A Journal of the Gesellschaft Deutscher Chemiker

Angewandte Chemie

GDCh

International Edition

www.angewandte.org

Accepted Article

Title: Science-Driven Atomistic Machine Learning

Authors: Johannes Margraf

To be cited as: *Angew. Chem. Int. Ed.* **2023**, e202219170

Link to VoR: https://doi.org/10.1002/anie.202219170

WILEY-VCH

Science-Driven Atomistic Machine Learning

Johannes T. Margraf*
Fritz-Haber-Institute of the Max-Planck Society

(Dated: March 6, 2023)

Machine learning (ML) algorithms are currently emerging as powerful tools in all areas of science. Conventionally, ML is understood as a fundamentally *data-driven* endeavour. Unfortunately, large well-curated databases are sparse in chemistry. In this contribution, I therefore review *science-driven* ML approaches which do not rely on "big data", focusing on the atomistic modelling of materials and molecules. In this context, the term science-driven refers to approaches that begin with a scientific question and then ask what training data and model design choices are appropriate. As key features of science-driven ML, the automated and purpose-driven collection of data and the use of chemical and physical priors to achieve high data-efficiency are discussed. Furthermore, the importance of appropriate model evaluation and error estimation is emphasized.

I. INTRODUCTION

Machine learning (ML) is now an established part of several key areas of chemical research, *e.g.* in the development of interatomic potentials[1, 2], the analysis of complex simulation data[3] or the design of novel drugs[4] and materials[5]. Beyond being a methodological novelty, atomistic ML has enabled real scientific breakthroughs, *e.g.* in predicting protein structures[6] or understanding the properties of water[7, 8], silicon[9], and hydrogen under extreme conditions[10].

While chemical ML is an extraordinarily diverse subject (including applications in so-called self-driving labs[11] or in the analysis of experimental data[12]), atomistic ML is arguably one of its most mature subfields. Here, atomically resolved structural data serve as the main in- or outputs of a model. Among other reasons, the success of atomistic ML can be attributed to the facts that modern ML methods are inherently well suited for such high-dimensional problems, and that electronic structure calculations (most often using Density Functional Theory, DFT) offer a relatively straightforward way for generating high quality reference data.

Indeed, there is currently a veritable hype around ML for atomistic systems, with a multitude of new applications being reported every day. As is commonly the case with hypes, not all the reported benefits of ML hold up to scrutiny, however. For instance, comparisons with adequate (non-ML) baselines are often not performed and the applicability of the proposed methods beyond the scope of the training data is often unclear.[13]

Here, a certain disconnect between common practices in method development and the practical demands of atomistic modelers can be observed. For understandable reasons, the former prefer to focus on well established benchmark datasets. These are readily available and allow rigorously comparing new methods with the state-of-the-art. Unfortunately, these benchmark problems merely represent an imperfect proxy to real chemical research questions. Consequently, many proposed methods do not find their way into practical applications. Even more critically, the focus on specific benchmarks leads to certain trends in atomistic ML research (such as the development of ever larger deep learning models with millions of parameters) that may actually be detrimental for many practical applications.[14]

In this contribution, I aim to provide an overview of how the availability of data shapes research in atomistic ML, with a focus on the use of supervised learning in atomistic simulations. Based on this, I differentiate between data-driven and science-driven ML approaches and argue that the latter are essential for addressing many pressing scientific questions. Finally, key aspects of science-driven ML approaches are reviewed and promising future research directions are discussed.

II. BIG AND SMALL DATA

There is a famous quote attributed to Ernest Rutherford, that "all science is either physics or stamp collecting". This is often interpreted as disparaging sciences like chemistry and biology for lacking a deep understanding of the physical world and merely describing and categorizing a large variety of phenomena and observations. The authenticity of this quote is dubious and it is actually rather unlikely that Rutherford of all people would have belittled the value of empirical observations in science. Nevertheless, chemical research sometimes undeniably has a certain resemblance with stamp collecting (*e.g.* starting with "Beilstein's Handbook of Organic Chemistry" first published in 1881[15]). Rather than being a frivolous hobby, however, such efforts have led to the formation of essential databases that are used by millions of chemists every day.

The longest tradition of this can be found in organic chemistry, particularly in the field of molecular synthesis. Beyond the already mentioned Beilstein Handbook (now part of Reaxys), there are several public domain efforts like PubChem[16] or ChemSpider[17], each containing data on hundreds of millions of small (*i.e.* consisting of ca. 100 atoms or less) organic molecules. Such

* email: margraf@fhi.mpg.de

databases are far from complete given the size of chemical space, estimated to be on the order of 10^{60} molecules (even when only considering CHNOS-containing drug-like molecules[18]). Nonetheless, they have played a key role in the development of ML models in chemoinformatics. Here typical applications include synthesis planning or the generation of new molecules.[19–22]

As useful as these databases are, the kind of information they contain limits their applicability in atomistic ML. For example, molecules are usually represented in terms of strings (SMILES or InChI) or graphs, lacking full three dimensional information (although approximate 3D geometries are available in many cases). Furthermore, for many of the contained molecules the only available experimental information is the fact that it has been reported in the literature. Structure-property relations and regression models (*e.g.* for biological activities) are thus usually obtained from smaller annotated subsets of these resources. Meanwhile, ML applications aiming to predict 3D geometries or generate full dimensional structure-property mappings (*e.g.* potential energy surfaces) cannot rely on them at all.

This lack is to some extent addressed by databases collecting experimentally determined structures (mostly from X-ray diffraction), such as the Cambridge Structural Database (CSD),[23] the Protein Data Bank (PDB),[24] or the Inorganic Crystal Structure Database (ICSD)[25]. These provide high quality insights into the 3D structure of molecules, proteins and inorganic solids, respectively. Given the expense and technical difficulty of performing such experiments, these databases are orders of magnitude smaller than the aforementioned ones, however (between 100,000 and one million entries). At the same time, each entry contains a much greater depth of information, so that powerful ML models can nevertheless be trained on them, as prominently demonstrated by the recent success of the AlphaFold2 model trained on the PDB.[6]

Efforts like the PDB are thus immensely valuable. Unfortunately, they are hard to reproduce in other fields. They depend on long-term funding and the collective contributions from an entire scientific community over several decades. Experimental databases in other fields are therefore usually much smaller, often containing only tens or hundreds of datapoints, if they exist at all. This is not necessarily due to a lack of reported experiments in principle but rather due to the difficulty of extracting the results from the literature and, crucially, due to the inconsistency of experimental results obtained in different labs or with different techniques.[26] On top of this, publication bias is a real problem in some cases, *e.g.* when only active catalyst materials are published, while ‘failed’ experiments remain unreported.[27] To address pressing chemical questions such as the prediction of catalytic activities or solvation effects with ML, we thus cannot wait for a project equivalent to the PDB to materialize in the respective fields.

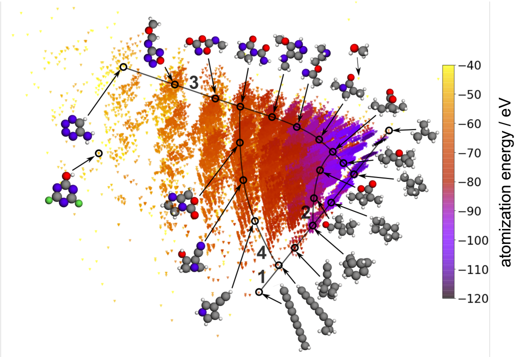


Figure 1. Visualization of the QM9 dataset of small organic molecules using the kernel Pricipal Component Analysis method. Each point represents a molecule and the distance between points indicates their structural similarity. This is emphasized by the paths in the figure, highlighting systematic structural changes. The colormap reveals that electronic properties like the atomization energy vary smoothly across chemical space. Adapted with permission from Ref. [39]. Copyright 2020 American Chemical Society.

Since the mid-2000s, the increased availability of computational resources and first-principles electronic structure methods (in particular DFT) has changed this situation somewhat, by enabling the creation of computational databases. Prominent examples of this include the Materials Project,[28] AFLOW,[29] the Materials Cloud,[30] the Open Quantum Materials Database,[31] and NOMAD,[32] all of which mostly focus on the properties of bulk solids. Here, the Materials Project is a particularly insightful example as it expands an experimental database (the ICSD) with additional materials and properties. The relative ease of running a DFT calculation means that this can be achieved much more quickly and at a fraction of the cost of the corresponding experiments. Similarly, a series of computational molecular databases have been developed, initially focusing on the small organic molecules of the GDB-17 universe.[33] The QM9 database containing ground state properties of over 100k molecules was a pioneering achievement in this context (see Fig. 1).[34] The landscape of molecular databases has since been expanded further by including non-equilibrium configurations and conformers,[35, 36] ionized states[37], and radicals[38].

To summarize this bird’s-eye view of atomistic data, we can say that in spite of chemist’s supposed fondness of stamp collecting, there are only a few experimental databases that can reasonably be called ‘big data’. This is because large, community-wide efforts are required to generate them. In most fields, high quality experimental databases are thus decidedly ‘small data’. Computational databases based on high-throughput electronic structure calculations provide an important alternative in this context, as they can be generated much more cheaply. Even here, full coverage of chemical space cannot be expected, however.

10.1002/anie.202219170

III. DATA-DRIVEN AND SCIENCE-DRIVEN MACHINE LEARNING

It is clear that the landscape of available databases just described impacts the direction that research in atomistic ML is taking. In particular, computational databases like QM9[34] have been instrumental in the development of atomistic ML models. This reflects the highly competitive nature of ML research, where exceeding state-of-the-art performance on a well defined dataset and task is one of the main goals when developing new methods. This kind of ML research is thus in a literal sense data-driven, *i.e.* the available datasets and associated tasks determine how new methods are designed. This has undeniably led to significant progress, but it also represents a rather artificial setting compared to real chemical research. Most importantly, for the reasons outlined above it is usually not the case that a well-curated dataset exists that can be used to develop an ML model.

Approaches that are optimized in a data-driven setting are therefore of limited use for answering questions like: “What is the structure of an interface between two materials?” or “What is the free energy barrier for a particular heterogeneous catalytic reaction?”. The available sources of big data contain little to no information about these questions. While it would in principle be possible to address this lack by generating new extensive databases dedicated to a certain material class or target property (as was recently done with the OC20 database focusing on heterogeneous catalysis)[40], this requires massive investments of time and money. Furthermore, it is not trivial to predict what size and shape such a database should have, in order to cover the target domain in a satisfactory manner.

Fortunately, there is an alternative to the data-driven approach, which I will term “science-driven” in the following (see Fig. 2). The key feature of science-driven ML is that it begins with a scientific question and then asks what training data and model design choices are appropriate. This is particularly important when scientific questions are not reducible to a simple metric like the mean absolute error (MAE) with respect to some predefined test set. Indeed, this is a common situation in atomistic ML, where training data is usually generated by first-principles electronic structure calculations (*e.g.* of total energies and forces), while the property of interest is often a macroscopic observable, such as a reaction rate, melting point or diffusion coefficient at finite temperature.[8, 38] Accurately predicting energies and forces for a fixed test set does not guarantee that the observable is accurately predicted, since the dynamic simulation required to predict the observable can lead far away from the configurations in the test set. Furthermore, it is not trivial to determine *a priori* how errors on energies and forces translate to errors on the desired observables. In other words, even if the test set provides an accurate estimate of how the model performs for unseen data, it is unclear how low the corresponding test error should be for any given application.

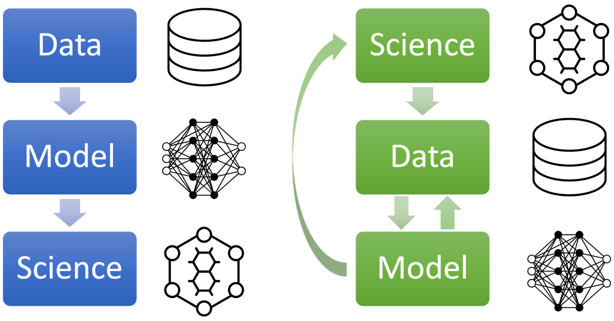


Figure 2. Schematic illustration of data-driven and science-driven machine-learning approaches. In the science-driven approach, data collection is driven by specific scientific questions and a feedback between model fitting and data collection is implemented.

Having established the scope of science-driven ML, what are the main challenges towards developing such models? First, since predefined training sets are not available, adaptive algorithms for data generation and high data-efficiency of the models are required. Second, robust extrapolative capabilities are essential, since the configurational space of interest is also not known at the outset. Finally, it would be benefitial to gain insights into how errors propagate from atomistic predictions to macroscopic observables. In the remainder of this manuscript, I will discuss how these requirements can be achieved in practice, with the help of active learning, physical priors and uncertainty estimation.

Note that the concepts discussed herein are largely agnostic towards the technical details of the ML models themselves (*e.g.* regarding neural networks *vs.* Kernel methods). For in-depth discussions of different methodological approaches to atomistic ML, the reader is referred to several recent review articles.[1, 2, 41–46]

IV. ACTIVE AND ITERATIVE LEARNING

The development of ML interatomic potentials almost immediately revealed the limitations of a purely data-driven approach in chemistry. Indeed, it is almost impossible to generate a training database that adequately covers the phase space of any reasonably complex molecule or material *a priori.*[47] Even for the relatively simple case of non-reactive closed-shell organic molecules in the gas-phase, this requires a sufficiently representative set of molecules, an extensive exploration of the configuration space of each molecule and a set of non-equilibrium configurations for each conformer (*e.g.* from molecular dynamics or normal mode displacements).

To put this into perspective, the QM7-x database of Hoja et al. provides over 4 million configurations that fulfill these criteria for organic molecules with up to seven heavy atoms.[36] While this affords sufficient coverage to

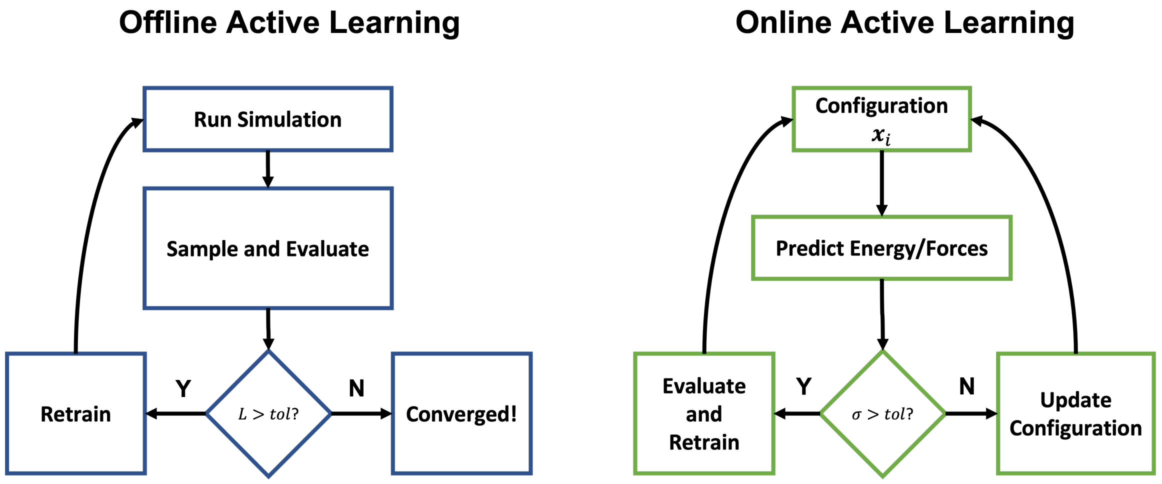


Figure 3. Illustration of offline and online active learning. Left: In the offline approach, the simulation of interest is performed with an ML model, generating new configurations. By evaluating samples of these, the accuracy of the ML model can be checked. If the obtained error L is above a tolerance criterion, new configurations are added to the training set and the simulation is repeated. Right: In the online approach, the model itself provides an uncertainty measure for each prediction made, *e.g.*, during a molecular dynamics or Monte Carlo simulation. If the predictive uncertainty σ is above a tolerance factor, a reference calculation is performed and the model is retrained based on this new information.

train robust interatomic potentials for small gas-phase molecules, such potentials will not be able to extrapolate to condensed systems (*e.g.* molecular liquids or crystals), macromolecules (*e.g.* proteins or polymers) or chemical reactions. Furthermore, data requirements tend to rise non-linearly with the number of elements in the system due to the curse of dimensionality. Generating a dataset with similar coverage as QM7-x for condensed phase systems, inorganic materials or biomolecules in solution would consequently require a staggering computational effort.

A hallmark of most science-driven ML approaches is therefore that data collection and model construction are not decoupled from each other. Instead, multiple models are fitted in an iterative fashion so that the training set is expanded at each step, based on the predictions of the current model. Because the model itself influences the training set, this is often termed *active learning* (AL) or *iterative training* (see Fig. 3).

The key ingredient of any AL approach is a criterion according to which new datapoints are selected. Here, the most common choices either leverage data diversity or predictive uncertainties. In the former case, a measure for similarity between datapoints is used to ensure that new configurations added to the training set are as dissimilar as possible to the already known configurations.[48] In the latter case, the ML model provides a measure of uncertainty along with each prediction. This way, highly uncertain predictions can be checked with accurate reference calculations and subsequently added to the training set. These uncertainty estimates are often obtained by fitting ensembles of models with different weight initializations and/or different subsamples of the training set.[49] Alternatively, Bayesian ML methods like Gaussian Process Regression directly provide predictive uncertainties.[2, 50]

While the AL concept is rather simple, it is not necessarily trivial to implement in practice. In particular, uncertainty measures must be well calibrated in order to provide reliable and useful error estimates.[49, 50] There is also a question of resolution: the predicted uncertainty on the atomization energy of a large molecule may be small, even if a certain functional group is poorly described by the model. For both uncertainty and diversity-driven workflows it can therefore be appropriate to use per-atom rather than per-configuration estimates, depending on the application.[51–53]

The most common use-case for AL in chemistry is the development of interatomic potentials. Here, a preliminary potential can be used to run exploratory simulations (often dynamics or structure searches), generating novel configurations. These can in turn be evaluated with first-principles calculations and added to the training set.

Since it is impossible to know *a priori* what configurations are required for fitting an interatomic potential, AL has always been used for in this context, though not necessarily in fully automated workflows. This is sometimes termed "offline" AL, since the simulation, data selection and training are performed in separate steps (see Fig. 3).[54] More recently, several groups have also shown that "online" active learning can be used in some cases, *e.g.* for molecular dynamics (MD) simulations.[51, 55–57]

As an example, an online AL potential for Methylammonium Lead Iodide (MAPI) by Jinnouchi et al. is shown in Fig. 4.[51] It can be seen that the uncertainty estimate in this plot correlates well with the real error of the potential, so that first-principles calculations can be invested effectively. As a consequence, 99% of the calculations

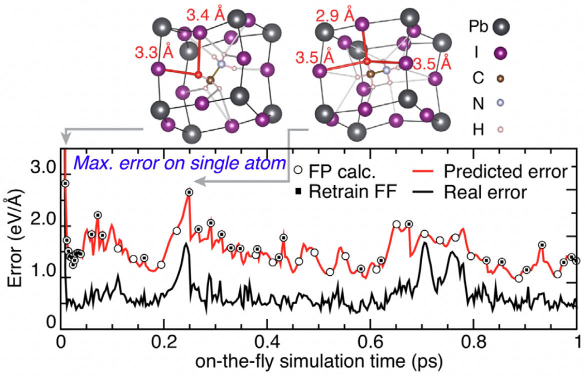


Figure 4. Online active learning molecular dynamics for a Methylammonium Lead Halide perovskite. Shown are the real (black) and estimated (red) errors of a machine learned force field (FF). Large estimated errors trigger first principles (FP) calculations, which are used to retrain the FF. The structures on top highlight the hydrogen atom with the highest error in red, for two snapshots. Reprinted with permission from Ref. [51]. Copyright 2019 by the American Physical Society.

necessary for the corresponding ab initio MD trajectory are saved, enabling the application of the potential to study complex phase transitions.

Similarly, offline active learning was recently shown to drastically increase the efficiency of global structure searches for large molecular adsorbates on transition metal surfaces.[58] Due to the conformational flexibility of these molecules, this is a complex global optimization problem. In an AL workflow, interatomic potentials were used to run extensive Minima Hopping simulations to explore this configuration space. By retraining the models on the thus generated configurations, high accuracy and data-efficiency could be achieved. The converged potentials were then used for extensive structure searches for a series of molecules and fragments on different Rh surfaces. This revealed that the stability of some adsorbates was previously underestimated by more than 1 eV, with significant implications for catalysis.

In this context, interatomic potentials are a special case as they can directly be used to generate new configurations (*e.g.* through simulations). This is different for more general regression or classification models, *e.g.* when predicting electronic properties like reorganization energies[59] or biological activities[60]. Here, a similar concept can nonetheless be applied, when a large pool of unlabeled, potentially interesting systems is available. An uncertainty or diversity measure can be used to efficiently draw samples from this pool, again iteratively expanding the training set. This strategy was, *e.g.*, used for the ANI-1x potential, leading to higher accuracy with a fraction of the data used for its predecessor.[35] A similar concept is used in the self-correcting ML of Dral and coworkers.[61]

Another important application of AL and related techniques (namely Bayesian Optimization) is for optimization tasks.[62, 63] These include the already mentioned structure searches (*i.e.* finding the most stable geometry of a system)[52, 62–66] but also more general molecular or materials design tasks (*i.e.* finding a compound with desired properties)[59, 60, 67, 68]. Here, the goal is not just to add diverse configurations to the training set but also to guide the optimizer towards favourable configurations or systems. Data selection is thus governed by a acquisition function that balances exploration (as quantified by uncertainty or diversity measures) and exploitation (as quantified by favourable predicted properties of a candidate).

V. INDUCTIVE BIASES AND PHYSICAL PRIORS

It is often claimed that ML models merely interpolate the training data. While this is true in some sense (though not strictly speaking, as shown by Zeni *et al.*[69]), it is also vastly oversimplifying. Depending how an ML model is set up, it will perform predictions on unknown data (induction) in vastly different ways. As a consequence, we can influence how robustly an ML model will extrapolate beyond the current training set by making the right design choices. As discussed in the previous section, AL frameworks use coarse initial models to generate or select training configurations. It is therefore of particular importance to use models that work well in low-data regimes in this context.

In the ML literature, the set of assumptions that determines how an algorithm performs predictions is collectively termed the *inductive bias* of a model. This is illustrated in Fig. 5. There is usually a space of possible ML models that can fit a given training set equally well. However, since each model has its own inductive bias, their predictions for unseen data will in general be different. This variation is particularly large when little data is available.

In more concrete terms, inductive biases can be related to how input features are passed to the model (*e.g.* sequentially or all at once), how they are processed (*e.g.* taking spatial locality into account) and how the output is produced (*e.g.* respecting permutational invariance of the input features). All of these choices influence the predictions of the resulting models and different applications call for different model architectures. For example, computer vision models benefit from other inductive biases than natural language processing models.

Over the last decade several powerful inductive biases for atomistic machine learning have been found. Perhaps the most fundamental of these relate to mathematical invariances that most chemical properties (particularly the total energy of a system) fulfill. Specifically, these are invariance to permutations of atoms of the same element, as well as global rotations and translations of a system.[41, 42] By rigorously enforcing these invariances when building representations of chemical structures, all

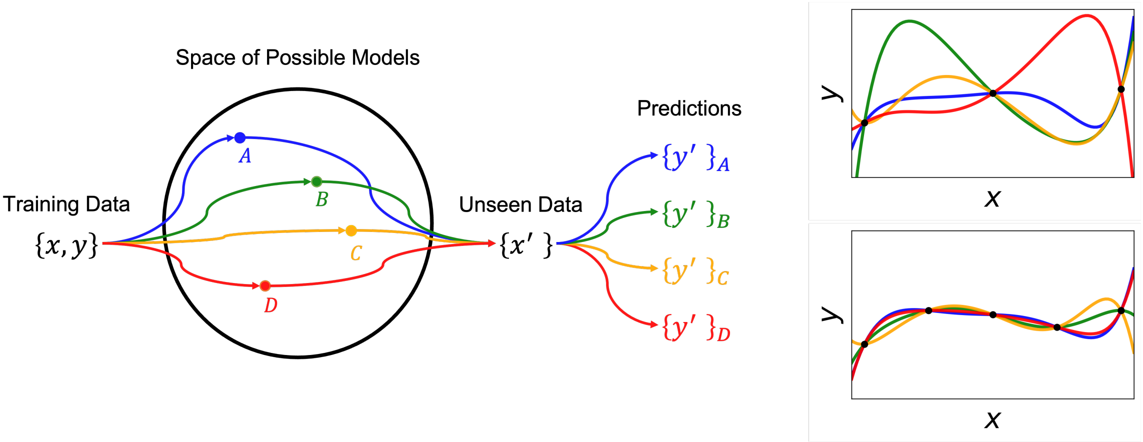


Figure 5. Illustration of inductive bias. Left: There is a space of possible machine learning models that can fit the training data similarly well. However, each of these models will make different predictions for unknown data. Right: As more data is added (from top to bottom), the variation of different models that perfectly fit the training data decreases. Consequently, inductive biases are particularly important in the 'small data' regime.

subsequent ML models automatically fulfill them. As a consequence, they do not need to be learned from data, making the corresponding models more data-efficient.

While invariance is thus a key property, it has recently been found that important structural information can be lost in the process of making representations rotationally invariant. In particular, degeneracies or near-degeneracies can occur, meaning that different structures (with different properties) are mapped to the same representation.[70–74] This is highly problematic for ML models, which obviously cannot assign different outputs to identical inputs. Furthermore, it is clear that not all molecular properties are invariant to rotations. Instead, tensorial properties like (hyper-)polarizabilities or dipole vectors are equivariant, meaning that they rotate with the molecule.

To address this, a series of equivariant ML models have been proposed in recent years.[75–77] These are invaluable for the rigorous prediction of properties ranging from dipole moments to full electron densities.[78–81] Perhaps surprisingly, equivariant neural networks can also display significant advantages when predicting invariant properties like the potential energy of a system.[75, 82] In this case, the models are internally equivariant until the last layer of the network, where an invariant output is produced. This way, the loss of structural information that plagues some invariant representations is avoided.

Alternatively, it has also been found that higher-order invariant representations can be generated efficiently using the atomic cluster expansion (ACE).[83, 84] This approach is closely related to the classic cluster and many-body expansion methods used in chemistry and materials modelling.[85] Importantly, the ACE invariants form a systematically convergent basis so that full structural information can be retained in an invariant representation.[86] Indeed, due to its completeness and efficiency, ACE allows the development of highly accurate interatomic potentials using (regularized) linear regression, providing an ideal trade-off between accuracy and speed in many cases.[84, 87] In a similar vein, models like UF3 or ChiMES use explicit body-order expansions of the energy in terms of products of two-body functions.[88, 89]

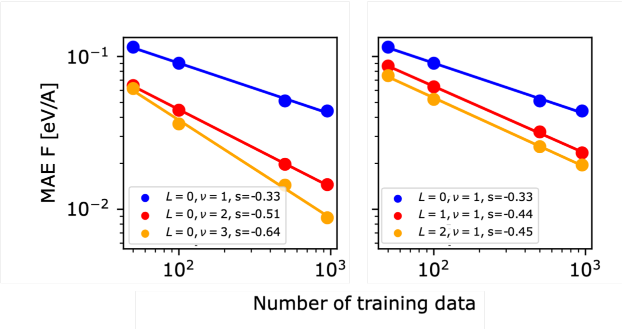


Figure 6. Mean absolute error (MAE) of force predictions for Aspirin configurations with MACE potentials, as a function of the training set size. Left: Without equivariance ($L = 0$), the data efficiency can be increased by increasing the body-order of the potential (ν). Right: For a fixed body-order, introducing equivariance ($L > 0$) also increases the data-efficiency. In both cases, the slopes (s) of the learning curves increase with better inductive biases. Figure adapted from Ref. [82] with permission.

To illustrate the benefits of inductive biases such as equivariance and high body-order for science-driven ML, it is instructive to consider the MACE approach of Batatia et al.[82] In MACE, both the equivariance and the body-order of the model can be controlled via hyperparameters. As shown in Fig. 6, both of these factors lead to improved predictive accuracy. Interestingly, this is not merely reflected in a consistently lower MAE, but in a steeper slope of the learning curves, indicating that equivariant and high body-order models learn more effectively from the data.

It should be noted that an alternative approach to predicting tensorial properties is to take Cartesian derivatives of invariant model outputs. Indeed, this is the standard approach for predicting force vectors, which are rigorously defined as energy derivatives. More recently, this idea has been generalized to predicting dipole moments, coupling vectors and electronic friction tensors.[90–92]

The above considerations relate to how structural information is received and transformed by the model. Inductive biases can also be related to the fitting target. This can be as simple as choosing the adequate scale when fitting energies. By predicting atomization energies and requiring that the energies of isolated atoms are strictly zero, a model will automatically predict bond energies that are on the correct order of magnitude, even when no bond-breaking events are in the training set.[2] Similarly, size-extensivity of ML models can be enforced by adequately normalizing the representation and fitting target.[93] This also has important consequences for predicting reaction energies in complex reaction networks.[38] Here, fitting energies per atom is beneficial since small but chemically important molecules like H_2 and CO are described less accurately otherwise.

A particularly powerful type of inductive bias is the explicit inclusion of physical priors. This is often achieved via the Δ-ML approach,[94] where the predictions of a computationally efficient physical model (often a semiempirical method[95–97]) are used as a baseline. The ML model then merely predicts the difference between the target method and the baseline, which can dramatically decrease the amount of data required to achieve a given accuracy.[98–100] The inclusion of an explicit physical baseline furthermore often ensures better transferability of the model.[101, 102]

An additional advantage of Δ-ML is that it enables the inclusion of effects at the baseline level, which cannot be described by the ML model at all. A prominent example of this are long-range electrostatic and dispersion interactions, which are missing in many common ML models based on local atomic environments.[102, 103] In the context of dispersion interactions, a related idea is to train a short-ranged model on (long-range) dispersion-free DFT data. These effects can then be treated separately via physical van-der-Waals corrections. The latter may in turn also be coupled to ML models predicting charges or Hirshfeld volumes.[104, 105]

Such Δ-ML models were used in Ref. 103 to predict the structures of organic molecular crystals. Here, a dispersion corrected density functional tight-binding (DFTB) baseline was combined with a local ML correction[106, 107]. While the baseline alone was not sufficiently accurate to reliably rank potential crystal polymorphs (or predict their structures), it did provide a reasonable prior for the relevant inter- and intramolecular interactions. Meanwhile, a pure local ML model would also be inadequate here, since relative crystal stabilities are known to depend on long-range interactions. Combining DFTB and ML, highly accurate and data-efficient models could be obtained.

As an alternative to Δ-ML, semiempirical models can also be used to generate more powerful, physics-based input features for ML models. An example of this are the OrbNet models, which use semiempirical electronic properties to this end.[108, 109] Such features were also found to be advantageous when predicting molecular reorganization energies, which do not depend on the ground state structure alone.[100]

Finally, the arguably most sophisticated way to use physical priors is to build so-called physics-enhanced ML models. In this case, an ML model is intimately connected to a physical model. Indeed, it would be equally valid to talk about ML-enhanced physical models. This is a highly active field of research which spans ML-predicted Hamiltonians[110–112], semiempirical models with environment-dependent parameters[113] and machine-learned quantum chemical methods[114, 115].

In this context, the electron density plays a central role. Several groups have reported models for predicting electron densities in materials and molecules.[80, 116, 117] This is of great interest since a range of important electronic properties (such as multipole moments or molecular electrostatic potentials) can be obtained directly from the density. Furthermore, accurate learned densities can accelerate the convergence of DFT calculations or avoid self-consistency loops altogether.[116, 118] A key inductive bias for such models is how the electron density is represented (see Fig. 7).

Arguably, the most straightforward solution is to use real space grids, which are already implemented in most DFT codes. The feasibility of this was demonstrated in Ref. [117] for uniform grids. Such grids are only efficiently applicable for valence electron densities of dense, condensed phase systems, however. Generalization of density prediction to the non-uniform grids used in all-electron DFT codes with open boundary conditions is not straightforward. Furthermore, real space grids have enormous memory demands for large systems. In a seminal paper, Brockherde et al. showed that the density can instead be predicted efficiently in a plane-wave basis.[116] Here, the orthogonality of the basis is mathematically convenient, as a separate ML model can be fitted for each Fourier component. On the flipside, this restricts the prediction to fixed unit-cells and relatively rigid systems.

In Ref. [80] these limitations were overcome by representing the electron density in terms of atom-centered basis functions. A key feature of this approach is that it naturally decomposes the density into atomic contributions. This enables highly transferable and equivariant density predictions, where models can be trained on small systems and applied to large ones. This advance was subsequently used for density prediction in large noncovalently bonded systems[120] and periodic cells[81].

Even more physics can be introduced into density-

10.1002/anie.202219170

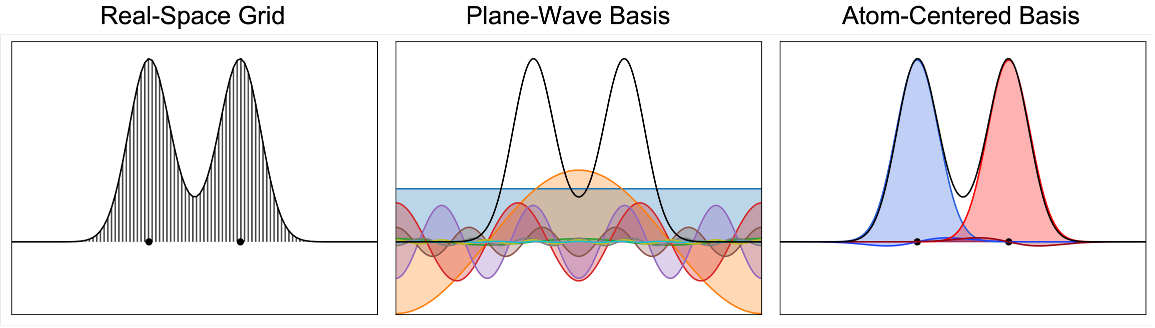


Figure 7. Representations of electron densities in machine learning, illustrated for a one-dimensional hydrogen molecule. Left: The most straightforward approach is to map the density onto a real space grid. Center: A plane-wave basis is more compact and mathematically convenient due to the orthogonality of the basis functions. Right: An atom-centered basis is highly compact and allows decomposition of the density into atomic contributions, which enables scaling to large systems. Note that all representations yield the same density using 100, 10 and 4 basis functions, respectively. Figure adapted with permission from Ref. [119].

based ML models by using the variational principle as an inductive bias. This means that the electron density is predicted by minimizing an ML-based energy functional. In this case the model is a full-blown density functional approximation, meaning that it offers a route to all first-order properties, energies, and forces on an equal footing. Here, flexible ML models can overcome the well-known self-interaction problems of conventional semilocal functionals.[121] In this context, the main focus has been on machine-learned exchange-correlation functionals, and to a lesser degree on kinetic energy functionals for orbital-free DFT.

A key question for ML DFT is how non-locality can be introduced into the functionals. One alternative is to use the same non-local ingredients also used in conventional DFT, such as the kinetic energy density and the exact exchange density. This was exploited in the recent DM21 functional.[122] While DM21 shows impressive performance on a wide variety of benchmarks, Becke subsequently showed that physics-based functionals using the same ingredients can be equally or more accurate.[123, 124] More critically, the local exchange density is computationally involved to calculate, so that these functionals are much less widely applicable than conventional ones. Much interest has therefore been devoted to the development of non-local ML functionals that depend on the electron density alone (*pure* density functionals).

Here, one strategy is to use convolutions, so that the local exchange-correlation energy density on any given point depends on the electron densities at nearby grid points.[125–128] Bystrom and Kozinsky showed that these convolutions can be designed to obey scaling constraints, a critical step towards more rigor in ML DFT.[129] Indeed, a substantial advantage of the convolutional approach is that it can be formulated as a straightforward generalization of semilocal functionals, so that the same exact constraints can be enforced. It should also be noted, however, that performing the convolutions represents substantial computational overhead.

As with density prediction, moving away from the grid-based representation has some advantages in this context. In both plane-wave and atom-centered basis sets, non-locality is automatically included. Bogojeski et al. showed that highly accurate non-local functionals can be fitted in a plane wave basis.[118] This completely avoids numerical quadrature on a grid and thus has the potential to be more efficient than conventional DFT, especially when combined with density prediction. Unfortunately, the resulting functionals are not size-extensive, however.

Here, the use of atom-centered basis functions again offers a promising route. Dick and Fernandez-Serra showed that density projections could be used to create non-local, atom-centered density representations for size-extensive exchange-correlation functionals, using the semilocal PBE functional as a baseline.[130] Similarly, Margraf and Reuter used density-fitting to obtain pure, non-local and size-extensive correlation functionals that achieve energy errors below 1 kcal/mol with less than 100 training samples.[119]

A key advantage of using such atom-centered representations is that the corresponding models naturally scale to large systems and are computationally highly efficient. A downside compared to grid based methods is that they are not as transferable across the periodic table, since the basis functions are to an extent element specific. Furthermore, these functionals currently do not respect any exact constraints.

A big advantage of the physics-enhanced approach in general is that it leads to much higher interpretability of the predictions. A pure ML model may be able to accurately predict dipole moments, but a physics-enhanced model additionally allows understanding these dipole moments in terms of more fine-grained charge distributions. Another advantage is that physics-enhanced models often predict more fundamental quantities that allow predicting multiple molecular properties on an equal footing. As an example, predicted electron densities give access to multipole moments, electrostatic and exchange-correlation energies, as well as topological properties like partial charges and bond orders. With the appropriate physical model, it is even possible to learn properties indirectly (*e.g.* electron densities from energies[131]). Finally, as with Δ-ML, physics-enhanced models tend to be highly data-efficient since they incorporate strong priors.

On the flip-side, including physical priors in this manner usually leads to lower computational efficiency when performing induction. Learned DFT functionals can surpass the accuracy of conventional approximations in many cases, but they are usually equally or more expensive to evaluate. In contrast, pure ML models are typically several orders of magnitude faster than DFT calculations. In this context, the scientific question of interest must decide which approach is best suited. Fortunately, this is not a binary question between highly data-efficient and interpretable models on one hand to ultra-fast black-box models on the other.

As an example of an intermediate approach between these extremes, it can be noted that it is not necessary to know the full details of the electron density in order to adequately describe long-range Coulomb interactions. Artrith et al. showed that short-ranged ML potential can instead be combined with learned atomic partial charges.[90, 132, 133] However, when these charges are directly predicted by an ML model, non-local charge transfer effects or different total charge states cannot be described.

By invoking the variational principle as an inductive bias, Goedecker and co-workers showed that ML-based charge equilibration models can overcome this limitation.[134] This idea has since been further developed, *e.g.* in the fourth generation neural network potentials of Behler and co-workers[135] and our recently proposed kernel charge equilibration method[136]. It should also be noted that ML models using global descriptors provide a different path towards including non-local effects, by explicitly correlating all atomic degrees of freedom.[137]

VI. MODEL EVALUATION AND ERROR PROPAGATION

The previous sections described how science-driven ML approaches can take advantage of active learning, inductive biases, and physical priors to overcome the need for large predefined databases. Up to this point we tacitly assumed that there is a clear performance metric with respect to which the ML models should be optimized. In some cases defining this metric is fairly straightforward: Molecular or materials design requires an accurate prediction of the target property for the candidates of interest. This can be quantified in terms of a MAE on an unseen test set, provided that the test set is representative of the full design space (a non-trivial caveat).

In other cases, the ML model only addresses the main scientific question in an indirect way, however.[138] A prime example of this are interatomic potentials. From an ML perspective these are simply regression models fitted to energies and forces. From a scientific perspective, the energies and forces are not really of much interest. Instead, the interatomic potential is a tool used to propagate atomic coordinates, *e.g.* in MD or Monte Carlo simulations. The observable of interest can then be derived from these simulations in the form of an average density, a melting point, a diffusion constant, or a free energy difference. This raises the question how the force MAE of the potential relates to this observable.

This question may appear somewhat academic, since one could argue that as long as the predicted PES matches the target one (as quantified by the MAE), all derived quantities should also match. However, it turns out that the force MAE on a test set is not even a good predictor for the general force error of an interatomic potential. In Ref. [139], a series of interatomic potentials were fitted to subsets of the QM7-x database.[36] Graph neural networks based on the recent GEMNet architecture[140] displayed the best performance in this context, with force errors below 0.005 eV/Å for the test set when training on 3.2 million configurations. Interestingly, even the smallest training set used (3.2 thousand configurations) yielded quite low force errors, on average below 0.05 eV/Å. However, when running MD simulations with these potentials and reevaluating the obtained configurations with DFT, the observed error was found to be several orders of magnitude larger for the models trained on 32,000 configurations or less (see Fig. 8).

The problem here is that these potentials have no information about unphysical regions of the potential energy surface. If the trajectory leaves the scope of the training set (which is unavoidable in high dimensional PESs), such unphysical configurations (*e.g.* doubly coordinated hydrogen atoms in organic molecules) may erroneously be assigned low energies. At this point, the simulation becomes stuck in an unphysical region of the PES and the trajectory is useless. Importantly, this may only become apparent when performing rather long MD simulations (on the order of nanoseconds), as shown in Fig. 8.

The ML potential trained on 3.2 million configurations extrapolates quite robustly in this test, indicating that these pathologies can to some extent be avoided with enough data (or better yet, with improved data selection using active learning). The point is, however, that the only way to reliably evaluate the suitability of an atomistic ML model is by running real simulations with it. From this perspective, the common practice of merely reporting improvements on static benchmark databases should be questioned. This is another advantage of the offline active learning approach described above, since it includes atomistic simulations in the model fitting process by construction.

Once a robust interatomic potential has been obtained, the question how the force MAE translates to uncertainty

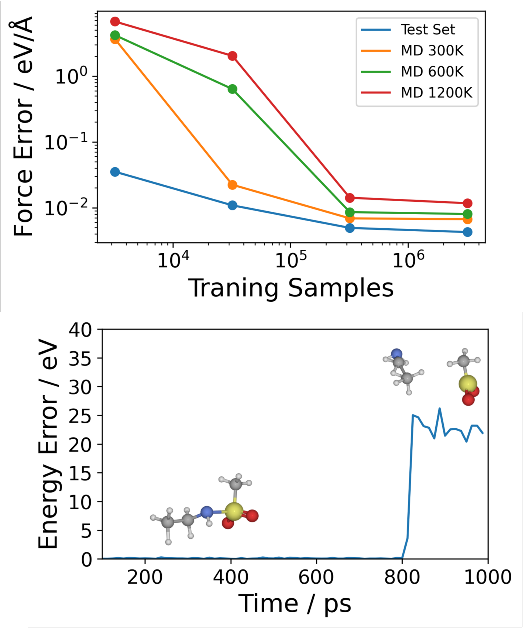


Figure 8. Robustness of Learned Potentials. Top: The force error on a static test set can be orders of magnitude lower than the error observed during long and hot molecular dynamics (MD) trajectories. Bottom: The large MD errors are only observed when running the simulation sufficiently long, since they stem from unphysical behaviour for particular regions of the potential energy surface. Figure adapted with permission from Ref. [139].

in the predicted observable can be addressed. Beyond the intrinsic usefulness of such an uncertainty estimate, this is important in order to determine how accurate the underlying potential needs to be to achieve the desired precision on the target observable. Here, Imbalzano et al. have shown that ensemble based uncertainty estimates can be propagated through MD simulations.[141] This is show in Fig. 9 for the example of predicting the melting point of hexagonal ice. While this type of calculation is not yet common practice in the field, this approach holds great promise for increasing the rigor of ML based predictions.

An important aspect that has not been addressed up to this point is that an ML model can only ever be as good as the reference data it is trained on. In atomistic ML this data most often stems from DFT calculations, which do not yield experimental accuracy in many cases. Here, the development of data-efficient ML approaches and the increasing availability of high-level quantum chemistry methods for large and even periodic systems present an opportunity to exceed DFT accuracy in complex atomistic simulations. Examples of this include the prediction of surface adsorbate coverages[56], the properties of liquid water[142] and crystal structure prediction[103].

An alternative route to overcome the limitations of DFT references is to incorporated experimental data into the training process, *e.g.* by biasing simulations towards known macroscopic properties.[143] Indeed, this concept is already well established in the domain of classical MD simulations. For example, the non-bonding parameters in the OPLS force field were fitted to structural and thermodynamic properties of liquids.[144] More recently, minimal biasing methods were developed which modify existing potentials to reproduce experimental data.[145] Experimental information can also be incorporated at a non-atomistic scale, *e.g.* in coarse grained potentials[146], augmented Markov Models[147] or microkinetic models of catalytic processes[148]. Many of these methods could in principle be directly applied to ML-based simulations.

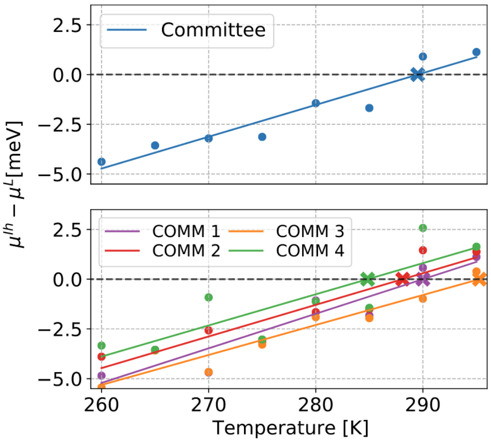


Figure 9. Propagating the uncertainty of learned potentials to physical observables. Top: The melting point of ice is estimated by computing the point of equal chemical potential between hexagonal ice and liquid water, using a committee (ensemble) of machine learning (ML) potentials. Bottom: By interrogating the individual members of the committee, the uncertainty of this estimate can be obtained. Note that this only captures the uncertainty due to the ML potentials, whereas functional and sampling errors are not included. Reprinted with permission from Ref. [141]. Copyright 2021 by the American Institute of Physics.

VII. SUMMARY AND OUTLOOK

In this review I have argued that the sparsity of large, curated databases precludes the use of purely data-driven ML in many areas of chemistry. In contrast to this, science-driven ML approaches can be used to answer concrete scientific questions, even in the absence of pre-existing databases. To this end, active and iterative learning schemes are leveraged and data-efficiency of the underlying ML models is an important requirement. Furthermore, the use of physical priors is often helpful since it improves the extrapolative capabilities of the models and reduces the need for large amounts of training data. These hallmarks of science-driven ML have some important implications for method development at the interface of chemistry and ML.

First, iterative training workflows depend on the capability to (re-)train a model many times on small to mid-sized datasets, whereas the typical data-driven model is trained only once on a very big dataset. In the latter case, investing weeks to train a single model is possible, but for an active learning protocol this is prohibitive. The move to ever larger deep learning models that is observed in many ML applications is potentially a worrying development in this context.[14] Similarly, the selection of appropriate hyperparameters for a model can be problematic as the training set is continuously changed. In particular, common techniques like cross-validation are not robust in early iterations, when the training set is extremely small. Here, reliable heuristics or defaults are necessary.[58]

Second, semiempirical models are currently experiencing a surprising revival, just when it seemed they would become irrelevant with the rise of ML potentials. On one hand, this is because they are invaluable for cheap exploratory structure searches for complex molecules and materials.[95, 149] On the other hand, they are also highly useful for describing long-range interactions to complement short-ranged ML potentials, providing baselines for Δ-ML or computing inexpensive electronic structure features for ML models.[97, 103, 109] Transfer and multi-fidelity learning approaches can also be used to increase the accuracy on high-level targets by (pre-)training on lower-level reference data.[150]

Third, the quality of a science-driven ML model should mainly be assessed by how well it performs its task, not necessarily by how well it fits some particular dataset. For example, a reasonably low test set error on atomic forces is a necessary but insufficient condition for accurately predicting macroscopic observables with ML-based MD simulations. Method developers should therefore take into account the wider context of where the proposed methods are supposed to be applied. A neural network predicting energies and forces for an atomistic system is not just another regression model, it is an interatomic potential. It should therefore also be tested in a realistic use case for interatomic potentials, such as a (sufficiently long) MD simulation.

Fourth, the use of error and uncertainty estimation is still somewhat underdeveloped in the field, although the corresponding methodology is now quite mature. Beyond the quantification of uncertainty due to the ML fit, the incorporation of experimental or high-level quantum chemical data represents the next step towards quantitative predictions with science-driven ML methods. Interestingly, ML potentials also play a central role in quantifying the accuracy of electronic structure methods. In ab initio MD studies, it is usually not possible to disentangle basis set incompleteness, finite size and statistical sampling errors. By training ML potentials, these can often be overcome, leaving an unobstructed view of the real functional error.

[1] J. Behler, *Chem. Rev.* **2021**, *121*, 10037–10072.

[2] V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, G. Csányi, *Chem. Rev.* **2021**, *121*, 10073–10141.

[3] A. Glielmo, B. E. Husic, A. Rodriguez, C. Clementi, F. Noé, A. Laio, *Chem. Rev.* **2021**, *121*, 9722–9758.

[4] M. Staszak, K. Staszak, K. Wieszczycka, A. Bajek, K. Roszkowski, B. Tylkowski, *WIREs Comput Mol Sci* **2022**, *12*, 1.

[5] S. Kim, J. Noh, G. H. Gu, A. Aspuru-guzik, Y. Jung, *ACS Cent. Sci.* **2020**, *6*, 1412–1420.

[6] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, *Nature* **2021**, *596*, 583–589.

[7] T. Morawietz, A. Singraber, C. Dellago, J. Behler, *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113*, 8368–8373.

[8] V. Kapil, C. Schran, A. Zen, J. Chen, C. J. Pickard, A. Michaelides, *Nature* **2022**, *609*, 512–516.

[9] V. L. Deringer, N. Bernstein, G. Csányi, C. Ben mahmoud, M. Ceriotti, M. Wilson, D. A. Drabold, S. R. Elliott, *Nature* **2021**, *589*, 59–64.

[10] B. Cheng, G. Mazzola, C. J. Pickard, M. Ceriotti, *Nature* **2020**, *585*, 217–220.

[11] B. P. Macleod, F. G. L. Parlane, T. D. Morrissey, F. Häse, L. M. Roch, K. E. Dettelbach, R. Moreira, L. P. E. Yunker, M. B. Rooney, J. R. Deeth, V. Lai, G. J. Ng, H. Situ, R. H. Zhang, M. S. Elliott, T. H. Haley, D. J. Dvorak, A. Aspuru-guzik, J. E. Hein, C. P. Berlinguette, *Sci. Adv.* **2020**, *6*,.

[12] L. Yao, Z. Ou, B. Luo, C. Xu, Q. Chen, *ACS Cent. Sci.* **2020**, *6*, 1421–1430.

[13] A. Bender, N. Schneider, M. Segler, W. Patrick walters, O. Engkvist, T. Rodrigues, *Nat Rev Chem* **2022**, *6*, 428–442.

[14] D. Probst, *ChemRxiv:10.26434/chemrxiv-2022-z6s5m* **2022**.

[15] R. Luckenbach, *Chem. Unserer Zeit* **1981**, *15*, 47–51.

[16] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, E. E. Bolton, *Nucleic Acids Res.* **2019**, *47*, D1102–D1109.

[17] H. E. Pence, A. Williams, *J. Chem. Educ.* **2010**, *87*, 1123–1124.

[18] P. Kirkpatrick, C. Ellis, *Nature* **2004**, *432*, 823–823.

[19] M. H. S. Segler, M. Preuss, M. P. Waller, *Nature* **2018**, *555*, 604–610.

[20] G. Schneider, U. Fechner, *Nat Rev Drug Discov* **2005**, *4*, 649–663.

[21] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas, A. A. Lee, *ACS Cent. Sci.* **2019**, *5*, 1572–1583.

[22] N. Brown, M. Fiscato, M. H. Segler, A. C. Vaucher, *J.*

10.1002/anie.202219170

Chem. Inf. Model. **2019**, *59*, 1096–1108.

[23] C. R. Groom, I. J. Bruno, M. P. Lightfoot, S. C. Ward, *Acta Crystallogr B Struct Sci Cryst Eng Mater* **2016**, *72*, 171–179.

[24] H. Berman, K. Henrick, H. Nakamura, J. L. Markley, *Nucleic Acids Res.* **2007**, *35*, D301–D303.

[25] D. Zagorac, H. Müller, S. Ruehl, J. Zagorac, S. Rehme, *J Appl Crystallogr* **2019**, *52*, 918–925.

[26] L. Foppa, L. M. Ghiringhelli, F. Girgsdies, M. Hashagen, P. Kube, M. Hävecker, S. J. Carey, A. Tarasov, P. Kraus, F. Rosowski, R. Schlögl, A. Trunschke, M. Scheffler, *MRS Bulletin* **2021**, *46*, 1016–1026.

[27] F. Strieth-Kalthoff, F. Sandfort, M. Kühnemund, F. R. Schäfer, H. Kuchen, F. Glorius, *Angew Chem Int Ed* **2022**, *61*, 8660.

[28] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. A. Persson, *APL Materials* **2013**, *1*, 011002.

[29] R. H. Taylor, F. Rose, C. Toher, O. Levy, K. Yang, M. Buongiorno Nardelli, S. Curtarolo, *Comput. Mater. Sci.* **2014**, *93*, 178–192.

[30] L. Talirz, S. Kumbhar, E. Passaro, A. V. Yakutovich, V. Granata, F. Gargiulo, M. Borelli, M. Uhrin, S. P. Huber, S. Zoupanos, C. S. Adorf, C. W. Andersen, O. Schütt, C. A. Pignedoli, D. Passerone, J. Vandevondele, T. C. Schulthess, B. Smit, G. Pizzi, N. Marzari, *Sci Data* **2020**, *7*, 17.

[31] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, C. Wolverton, *npj Comput Mater* **2015**, *1*, 864.

[32] C. Draxl, M. Scheffler, *J. Phys. Mater.* **2019**, *2*, 036001.

[33] L. Ruddigkeit, R. Van Deursen, L. C. Blum, J.-L. Reymond, *J. Chem. Inf. Model.* **2012**, *52*, 2864–2875.

[34] R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld, *Sci Data* **2014**, *1*, 191.

[35] J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, A. E. Roitberg, *J. Chem. Phys.* **2018**, *148*, 241733.

[36] J. Hoja, L. Medrano Sandonas, B. G. Ernst, A. Vazquez-Mayagoitia, R. A. Distasio, A. Tkatchenko, *Sci Data* **2021**, *8*, 649.

[37] A. Stuke, C. Kunkel, D. Golze, M. Todorović, J. T. Margraf, K. Reuter, P. Rinke, H. Oberhofer, *Sci Data* **2020**, *7*, 241722.

[38] S. Stocker, G. Csányi, K. Reuter, J. T. Margraf, *Nat Commun* **2020**, *11*, 227.

[39] B. Cheng, R.-R. Griffiths, S. Wengert, C. Kunkel, T. Stenczel, B. Zhu, V. L. Deringer, N. Bernstein, J. T. Margraf, K. Reuter, G. Csanyi, *Acc. Chem. Res.* **2020**, *53*, 1981–1991.

[40] L. Chanussot, A. Das, S. Goyal, T. Lavril, M. Shuaibi, M. Riviere, K. Tran, J. Heras-domingo, C. Ho, W. Hu, A. Palizhati, A. Sriram, B. Wood, J. Yoon, D. Parikh, C. L. Zitnick, Z. Ulissi, *ACS Catal.* **2021**, *11*, 6059–6072.

[41] M. F. Langer, A. Goeßmann, M. Rupp, *npj Comput Mater* **2022**, *8*, 8732.

[42] F. Musil, A. Grisafi, A. P. Bartók, C. Ortner, G. Csányi, M. Ceriotti, *Chem. Rev.* **2021**, *121*, 9759–9815.

[43] O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, K.-R. Müller, *Chem. Rev.* **2021**, *121*, 10142–10186.

[44] A. M. Miksch, T. Morawietz, J. Kästner, A. Urban, N. Artrith, *Mach. Learn.: Sci. Technol.* **2021**, *2*, 031001.

[45] B. Huang, O. A. von Lilienfeld, *Chem. Rev.* **2021**, *121*, 10001–10036.

[46] O. A. von Lilienfeld, *Angew. Chem. Int. Ed.* **2018**, *57*, 4164–4169.

[47] J. Behler, M. Parrinello, *Phys. Rev. Lett.* **2007**, *98*, 583.

[48] N. Bernstein, G. Csányi, V. L. Deringer, *npj Comput Mater* **2019**, *5*, aad3000.

[49] F. Musil, M. J. Willatt, M. A. Langovoy, M. Ceriotti, *J. Chem. Theory Comput.* **2019**, *15*, 906–915.

[50] K. Tran, W. Neiswanger, J. Yoon, Q. Zhang, E. Xing, Z. W. Ulissi, *Mach. Learn.: Sci. Technol.* **2020**, *1*, 025006.

[51] R. Jinnouchi, J. Lahnsteiner, F. Karsai, G. Kresse, M. Bokdam, *Phys. Rev. Lett.* **2019**, *122*, 225701.

[52] J. Timmermann, F. Kraushofer, N. Resch, P. Li, Y. Wang, Z. Mao, M. Riva, Y. Lee, C. Staacke, M. Schmid, C. Scheurer, G. S. Parkinson, U. Diebold, K. Reuter, *Phys. Rev. Lett.* **2020**, *125*, 206101.

[53] J. Timmermann, Y. Lee, C. G. Staacke, J. T. Margraf, C. Scheurer, K. Reuter, *J. Chem. Phys.* **2021**, *155*, 244107.

[54] M. Shuaibi, S. Sivakumar, R. Q. Chen, Z. W. Ulissi, *Mach. Learn.: Sci. Technol.* **2020**, *2*, 025007.

[55] J. Vandermause, Y. Xie, J. S. Lim, C. J. Owen, B. Kozinsky, *Nat Commun* **2022**, *13*, 307.

[56] P. Liu, J. Wang, N. Avargues, C. Verdi, A. Singraber, F. Karsai, X.-Q. Chen, G. Kresse, *Phys. Rev. Lett.* **2023**, *130*, 078001.

[57] J. Vandermause, S. B. Torrisi, S. Batzner, Y. Xie, L. Sun, A. M. Kolpak, B. Kozinsky, *npj Comput Mater* **2020**, *6*, 104108.

[58] H. Jung, L. Sauerland, S. Stocker, K. Reuter, J. T. Margraf, *ChemRxiv:10.26434/chemrxiv-2022-q3j0s* **2022**.

[59] C. Kunkel, J. T. Margraf, K. Chen, H. Oberhofer, K. Reuter, *Nat Commun* **2021**, *12*, 675.

[60] D. E. Graff, E. I. Shakhnovich, C. W. Coley, *Chem. Sci.* **2021**, *12*, 7866–7881.

[61] P. O. Dral, A. Owens, S. N. Yurchenko, W. Thiel, *J. Chem. Phys.* **2017**, *146*, 244108.

[62] L. Hörmann, A. Jeindl, A. T. Egger, M. Scherbela, O. T. Hofmann, *Comput. Phys. Commun.* **2019**, *244*, 143–155.

[63] M. Todorović, M. U. Gutmann, J. Corander, P. Rinke, *npj Comput Mater* **2019**, *5*, 1029.

[64] M. K. Bisbo, B. Hammer, *Phys. Rev. Lett.* **2020**, *124*, 086102.

[65] V. L. Deringer, C. J. Pickard, G. Csányi, *Phys. Rev. Lett.* **2018**, *120*, 156001.

[66] M.-P. V. Christiansen, N. Rønne, B. Hammer, *J. Chem. Phys.* **2022**, *157*, 054701.

[67] M. K. Warmuth, J. Liao, G. Rätsch, M. Mathieson, S. Putta, C. Lemmen, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 667–673.

[68] Y. Zhang, A. A. Lee, *Chem. Sci.* **2019**, *10*, 8154–8163.

[69] C. Zeni, A. Anelli, A. Glielmo, K. Rossi, *Phys. Rev. B* **2022**, *105*, 165141.

[70] S. N. Pozdnyakov, M. J. Willatt, A. P. Bartók, C. Ortner, G. Csányi, M. Ceriotti, *Phys. Rev. Lett.* **2020**, *125*, 166001.

[71] B. Parsaeifard, D. Sankar de, A. S. Christensen, F. A. Faber, E. Kocer, S. De, J. Behler, O. Anatole von Lilienfeld, S. Goedecker, *Mach. Learn.: Sci. Technol.* **2021**, *2*, 015018.

[72] B. Parsaeifard, S. Goedecker, *J. Chem. Phys.* **2022**, *156*, 034302.

[73] S. N. Pozdnyakov, M. J. Willatt, A. P. Bartók, C. Ortner, G. Csányi, M. Ceriotti, *J. Chem. Phys.* **2022**, *157*, 177101.

[74] B. Parsaeifard, M. Krummenacher, S. Goedecker, *J. Chem. Phys.* **2022**, *157*, 177102.

[75] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, B. Kozinsky, *Nat Commun* **2022**, *13*, 1.

[76] A. Grisafi, D. M. Wilkins, G. Csányi, M. Ceriotti, *Phys. Rev. Lett.* **2018**, *120*, 036002.

[77] K. Schütt, O. Unke, M. Gastegger, *Proceedings of the 38th International Conference on Machine Learning*, **2021**, pp. 9377–9388.

[78] P. B. Jørgensen, A. Bhowmik, *npj Comput Mater* **2022**, *8*, 736.

[79] M. Veit, D. M. Wilkins, Y. Yang, R. A. Distasio, M. Ceriotti, *J. Chem. Phys.* **2020**, *153*, 024113.

[80] A. Grisafi, A. Fabrizio, B. Meyer, D. M. Wilkins, C. Corminboeuf, M. Ceriotti, *ACS Cent. Sci.* **2019**, *5*, 57–64.

[81] A. M. Lewis, A. Grisafi, M. Ceriotti, M. Rossi, *J. Chem. Theory Comput.* **2021**, *17*, 7203–7214.

[82] I. Batatia, D. P. Kovacs, G. N. C. Simm, C. Ortner, G. Csanyi, *Advances in Neural Information Processing Systems*, **2022**.

[83] R. Drautz, *Phys. Rev. B* **2019**, *99*, 014104.

[84] Y. Lysogorskiy, C. v. d. Oord, A. Bochkarev, S. Menon, M. Rinaldi, T. Hammerschmidt, M. Mrovec, A. Thompson, G. Csányi, C. Ortner, R. Drautz, *npj Comput Mater* **2021**, *7*, 014104.

[85] R. M. Richard, K. U. Lao, J. M. Herbert, *Acc. Chem. Res.* **2014**, *47*, 2828–2836.

[86] G. Dusson, M. Bachmayr, G. Csányi, R. Drautz, S. Etter, C. Van der oord, C. Ortner, *J. Comput. Phys.* **2022**, *454*, 110946.

[87] D. P. Kovács, C. v. d. Oord, J. Kucera, A. E. A. Allen, D. J. Cole, C. Ortner, G. Csányi, *J. Chem. Theory Comput.* **2021**, *17*, 7696–7711.

[88] S. R. Xie, M. Rupp, R. G. Hennig, *arXiv:2110.00624v1* **2021**.

[89] R. K. Lindsey, L. E. Fried, N. Goldman, *J. Chem. Theory Comput.* **2017**, *13*, 6222–6229.

[90] M. Gastegger, J. Behler, P. Marquetand, *Chem. Sci.* **2017**, *8*, 6924–6935.

[91] J. Westermayr, M. Gastegger, P. Marquetand, *J. Phys. Chem. Lett.* **2020**, *11*, 3828–3834.

[92] Y. Zhang, R. J. Maurer, B. Jiang, *J. Phys. Chem. C* **2020**, *124*, 186–195.

[93] H. Jung, S. Stocker, C. Kunkel, H. Oberhofer, B. Han, K. Reuter, J. T. Margraf, *ChemSystemsChem* **2020**, *2*, 659.

[94] R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld, *J. Chem. Theory Comput.* **2015**, *11*, 2087–2096.

[95] M. Hennemann, T. Clark, *J Mol Model* **2014**, *20*, 4899.

[96] G. Fan, A. Mcsloy, B. Aradi, C.-Y. Yam, T. Frauenheim, *J. Phys. Chem. Lett.* **2022**, *13*, 10132–10139.

[97] P. Zheng, R. Zubatyuk, W. Wu, O. Isayev, P. O. Dral, *Nat Commun* **2021**, *12*, 479.

[98] M. Ruth, D. Gerbig, P. R. Schreiner, *J. Chem. Theory Comput.* **2022**, *18*, 4846–4855.

[99] C. Qu, Q. Yu, R. Conte, P. L. Houston, A. Nandi, J. M. Bomwan, *Digital Discovery* **2022**, *1*, 658–664.

[100] K. Chen, C. Kunkel, K. Reuter, J. T. Margraf, *Digital Discovery* **2022**, *1*, 147–157.

[101] G. Sun, P. Sautet, *J. Chem. Theory Comput.* **2019**, *15*, 5614–5627.

[102] S. Wengert, G. Csányi, K. Reuter, J. T. Margraf, *J. Chem. Theory Comput.* **2022**, *18*, 4586–4593.

[103] S. Wengert, G. Csányi, K. Reuter, J. T. Margraf, *Chem. Sci.* **2021**, *12*, 4536–4546.

[104] H. Muhli, X. Chen, A. P. Bartók, P. Hernández-León, G. Csányi, T. Ala-Nissila, M. A. Caro, *Phys. Rev. B* **2021**, *104*, 054106.

[105] J. Westermayr, S. Chaudhuri, A. Jeindl, O. T. Hofmann, R. J. Maurer, *Digital Discovery* **2022**, *1*, 463–475.

[106] A. P. Bartók, M. C. Payne, R. Kondor, G. Csányi, *Phys. Rev. Lett.* **2010**, *104*, 136403.

[107] M. Gaus, A. Goez, M. Elstner, *J. Chem. Theory Comput.* **2013**, *9*, 338–354.

[108] A. S. Christensen, S. K. Sirumalla, Z. Qiao, M. B. O'Connor, D. G. A. Smith, F. Ding, P. J. Bygrave, A. Anandkumar, M. Welborn, F. R. Manby, T. F. Miller, *J. Chem. Phys.* **2021**, *155*, 204103.

[109] Z. Qiao, M. Welborn, A. Anandkumar, F. R. Manby, T. F. Miller, *J. Chem. Phys.* **2020**, *153*, 124111.

[110] K. T. Schütt, M. Gastegger, A. Tkatchenko, K.-R. Müller, R. J. Maurer, *Nat Commun* **2019**, *10*, 146401.

[111] J. Westermayr, M. Gastegger, D. Vörös, L. Panzenboeck, F. Joerg, L. González, P. Marquetand, *Nat. Chem.* **2022**, *14*, 914–919.

[112] J. Nigam, M. J. Willatt, M. Ceriotti, *J. Chem. Phys.* **2022**, *156*, 014115.

[113] P. O. Dral, O. A. von Lilienfeld, W. Thiel, *J. Chem. Theory Comput.* **2015**, *11*, 2120–2125.

[114] J. T. Margraf, K. Reuter, *J. Phys. Chem. A* **2018**, *122*, 6343–6348.

[115] J. Hermann, Z. Schätzle, F. Noé, *Nat. Chem.* **2020**, *12*, 891–897.

[116] F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke, K.-R. Müller, *Nat Commun* **2017**, *8*, A1133.

[117] A. Chandrasekaran, D. Kamal, R. Batra, C. Kim, L. Chen, R. Ramprasad, *npj Comput Mater* **2019**, *5*, 436.

[118] M. Bogojeski, L. Vogt-Maranto, M. E. Tuckerman, K.-R. Müller, K. Burke, *Nat Commun* **2020**, *11*, 058301.

[119] J. T. Margraf, K. Reuter, *Nat Commun* **2021**, *12*, B864.

[120] A. Fabrizio, A. Grisafi, B. Meyer, M. Ceriotti, C. Corminboeuf, *Chem. Sci.* **2019**, *10*, 9424–9432.

[121] A. J. Cohen, P. Mori-Sánchez, W. Yang, *Science* **2008**, *321*, 792–794.

[122] J. Kirkpatrick, B. McMorrow, D. H. P. Turban, A. L. Gaunt, J. S. Spencer, A. G. D. G. Matthews, A. Obika, L. Thiry, M. Fortunato, D. Pfau, L. R. Castellanos, S. Petersen, A. W. R. Nelson, P. Kohli, P. Mori-Sánchez, D. Hassabis, A. J. Cohen, *Science* **2021**, *374*, 1385–1389.

[123] A. D. Becke, *J. Chem. Phys.* **2022**, *156*, 214101.

[124] A. D. Becke, *J. Chem. Phys* **2022**, 234102.

[125] J. Schmidt, C. L. Benavides-Riveros, M. A. L. Marques, *J. Phys. Chem. Lett.* **2019**, *10*, 6425–6431.

[126] R. Nagai, R. Akashi, O. Sugino, *npj Comput Mater* **2020**, *6*, 145301.

[127] X. Lei, A. J. Medford, *Phys. Rev. Materials* **2019**, *3*, 063801.

[128] K. Ryczko, D. A. Strubbe, I. Tamblyn, *Phys. Rev. A* **2019**, *100*, 022512.

[129] K. Bystrom, B. Kozinsky, *J. Chem. Theory Comput.* **2022**, *18*, 2180–2192.

[130] S. Dick, M. Fernandez-Serra, *Nat Commun* **2020**, *11*, 897.

[131] M. Tsubaki, T. Mizoguchi, *Phys. Rev. Lett.* **2020**, *125*, 206401.

[132] N. Artrith, T. Morawietz, J. Behler, *Phys. Rev. B* **2011**, *83*, 153101.

[133] T. Morawietz, V. Sharma, J. Behler, *J. Chem. Phys.* **2012**, *136*, 064103.

[134] S. A. Ghasemi, A. Hofstetter, S. Saha, S. Goedecker, *Phys. Rev. B* **2015**, *92*, 045131.

[135] T. W. Ko, J. A. Finkler, S. Goedecker, J. Behler, *Nat Commun* **2021**, *12*, 585.

[136] C. G. Staacke, S. Wengert, C. Kunkel, G. Csányi, K. Reuter, J. T. Margraf, *Mach. Learn.: Sci. Technol.* **2022**, *3*, 015032.

[137] S. Chmiela, V. Vassilev-Galindo, O. T. Unke, A. Kabylda, H. E. Sauceda, A. Tkatchenko, K.-R. Müller, *Sci. Adv.* **2023**, *9*, 1875.

[138] M. Ceriotti, *arXiv:2208.06139v1* **2022**.

[139] S. Stocker, J. Gasteiger, F. Becker, S. Günnemann, J. T. Margraf, *Mach. Learn.: Sci. Technol.* **2022**, *3*, 045010.

[140] J. Gasteiger, F. Becker, S. Günnemann, *Conference on Neural Information Processing Systems (NeurIPS)*, **2021**.

[141] G. Imbalzano, Y. Zhuang, V. Kapil, K. Rossi, E. A. Engel, F. Grasselli, M. Ceriotti, *J. Chem. Phys.* **2021**, *154*, 074102.

[142] M. S. Chen, J. Lee, H.-Z. Ye, T. C. Berkelbach, D. R. Reichman, T. E. Markland, *arXiv:2211.16619v1* **2022**.

[143] S. Thaler, J. Zavadlav, *Nat Commun* **2021**, *12*, 230902.

[144] W. L. Jorgensen, D. S. Maxwell, J. Tirado-Rives, *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

[145] A. D. White, G. A. Voth, *J. Chem. Theory Comput.* **2014**, *10*, 3023–3030.

[146] J. Chen, J. Chen, G. Pinamonti, C. Clementi, *J. Chem. Theory Comput.* **2018**, *14*, 3849–3858.

[147] S. Olsson, H. Wu, F. Paul, C. Clementi, F. Noé, *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, 8265–8270.

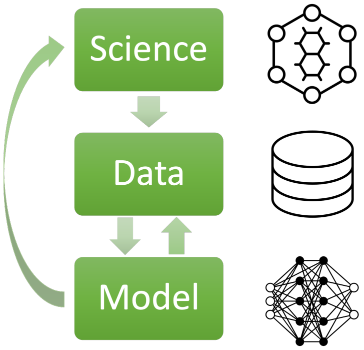
[148] B. Kreitz, K. Sargsyan, K. Blöndal, E. J. Mazeau, R. H. West, G. D. Wehinger, T. Turek, C. F. Goldsmith, *JACS Au* **2021**, *1*, 1656–1673.

[149] J. T. Margraf, M. Hennemann, T. Clark, *J Mol Model* **2020**, *26*, 1.

[150] J. S. Smith, B. T. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev, A. E. Roitberg, *Nat Commun* **2019**, *10*, A311.

Keywords: Machine Learning, Atomistic Simulations, Artificial Intelligence, Chemical Data, Molecular Dynamics

Table of Contents: Machine learning algorithms are currently emerging as powerful tools in all areas of science. This review covers atomistic machine learning approaches in chemistry beyond the conventional data-driven perspective.



10.1002/anie.202219170

Johannes T. Margraf studied chemistry at the University of Erlangen, where he also obtained his PhD. Subsequently he joined the Quantum Theory Project at the University of Florida as a PostDoc, funded by a Feodor-Lynen fellowship. This was followed by another postdoctoral fellowship at the Technical University of Munich. Since 2021, he is a group leader at the Theory Department of the Fritz-Haber-Institute in Berlin. His group focuses on using and developing machine-learning and electronic structure methods to study chemical reactions and discover new functional materials.

10.1002/anie.202219170