How robust are modern graph neural network potentials in long and hot molecular dynamics simulations?

IOP Publishing Mach. Learn.: Sci. Technol. 3 (2022) 045010 https://doi.org/10.1088/2632-2153/ac9955

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
26 April 2022

REVISED
22 July 2022

ACCEPTED FOR PUBLICATION
11 October 2022

PUBLISHED
1 November 2022

PAPER

How robust are modern graph neural network potentials in long and hot molecular dynamics simulations?

Sina Stocker[1,2,3], **Johannes Gasteiger**[2,3], **Florian Becker**[2], **Stephan Günnemann**[2] and **Johannes T Margraf**[1,*]

[1] Fritz-Haber-Institute of the Max-Planck-Society, Berlin, Germany
[2] Technical University of Munich, Munich, Germany
[3] The first two authors contributed equally.
[*] Author to whom any correspondence should be addressed.

E-mail: margraf@fhi.mpg.de

Keywords: graph neural networks, interatomic potentials, molecular dynamics

Supplementary material for this article is available online

Abstract

Graph neural networks (GNNs) have emerged as a powerful machine learning approach for the prediction of molecular properties. In particular, recently proposed advanced GNN models promise quantum chemical accuracy at a fraction of the computational cost. While the capabilities of such advanced GNNs have been extensively demonstrated on benchmark datasets, there have been few applications in real atomistic simulations. Here, we therefore put the robustness of GNN interatomic potentials to the test, using the recently proposed GemNet architecture as a testbed. Models are trained on the QM7-x database of organic molecules and used to perform extensive molecular dynamics simulations. We find that low test set errors are not sufficient for obtaining stable dynamics and that severe pathologies sometimes only become apparent after hundreds of ps of dynamics. Nonetheless, highly stable and transferable GemNet potentials can be obtained with sufficiently large training sets.

Atomistic simulations are an invaluable tool for gaining mechanistic and structural insight into chemical systems, including solid materials [1], interfaces [2, 3], liquids [4] or even complex biological systems like the SARS-CoV-2 virus [5]. They are also becoming increasingly important in the design of new materials and drugs [6, 7]. In many ways, the prototypical atomistic simulation is a molecular dynamics (MD) trajectory, which propagates the atomic coordinates of a system in time, starting from some initial conditions. MD simulations are extremely common, both by themselves and as part of more elaborate sampling procedures like parallel tempering or metadynamics.

In principle, highly accurate MD trajectories can be obtained from electronic structure methods like density functional theory (DFT). Unfortunately, such *ab initio* MD (AIMD) simulations require the (approximate) solution of the electronic Schrödinger equation at every time step. This makes them very expensive from a computational perspective and ultimately limits the applicability of AIMD to a few hundreds of atoms and relatively short (i.e. ps) timescales. For many scientific questions, simulations of much larger systems, longer timescales or (usually) both are required. To this end, empirical interatomic potentials are typically used. These provide an analytical expression for high-dimensional potential energy surfaces which can be evaluated in a small fraction of the time required for a DFT calculation. This gain in efficiency invariably comes at the expense of a decrease in accuracy and/or transferability, however.

To bridge this gap between computational cost and accuracy, machine learned interatomic potentials have recently gained popularity in computational chemistry [8–11] and materials science [12–14]. In particular, a range of neural network [15–17] and kernel based potentials [18, 19] have been developed and applied to a wide variety of chemical systems. While somewhat more expensive than classical force fields, these potentials are able to predict energies and forces with DFT accuracy and have thus become an important part of the toolbox of computational chemistry.

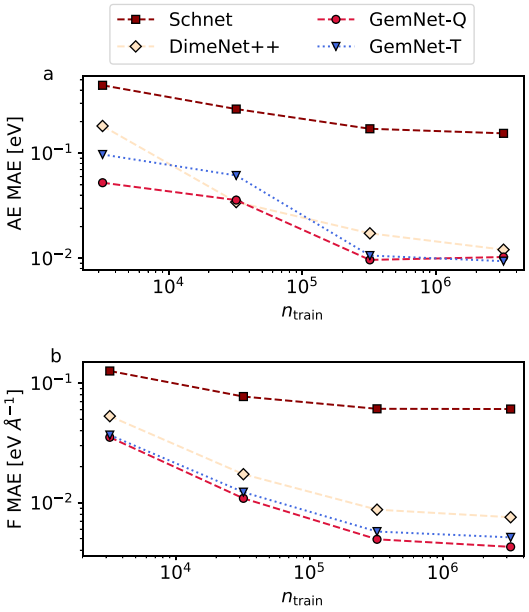

Figure 1. Learning curves. (a) Mean absolute errors (MAEs) of atomization energies (AEs) and (b) forces (F) against the number of training configurations. MAEs are calculated for a test set consisting of 10 100 random configurations from the QM7-x database.

One of the most recent additions to this family of methods are potentials based on graph neural networks (GNNs), such as SchNet, DimeNet, GemNet and NequIP [20–31]. Here, much progress towards ever more accurate and expressive potentials has been made, e.g. by using equivariant formulations or embedding atom pairs and triplets. While such efforts naturally focus on established benchmark databases like QM9 [32, 33], MD17 [34] or OC20 [35], comparatively little research has been conducted to show the applicability of such advanced GNN potentials in real atomistic simulations. A notable exception to this is a recent paper of Batzner and coworkers [22], which demonstrated that potentials based on the equivariant NequIP architecture could be used in stable and accurate MD simulations, when trained on AIMD data for the respective system.

In this contribution, we aim to provide an in-depth exploration of the robustness of state-of-the-art GNN potentials based on the GemNet architecture [21] in MD simulations. To this end, we ran a total of 280 ns of dynamics (more than 500 million timesteps) across a wide range of temperatures and organic molecules. By checking samples from these large ensembles with DFT reference calculations, the extrapolative capabilities of the potentials in configuration and chemical space was tested simultaneously. Furthermore, the impact of training set size on the robustness of the potentials was explored.

GNNs treat chemical systems as graphs, with nodes representing atoms and edges representing interactions between atom pairs. While traditional chemical graph representations usually equate edges with covalent bonds, GNNs assume edges between all atoms within a given cutoff distance. Most of the potentials discussed in the following are based on the geometric message passing neural network (GemNet) [21], which shows excellent performances on established benchmark data sets like MD17 and OC20 as well as QM7-x (see figure 1). GemNet embeds both the atoms and the interatomic edges via high-dimensional vectors. Both kinds of embeddings are then updated in multiple layers using learnable weight matrices and by passing messages between the edges and atoms within a given cutoff distance. GemNet leverages the full geometric information for this: the interatomic distances, the angles between neighboring edges, and the dihedral angles defined via triplets of edges. From the learned embeddings, energy contributions for each atom and layer are obtained, which are subsequently summed up to calculate the total energy of the system (see SI for details). The whole model is continuously differentiable, which allows calculating the forces via $\mathbf{F}_i = -\frac{\partial}{\partial \mathbf{x}_i} E$. As for all GNNs, the use of a finite cutoff and per-atom energy contributions makes the predictions size-extensive and the computational cost scale linearly with the number of atoms. Note that for comparison we also use a slightly simplified model termed GemNet-T [21], which does not explicitly use dihedral information. To avoid confusion, the full GemNet model is termed GemNet-Q.

Herein, we trained several GemNet potentials on different subsets of the recently published QM7-x database [36]. This dataset consists of around 4.2 million configurations sampled from small organic molecules consisting of up to seven non-hydrogen atoms (i.e. C, O, N, S, Cl), with 4–23 atoms in total. Importantly, QM7-x covers both equilibrium and non-equilibrium structures. Starting from 6950 structural

formulas, it contains around 41 500 equilibrium structures (including stereoisomers and conformers) and 100 additional non-equilibrium structures for each equilibrium geometry. The latter were generated by applying linear combinations of normal mode displacements to each configurations, thus approximately mimicking MD within the harmonic approximation. For each configuration, total energies and forces at the hybrid DFT (PBE0) [37] level with a many-body dispersion correction [38] are provided, computed with tightly converged numerical atom-centered basis sets and integration grids [39, 40] (see [36] for full details).

All potentials were trained on atomization energies (AEs) and forces (F) simultaneously. Since forces are ultimately the driver of MD simulations and contain more fine-grained information than energies, forces were weighted more strongly in our fits, so that the AE only contributes 0.1% to the loss function (see SI for details). This essentially follows the philosophy of gradient domain machine learning [34, 41], which exclusively uses forces. However, we do include a small AE contribution to the loss, as energy differences across chemical space cannot be learned effectively from forces alone [42]. For training, the QM7-x dataset was randomly split into a test set of 10 100 configurations, training sets of 3.2 k, 32 k, 320 k and 3.2 Mio configurations and corresponding validation sets of 800, 8 k, 80 k and 800 k configurations (the latter being used for hyperparameter selection, see SI). In the interest of simplification, we will denote models trained on small (3.2 k and 32 k) and large (320 k and 3.2 Mio) training sets as 'sparse' and 'exhaustive' models respectively.

In figure 1, the corresponding learning curves for AE and F are shown. The force curve shows a roughly linear decrease on the log–log scale between 3.2 k and 320 k training configurations but levels off between 320 k and 3.2 Mio configurations. This indicates that the more exhaustive models approach the intrinsic accuracy that is possible given the precision of the data and limitations of the models themselves (e.g. due to the cutoffs employed or the incompleteness of structural descriptors [43]). Due to the lower weighting of energies in the loss the AE curve is somewhat more noisy but follows the same trend.

To put this performance into perspective, the most exhaustive GemNet-Q model yields a force MAE of 0.0043 eV Å^{-1}, which can be compared with an MAE of 0.015 eV Å^{-1} for the recently developed SpookyNet [23] architecture (in this case trained on 4.2 Mio molecules). In addition, GemNet-Q and GemNet-T outperform SchNet [28] and DimeNet++ [25] on QM7-x for nearly all points of the learning curve (with the only exception being the AE error of the 32 k model). Importantly, the energy errors are also very low (0.01 eV = 0.23 kcal mol^{-1}) despite the low weighting of AEs in the loss. It is furthermore notable that even the model trained on 3.2 k configurations displays quite good performance with MAEs of around 0.035 eV Å^{-1} and 0.05 eV ($\approx$1 kcal mol^{-1}). The performance of GemNet-Q and GemNet-T is very similar, in particular for the larger training sets.

To explore the robustness of the GemNet potentials within the scope of their training set, constant temperature MD simulations were performed for 20 representative molecules from QM7-x (see figure 2(a), and the SI for the corresponding GemNet-T results). Here, care was taken to include all atom types in the dataset. For each molecule, 1 ns trajectories were generated with a 0.5 fs timestep at three different temperatures (300, 600 and 1200 K), using all models presented in the learning curve (see SI for details on the MD simulations). The rationale for using these temperatures is that they lead to increasingly extensive exploration of phase space. Indeed, it is not uncommon to use high temperature dynamics for this purpose, e.g. in replica exchange MD [44]. From each trajectory, 72 configurations were uniformly sampled and the corresponding energies and forces computed with identical DFT settings to the ones used for the QM7-x set.

Figures 2(b) and (c) show the AE and F MAEs for these samples as a function of temperature and training set size. Here, opaque symbols and lines represent MAEs averaged over 20 different trajectories corresponding to a given model and temperature. Transparent symbols and lines illustrate the MAEs for each trajectory individually, to provide some insight into the spread of MAEs for different molecules (see SI for additional illustrations of the respective error distributions). Overall, we find quite consistent trends for both AE and F predictions. Whereas the exhaustive models (320 k and 3.2 Mio) only display a very slight increase of the MAEs with temperature, the errors of the sparse models (3.2 k and 32 k) increase dramatically. This is expected, as higher temperature MD simulations more extensively explore the phase space and consequently move away from the training configurations.

Notably, the 3.2 k model already displays a very large AE error of more than 10 eV at 300 K. The MD error is thus orders of magnitude larger than the test set error, even though these configurations should arguably fall within the scope of the training set. This mainly stems from the fact that the trajectories for certain molecules lead to completely unphysical configurations, for which the potential then displays extremely large errors. Such unphysical configurations also commonly lead to convergence issues in the reference DFT calculations. To quantify this, the percentage of converged DFT calculations for configurations obtained with a given potential and simulation temperature is shown in figure 2(d). We find that all DFT calculations converge for the 320 k and 3.2 Mio potentials, while the sparse models generate increasingly

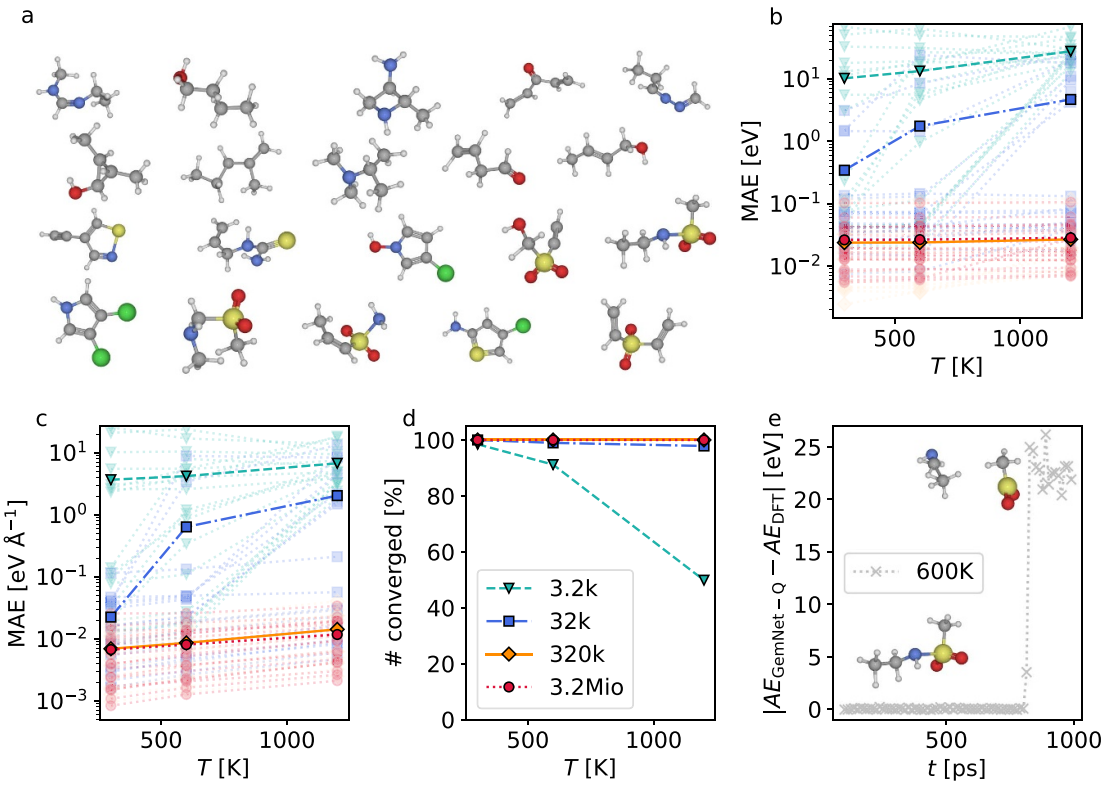

Figure 2. Robustness of GemNet-Q potentials in MD. (a) Representative molecules from QM7-x used in the MD tests. (b), (c) Mean absolute errors (MAEs) on atomization energy (AE) and force (F) predictions as a function of training set size and temperature. Opaque lines and symbols represent the average MAE over all molecules in (a). Transparent lines and symbols represent the MAE of one specific molecule. (d) Percentage of converged DFT calculations for configurations generated with different potentials and temperatures. (e) AE error as a function of simulation time for a 1 ns MD trajectory at 600 K, using a potential trained with 3.2 k configurations. This sub-panel shows a drastic deterioration in energy predictions after around 700 ps, when the molecule dissociates into fragments that cannot be accurately described by the model.

unphysical configurations with rising temperature. This is particularly evident for the 3.2 k model at 1200 K, where only about half of the DFT calculations converge.

The marked discrepancy between the test set and MD performance of the 3.2 k model underscores the limitations of using test configurations that are not generated by the potential itself. For a ML model to be useful in atomistic simulations, it is not sufficient to show that it provides accurate fits for physically reasonable configurations. It is equally important that the model avoids unphysical configurations in its own simulations. Note that testing this requires sufficiently long trajectories. This is illustrated for a representative example in figure 2(e). Here, the error of the 3.2 k model is actually quite low for the first 700 ps of the simulation after which it rises sharply to more than 20 eV due to an unphysical bond dissociation event. This behavior can be understood as a kind of 'hole' in the potential energy surface of the model, as previously described by Behler [45]. This hole can be rather small, but once the simulation reaches such a configuration the trajectory is completely unreasonable. The 'robustness' of a ML potential can thus be understood as a measure of how frequent and how large such holes in the potential energy surface are. Ultimately this can only be quantified by performing long MD simulations with the corresponding potential.

It should be noted that a common strategy to obtain more robust potentials is the use of active learning or diversity-driven data selection [45]. The idea behind this is that holes in the potential energy surface are associated with configurations that are significantly different from the training data. While active learning strategies are beyond the scope of the current work, it is nonetheless interesting to consider how the robustness of the sparse models is influenced by more sophisticated data selection schemes than uniform random sampling. To explore this, an additional model (referred to as the highE model in the following) was trained with a modified training set, also consisting of 3.2 k configurations. Here, the probability of drawing a configuration was weighted by the average force norm, favoring more distorted molecules.

As shown in the SI, this leads to similar force errors (MAE = 0.037 eV Å^{-1} vs. 0.035 meV Å^{-1} for uniform random sampling) but higher energy errors (MAE = 0.195 eV vs. 0.052 eV for uniform random sampling) for the test set, likely because the highE configurations are less representative of the test set overall. To test the robustness of the highE model, MD simulations at 1200 K were performed for three pathological molecules, which dissociated in the previous 3.2 k simulations (see SI). This leads to somewhat better energy

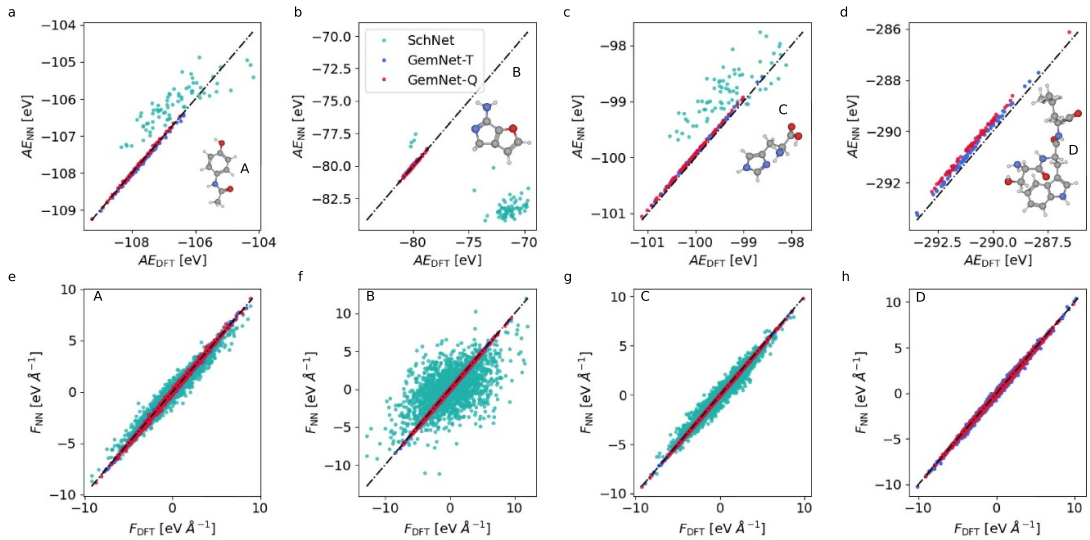

Figure 3. Out-of-sample validation of the interatomic potentials trained on 3.2 Mio configurations. (a)–(d) Correlation plots of atomization energies (AEs) for four out-of-sample molecules (A = Paracetamol, B = 4 H-Furo[2,3-c]pyrrol-6-amin, C = Histidin and D = Ser-Trp-Leu-tripeptide), obtained at 1200 K using the 3.2 Mio GemNet-Q, GemNet-T and SchNet potentials. (e)–(h) Correlation plots of predicted and reference forces (F). Note that molecule D immediately dissociates into unphysical fragments for SchNet, so that no DFT reference calculations could be obtained.

conservation (and thus a smoother potential energy surface) than for the original 3.2 k model. Nevertheless, all three molecules dissociate within around 100 ps. Improved data selection alone therefore does not lead to robust models in this case.

It should be stressed that this notion of robustness is not necessarily correlated with the test MAE, despite the fact that the robust GemNet models also display much lower MAEs. Indeed, the robustness of traditional bio-organic forcefileds with fixed topologies is very high. However, in this case robustness is gained at the expense of model flexibility. The challenge for ML potentials is that they must be robust without sacrificing flexibility. Our tests show that this is not trivial. On a more positive note, we do find that GemNet potentials with sufficiently large training sets are very robust across the phase space of the QM7-x dataset and beyond.

Another way to illustrate this is to consider the performance of the 3.2 k model for a trajectory generated with the 3.2 Mio potential in comparison with its own trajectory. Specifically this means that we generate two independent trajectories with the 3.2 k and 3.2 Mio model and evaluate MAEs of the 3.2 k model for configurations drawn from each trajectory. Taking the molecule in figure 2(e) at 1200 K, the F MAE of the 3.2 k potential is 6.8 eV Å^{-1} for the 3.2 k trajectory but only 0.16 eV Å^{-1} when it is evaluated on the 3.2 Mio trajectory. Again, the sparse model performs quite well for the physically reasonable configurations generated with the 3.2 Mio model. The problem only becomes apparent when testing the sparse model on its own trajectory.

Having established the robustness of the exhaustive models within the scope of QM7-x, we now turn to simultaneous extrapolation in chemical and configuration space. To this end, we consider four molecules consisting of 9–29 heavy atoms (i.e. which are significantly larger than the training molecules). Again, 1 ns MD trajectories were generated with the 3.2 Mio SchNet, GemNet-T and GemNet-Q potentials at 1200 K. Figure 3 shows the corresponding AE and F correlations with the DFT reference data. Here, the GemNet-T and GemNet-Q AEs are systematically less negative than the DFT reference energies, most prominently for the large Ser-Trp-Leu tripeptide. Here, the mean GemNet-Q AE is shifted by 0.47 eV with respect to the reference, which is substantial when compared to an MAE of 0.0284 eV at 1200 K in figure 2(b).

This shift can potentially be explained by the absence of attractive long-range interactions (e.g. dispersion or electrostatics) in the GemNet-Q potential or by basis-set superposition errors in the DFT data. While message-passing neural networks can in principle include information from beyond their cutoff distance, the QM7-x database exclusively consists of small molecules so that long-range interactions simply cannot be learned from it. Methods to include long-range interactions are proposed in literature [23, 27, 46, 47] and could also be applied to the GemNet architecture. Nonetheless, GemNet and DFT energies are highly correlated ($R^2 = 0.998$ for GemNet-Q, see SI) and the standard deviation of the AE error distribution is only 0.045 eV so that the MD trajectory for this system should still be considered to be of high quality. While the long-range interactions are thus considerable in magnitude, they do not fluctuate very strongly [48]. This is

also the case for the other molecules, for which GemNet-Q displays very narrow AE error distributions. Similarly, force component errors are consistently small, with MAEs between even 0.012 and 0.036 eV $Å^{-1}$.

In contrast, the 3.2 Mio SchNet model is substantially less robust for these large molecules. This is particularly evident for the tripeptide which immediately dissociates, leading to completely unphysical fragments. For the other molecules the errors are also significantly larger and less systematic than for the GemNet models. This indicates that the additional geometric information used by the latter (angles and dihedrals) improves both the accuracy and the robustness of the corresponding interatomic potentials.

The remarkable robustness of the GemNet-Q potential raises the question at which point its extrapolative capabilities break down. We therefore ran additional MD simulations for the four large molecules at 1800 K, using the 3.2 Mio GemNet-Q model (see SI). Here, we indeed find that molecules C (Histidin) and D (the tripepeptide) decompose, while A and B remain stable throughout the simulations. Correspondingly, the C and D trajectories display considerable (several eV) deviations between the ML and DFT energies. Nonetheless, even under these conditions the trajectory leads fairly reasonable fragments, which can at least be characterized with DFT calculations.

In conclusion, we have explored the robustness of GNN potentials based on the recent GemNet architecture in MD simulations. We find that sufficiently large training sets are key to obtaining robust GNN potentials and that a low test set error does not guarantee that stable trajectories can be generated. Interestingly, in some cases severe instabilities were only discovered after hundreds of ps of dynamics. The test set error should thus not be taken at face value as a measure for the error one can expect in 'real' applications. Demonstrating 'chemical accuracy' on a test set is by itself not enough.

With large enough training sets, the GemNet potentials used herein are highly robust, however. This is demonstrated by applications in high-temperature MD simulations of systems that are significantly larger than the training molecules. In this extrapolative regime, errors are mostly systematic and explainable and no instabilities were observed. Interestingly, no significant improvements in terms of accuracy or robustness were observed when training on 3.2 Mio instead of 320 k samples, indicating that all relevant information about the underlying potential energy surface can be learned from less than 10% of the dataset. This is significant because robust ML potentials are often associated with iterative training procedures. Due to their size and complexity (the models used herein fit 2.2 million parameters), GNN models are *a priori* not ideal for such settings. Indeed, training times of several GPU weeks are not unusual, which is clearly impractical in an iterative workflow. Well curated databases like QM7-x and powerful model architectures like GemNet circumvent this issue.

As a final point, we note that the potentials discussed herein (as well as the underlying code) are freely available at https://www.cs.cit.tum.de/daml/gemnet/. We recommend the 3.2 Mio GemNet potential as a general-purpose force field for exploring the conformational space of small to medium organic molecules. Indeed, the accuracy and the robustness of the 320 k and 3.2 Mio models is high enough that they can be considered as a cost effective replacement of DFT calculations for this application. It remains to be seen whether equally accurate and robust models can be obtained for larger chemical spaces, broader sections of the periodic table and chemical reactions.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://www.cs.cit.tum.de/daml/gemnet/.

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) through TUM International Graduate School of Science and Engineering (IGSSE) (GSC 81).

ORCID iDs

Johannes Gasteiger https://orcid.org/0000-0001-9388-6389
Johannes T Margraf https://orcid.org/0000-0002-0862-5289

References

[1] Deringer V L, Bernstein N, Csányi G, Mahmoud C B, Ceriotti M, Wilson M, Drabold D A and Elliott S R 2021 Origins of structural and electronic transitions in disordered silicon *Nature* **589** 59–64
[2] Stegmaier S *et al* 2021 Nano-scale complexions facilitate Li dendrite-free operation in LATP solid-state electrolyte *Adv. Energy Mater.* **11** 2100707

[3] Timmermann J et al 2020 IrO_2 surface complexions identified through machine learning and surface investigations Phys. Rev. Lett. 125 206101
[4] Cheng B, Mazzola G, Pickard C J and Ceriotti M 2020 Evidence for supercritical behaviour of high-pressure liquid hydrogen Nature 585 217–20
[5] Zimmerman M I et al 2021 SARS-CoV-2 simulations Go exascale to predict dramatic spike opening and cryptic pockets across the proteome Nat. Chem. 13 651–9
[6] Zhong M et al 2020 Accelerated discovery of CO_2 electrocatalysts using active machine learning Nature 581 178–83
[7] Artrith N, Urban A and Ceder G 2017 Efficient and accurate machine-learning interpolation of atomic energies in compositions with many species Phys. Rev. B 96 014112
[8] Unke O T, Chmiela S, Sauceda H E, Gastegger M, Poltavsky I, Schütt K T, Tkatchenko A and Müller K-R 2021 Machine learning force fields Chem. Rev. 121 10142–86
[9] Keith J A, Vassilev-Galindo V, Cheng B, Chmiela S, Gastegger M, Müller K-R and Tkatchenko A 2021 Combining machine learning and computational chemistry for predictive insights into chemical systems Chem. Rev. 121 9816–72
[10] Xu J, Cao X-M and Hu P 2021 Accelerating metadynamics-based free-energy calculations with adaptive machine learning potentials J. Chem. Theory Comput. 17 4465–76
[11] Stocker S, Csányi G, Reuter K and Margraf J T 2020 Machine learning in chemical reaction space Nat. Commun. 11 5505
[12] Kim J, Kang D, Kim S and Jang H W 2021 Catalyze materials science with machine learning ACS Mater. Lett. 3 1151–71
[13] Deringer V L, Bartók A P, Bernstein N, Wilkins D M, Ceriotti M and Csányi G 2021 Gaussian process regression for materials and molecules Chem. Rev. 121 10073–141
[14] Behler J and Csányi G 2021 Machine learning potentials for extended systems: a perspective Eur. Phys. J. B 94 142
[15] Behler J and Parrinello M 2007 Generalized neural-network representation of high-dimensional potential-energy surfaces Phys. Rev. Lett. 98 146401
[16] Smith J S, Isayev O and Roitberg A E 2017 ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost Chem. Sci. 8 3192–203
[17] Smith J S, Nebgen B T, Zubatyuk R, Lubbers N, Devereux C, Barros K, Tretiak S, Isayev O and Roitberg A E 2019 Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning Nat. Commun. 10 2903
[18] Bartók A P, Payne M C, Kondor R and Csányi G 2010 Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons Phys. Rev. Lett. 104 136403
[19] Bartók A P, Kondor R and Csányi G 2013 On representing chemical environments Phys. Rev. B 87 184115
[20] Schütt K, Unke O and Gastegger M 2021 Equivariant message passing for the prediction of tensorial properties and molecular spectra Int. Conf. on Machine Learning
[21] Gasteiger J, Becker F and Günnemann S 2021 GemNet: universal directional graph neural networks for molecules Neural Information Processing Systems
[22] Batzner S, Musaelian A, Sun L, Geiger M, Mailoa J P, Kornbluth M, Molinari N, Smidt T E and Kozinsky B 2022 E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials Nat. Commun. 13 2453
[23] Unke O T, Chmiela S, Gastegger M, Schütt K T, Sauceda H E and Müller K-R 2021 SpookyNet: learning force fields with electronic degrees of freedom and nonlocal effects Nat. Commun. 12 7273
[24] Gasteiger J, Groß J and Günnemann S 2020 Directional message passing for molecular graphs Int. Conf. on Learning Representations (ICLR)
[25] Gasteiger J, Giri S, Margraf J T and Günnemann S 2020 Fast and uncertainty-aware directional message passing for non-equilibrium molecules Machine Learning for Molecules Workshop (NeurIPS)
[26] Park C W, Kornbluth M, Vandermause J, Wolverton C, Kozinsky B and Mailoa J P 2020 Accurate and scalable multi-element graph neural network force field and molecular dynamics with direct force architecture (arXiv:2007.14444)
[27] Unke O T and Meuwly M 2019 PhysNet: a neural network for predicting energies, forces, dipole moments and partial charges J. Chem. Theory Comput. 15 3678–93
[28] Schütt K T, Kindermans P-J, Sauceda H E, Chmiela S, Tkatchenko A and Müller K-R 2017 SchNet: a continuous-filter convolutional neural network for modeling quantum interactions Neural Information Processing Systems
[29] Schütt K T, Arbabzadah F, Chmiela S, Müller K R and Tkatchenko A 2017 Quantum-chemical insights from deep tensor neural networks Nat. Commun. 8 13890
[30] Schütt K T, Kessel P, Gastegger M, Nicoli K A, Tkatchenko A and Müller K-R 2019 SchNetPack: a deep learning toolbox for atomistic systems J. Chem. Theory Comput. 15 448–55
[31] Zubatyuk R, Smith J S, Leszczynski J and Isayev O 2019 Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network Sci. Adv. 5 eaav6490
[32] Ruddigkeit L, van Deursen R, Blum L C and Reymond J-L 2012 Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17 J. Chem. Inf. Model. 52 2864–75
[33] Ramakrishnan R, Dral P O, Rupp M and von Lilienfeld O A 2014 Quantum chemistry structures and properties of 134 kilo molecules Sci. Data 1 140022
[34] Chmiela S, Tkatchenko A, Sauceda H E, Poltavsky I, Schütt K T and Müller K-R 2017 Machine learning of accurate energy-conserving molecular force fields Sci. Adv. 3 e1603015
[35] Chanussot L et al 2021 Correction to the open catalyst 2020 (OC20) dataset and community challenges ACS Catal. 11 13062–5
[36] Hoja J, Medrano Sandonas L, Ernst B G, Vazquez-Mayagoitia A, DiStasio Jr R A and Tkatchenko A 2021 QM7-X, a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules Sci. Data 8 43
[37] Perdew J P, Ernzerhof M and Burke K 1996 Rationale for mixing exact exchange with density functional approximations J. Chem. Phys. 105 9982–5
[38] Tkatchenko A, DiStasio R A, Car R and Scheffler M 2012 Accurate and efficient method for many-body van der Waals interactions Phys. Rev. Lett. 108 1–5
[39] Blum V, Gehrke R, Hanke F, Havu P, Havu V, Ren X, Reuter K and Scheffler M 2009 Ab initio molecular simulations with numeric atom-centered orbitals Comput. Phys. Commun. 180 2175–96
[40] Ren X, Rinke P, Blum V, Wieferink J, Tkatchenko A, Sanfilippo A, Reuter K and Scheffler M 2012 Resolution-of-identity approach to Hartree–Fock, hybrid density functionals, RPA, MP2 and GW with numeric atom-centered orbital basis functions New J. Phys. 14 053020
[41] Chmiela S, Sauceda H E, Müller K-R and Tkatchenko A 2018 Towards exact molecular dynamics simulations with machine-learned force fields Nat. Commun. 9 3887

[42] Christensen A S and von Lilienfeld O A 2020 On the role of gradients for machine learning of molecular energies and forces *Mach. Learn. Sci. Technol.* **1** 045018
[43] Pozdnyakov S N, Willatt M J, Bartók A P, Ortner C, Csányi G and Ceriotti M 2020 Incompleteness of atomic structure representations *Phys. Rev. Lett.* **125** 166001
[44] Petraglia R, Nicolai A, Wodrich M D, Ceriotti M and Corminboeuf C 2016 Beyond static structures: putting forth REMD as a tool to solve problems in computational organic chemistry *J. Comput. Chem.* **37** 83–92
[45] Behler J 2015 Constructing high-dimensional neural network potentials: a tutorial review *Int. J. Quantum Chem.* **115** 1032–50
[46] Ko T W, Finkler J A, Goedecker S and Behler J 2021 A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer *Nat. Commun.* **12** 398
[47] Staacke C, Wengert S, Kunkel C, Csányi G, Reuter K and Margraf J 2022 Kernel charge equilibration: efficient and accurate prediction of molecular dipole moments with a machine-learning enhanced electron density model *Mach. Learn. Sci. Technol.* **3** 015032
[48] Staacke C, Heenen H, Scheurer C, Csányi G, Reuter K and Margraf J 2021 On the role of long-range electrostatics in machine-learned interatomic potentials for complex battery materials *ACS Appl. Energy Mater.* **4** 12562–9