition state, while the error is the difference between the activation energies predicted by the surrogate potential and DFT. ML potentials trained on Transition1x outperform all other tested surrogate potentials in terms of MAE, RMSE and RMSD. QM9x contains only equilibrium configurations, and its models have not learned the intricacies of the PES around reaction pathways. This is reflected in the results through low convergence rates and high errors.

Figure 1 compares cross sections of PESs, spanned by reactant, product, and transition state of the reactions, calculated using DFT, and PaiNN trained on Transition1x, for two different reactions. The x and y axes are basis vectors describing the plane in units of Å, and the z-axis and color-coding show the atomization energy of configurations in the plane in eV. Not only does PaiNN trained on the Transition1x accurately calculate the barrier energy for the reaction, it also predicts an almost identical PES in the vicinity of the MEP and correctly identifies the plane spanned by the configurations defining the reaction.

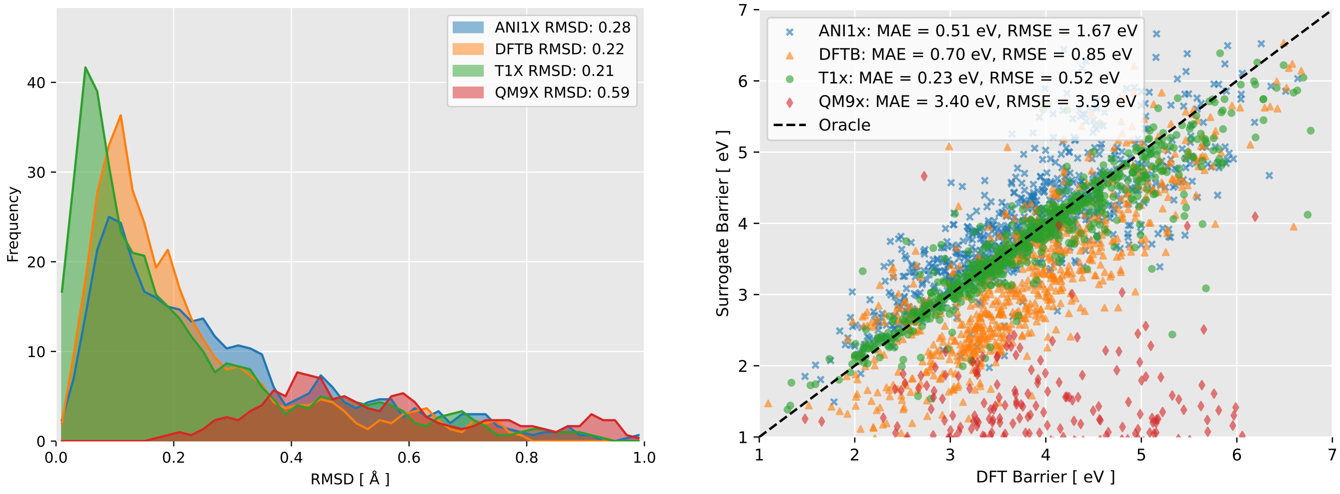
Figure 2 displays the performance of each surrogate potential. Panel *a* is the distribution of RMSDs between transition states predicted by surrogate potentials and DFT. The distributions are unnormalized to reflect convergence ratios. Panel *b* compares activation energies found by DFT on the x-axis with energies found by surrogate potentials on the y-axis. Models trained on ANI1x tend to overestimate activation energies as they have not seen the low energy valleys connecting equilibrium configurations, while DFTB tends to underestimate activation energies. Correcting ANI1x and DFTB



(a) Reaction involving C4N2H8. Follow this link to see GIF of reaction.

(b) Reaction involving C4NOH7. Follow this link to see GIF of reaction.

Figure 1: Minimal Energy Paths (MEPs) for two different reactions, found with the Nudged Elastic Band (NEB) algorithm, applying the Graph Neural Network (GNN) architecture PaiNN, trained on Transition1x, and DFT as potentials. The MEPs are projected onto intersections of the Potential Energy Surfaces (PESs) spanned by product, reactant and transition-state of the converged MEPs. The x- and y-axes are basis vectors describing the plane. The PESs have been calculated in the vicinity of the MEPs with the respective potential and is shown with colors and on the z-axis.



(a)

(b)

Figure 2: Performance of surrogate potentials compared to DFT. Panel *(a)* displays the unnormalized distribution of Root Mean Square Deviation of atomic positions (RMSD) between transition-states found by DFT and surrogate potentials. Panel *(b)* displays activation energies found by DFT on the x-axis, and found by surrogate potentials on the y-axis. Points that lie on the dashed line have been calculated perfectly.

for systematic error leads to a MAE of 0.48 eV and 0.48 eV and RMSE of 1.66 eV and 0.62 eV for ANI1x models and DFTB, respectively.

Conclusion

Relevant data is as important as expressive models for solving higher-order tasks in computational chemistry. We have presented the Transition1x dataset which contain force and energy calculations for 10 million molecular configurations on and around reaction pathways, and used it to train a fast and accurate ML calculator that can predict reaction pathways for general organic reactions. We achieved a MAE of 0.23 eV on activation energies on unseen reactions when compared to evaluating the PES with DFT, while simultaneously speeding up the MEP-search significantly.

Impact Statement

Transition1x and QM9x

We believe that Transition1x [24] is an important contribution to the completeness of available data in the literature for ML for molecular science. It provides a different type of data, facilitating new downstream tasks for ML models. It is calculated with the 6-31G(d) [21] basis set and ωB97x [22] functional, which makes it compatible with the ANI1x [17,18], and permits training models with rich representations by leveraging strengths from both datasets. The ubiquitous QM9 [25], dataset was recalculated with the appropriate level of theory and released under the name QM9x [24]. Data loaders, examples, and scripts for Transition1x and QM9x are available in their respective repositories - Transition1x: https://gitlab.com/matschreiner/Transition1x, and QM9x: https://gitlab.com/matschreiner/QM9x. The data collection procedure for Transition1x is scalable and can easily be extended to include new elements and reactions.

NeuralNEB

NeuralNEB is an accurate, inexpensive, and general ML calculator for transition-state and MEP search that outperforms NEB with DFTB [26] as potential, both in terms of accuracy and computational cost. It yields a new and better trade-off between speed and accuracy for screening of large reaction-networks.

References

[1] Daniel Sheppard, Rye Terrell, and Graeme Henkelman. Optimization methods for finding minimum energy paths. *The Journal of Chemical Physics*, 128:134106, 4 2008. ISSN 0021-9606. doi: 10.1063/1.2841941. URL https://aip.scitation.org/doi/abs/10.1063/1.2841941.

[2] Davide Bacciu, Federico Errica, Alessio Micheli, and Marco Podda. A gentle introduction to deep learning for graphs. *Neural Networks*, 129:203–221, 12 2019. doi: 10.1016/j.neunet.2020.06.006. URL http://arxiv.org/abs/1912.12693http://dx.doi.org/10.1016/j.neunet.2020.06.006.

[3] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. 2021. doi: 10.1016/j.aiopen.2021.01.001. URL https://doi.org/10.1016/j.aiopen.2021.01.001.

[4] Felix A. Faber, Luke Hutchison, Bing Huang, Justin Gilmer, Samuel S. Schoenholz, George E. Dahl, Oriol Vinyals, Steven Kearnes, Patrick F. Riley, and O. Anatole Von Lilienfeld. Prediction errors of molecular machine learning models lower than hybrid dft error. *Journal of Chemical Theory and Computation*, 13:5255–5264, 11 2017. ISSN 15499626. doi: 10.1021/ACS.JCTC.7B00577. URL https://pubs.acs.org/doi/abs/10.1021/acs.jctc.7b00577.

[5] Julia Westermayr, Michael Gastegger, Kristof T. Schütt, and Reinhard J. Maurer. Perspective on integrating machine learning into computational chemistry and materials science. *The Journal of Chemical Physics*, 154:230903, 6 2021. ISSN 0021-9606. doi: 10.1063/5.0047760. URL https://aip.scitation.org/doi/abs/10.1063/5.0047760.

[6] Stuart I Campbell, Daniel B Allan, and Andi M Barbour. Machine learning for the solution of the schrödinger equation. *Machine Learning: Science and Technology*, 1:013002, 4 2020. ISSN 2632-2153. doi: 10.1088/2632-2153/AB7D30. URL https://iopscience.iop.org/article/10.1088/2632-2153/ab7d30https://iopscience.iop.org/article/10.1088/2632-2153/ab7d30/meta.

[7] Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical review letters*, 98(14):146401, 2007.

[8] Julia Westermayr and Philipp Marquetand. Machine learning for electronically excited states of molecules. *Chemical Reviews*, 121:9873–9926, 8 2021. ISSN 15206890. doi: 10.1021/ACS.CHEMREV.0C00749. URL https://pubs.acs.org/doi/full/10.1021/acs.chemrev.0c00749.

[9] Jean Louis Reymond. The chemical space project. *Accounts of Chemical Research*, 48:722–730, 3 2015. ISSN 15204898. doi: 10.1021/AR500432K. URL https://pubs.acs.org/doi/full/10.1021/ar500432k.

[10] Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical Review Letters*, 98:146401, 4 2007. ISSN 00319007. doi: 10.1103/PHYSREVLETT.98.146401/FIGURES/4/MEDIUM. URL https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.98.146401.

[11] Jörg Behler. Constructing high-dimensional neural network potentials: A tutorial review. *International Journal of Quantum Chemistry*, 115:1032–1050, 8 2015. ISSN 1097-461X. doi: 10.1002/QUA.24890. URL https://onlinelibrary.wiley.com/doi/full/10.1002/qua.24890.

[12] Felix A. Faber, Luke Hutchison, Bing Huang, Justin Gilmer, Samuel S. Schoenholz, George E. Dahl, Oriol Vinyals, Steven Kearnes, Patrick F. Riley, and O. Anatole Von Lilienfeld. Prediction errors of molecular machine learning models lower than hybrid dft error. *Journal of Chemical Theory and Computation*, 13:5255–5264, 11 2017. ISSN 15499626. doi: 10.1021/ACS.JCTC.7B00577. URL https://pubs.acs.org/doi/abs/10.1021/acs.jctc.7b00577.

[13] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. 2017.

[14] Kristof T Schütt, Sch¨ Schütt, Oliver T Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. 2021.

[15] Jens Henriksson, Christian Berger, Markus Borg, Lars Tornberg, Sankar Raman Sathyamoorthy, and Cristofer Englund. Performance analysis of out-of-distribution detection on various trained neural networks. 2021. URL `https://www.iso.org/deliverables-all.html`.

[16] Lily H Zhang, Mark Goldstein, and Rajesh Ranganath. Understanding failures in out-of-distribution detection with deep generative models. 2021.

[17] Justin S. Smith, Ben Nebgen, Nicholas Lubbers, Olexandr Isayev, and Adrian E. Roitberg. Less is more: Sampling chemical space with active learning. *The Journal of Chemical Physics*, 148: 241733, 5 2018. ISSN 0021-9606. doi: 10.1063/1.5023802. URL `https://aip.scitation.org/doi/abs/10.1063/1.5023802`.

[18] Justin S. Smith, Roman Zubatyuk, Benjamin Nebgen, Nicholas Lubbers, Kipton Barros, Adrian E. Roitberg, Olexandr Isayev, and Sergei Tretiak. The ani-1ccx and ani-1x data sets, coupled-cluster and density functional theory properties for molecules. *Scientific Data 2020 7:1*, 7:1–10, 5 2020. ISSN 2052-4463. doi: 10.1038/s41597-020-0473-z. URL `https://www.nature.com/articles/s41597-020-0473-z`.

[19] Graeme Henkelman, Blas P. Uberuaga, and Hannes Jónsson. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *The Journal of Chemical Physics*, 113:9901, 11 2000. ISSN 0021-9606. doi: 10.1063/1.1329672. URL `https://aip.scitation.org/doi/abs/10.1063/1.1329672`.

[20] Colin A. Grambow, Lagnajit Pattanaik, and William H. Green. Reactants, products, and transition states of elementary chemical reactions based on quantum chemistry. *Scientific Data*, 7, 12 2020. ISSN 20524463. doi: 10.1038/s41597-020-0460-4.

[21] R. Ditchfield, W. J. Hehre, and J. A. Pople. 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, 9 2003. ISSN 0021-9606. doi: 10.1063/1.1674902. URL `https://aip.scitation.org/doi/abs/10.1063/1.1674902`.

[22] Jeng Da Chai and Martin Head-Gordon. Systematic optimization of long-range corrected hybrid density functionals. *The Journal of Chemical Physics*, 128:084106, 2 2008. ISSN 0021-9606. doi: 10.1063/1.2834918. URL `https://aip.scitation.org/doi/abs/10.1063/1.2834918`.

[23] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, pages 287–294, 1992. doi: 10.1145/130385.130417. URL `https://www.researchgate.net/publication/221497539_Query_by_Committee`.

[24] Schreiner M., Bhowmik A., Vegge T., and Busk J.and Winther O. Transition1x – a Dataset for Building Generalizable Reactive Machine Learning Potentials. 6 2022. doi: 10.6084/m9.figshare.19614657.v4. URL `https://figshare.com/articles/dataset/Transition1x/19614657`.

[25] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data 2014 1:1*, 1:1–7, 8 2014. ISSN 2052-4463. doi: 10.1038/sdata.2014.22. URL `https://www.nature.com/articles/sdata201422`.

[26] Gotthard Seifert and Jan Ole Joswig. Density-functional tight binding—an approximate density-functional theory method. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2:456–465, 5 2012. ISSN 1759-0884. doi: 10.1002/WCMS.1094. URL `https://onlinelibrary.wiley.com/doi/full/10.1002/wcms.1094`.

Checklist

1a) *Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?* We believe that the contribution of our dataset and NeuralNEB algorithm solves the problems we've outlined in the abstract and introduction.

1b) *Have you read the ethics review guidelines and ensured that your paper conforms to them?* Yes

1c) *Did you discuss any potential negative societal impacts of your work?* The work we have done is of an extremely abstract character and it is extremely hard to guess the impact it will have on society.

1d) *Did you describe the limitations of your work?* The limitations are limited elements and size of molecules in the dataset. We mention this.

3a*) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)?* Yes, we have made thoroughly documented repositories that can reproduce the datasets and run the algorithms presented in this paper.

3b) *Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)?* Yes.

3c) *Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)?* Yes.

3d) *Did you include the amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)?* Yes

4a) *If your work uses existing assets, did you cite the creators?* Yes

4b) *Did you mention the license of the assets?* No

4c) *Did you include any new assets either in the supplemental material or as a URL?* No

4d) *Did you discuss whether and how consent was obtained from people whose data you're using/curating?* No

4e) *Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content?* No, it is molecular data.