Batteries & Supercaps

Minireviews
doi.org/10.1002/batt.202000262

Chemistry Europe
European Chemical Societies Publishing

VIP *Very Important Paper*

Special Collection

Modelling Bulk Electrolytes and Electrolyte Interfaces with Atomistic Machine Learning

Yunqi Shao,[a] Lisanne Knijff,[a] Florian M. Dietrich,[a] Kersti Hermansson,[a] and Chao Zhang*[a]

Batteries and supercapacitors are electrochemical energy storage systems which involve multiple time-scales and length-scales. In terms of the electrolyte which serves as the ionic conductor, a molecular-level understanding of the corresponding transport phenomena, electrochemical (thermal) stability and interfacial properties is crucial for optimizing the device performance and achieving safety requirements. To this end, atomistic machine learning is a promising technology for bridging microscopic models and macroscopic phenomena. Here, we provide a timely snapshot of recent advances in this area. This includes technical considerations that are particularly relevant for modelling electrolytes as well as specific examples of both bulk electrolytes and associated interfaces. A perspective on methodological challenges and new applications is also discussed.

1. Introduction

Electrolytes occupy a central place in chemical science ever since Faraday invented the term in 1834 for describing "a body decomposable by the passage of the electric current"[1] and the rise of "the ionists" represented by Ostwald, van 't Hoff, and Arrhenius in the late 19th century.[2] Technology-wise, the electrolyte is an indispensable component in batteries (Li-ion,[3] Li-air,[4] Na-ion,[5] solid-state,[6,7] aqueous[8]) and supercapacitors.[9,10]

According to Goodenough, "an electrolyte is an ionic conductor and an electronic insulator".[11] In general, electrolytes can be divided into aqueous and nonaqueous electrolytes. Since the energy density and the power density of an electrochemical cell scale linearly and quadratically, respectively, with its operating voltage,[12] aqueous electrolytes, with the hydrolysis potential of water being 1.23 V, are not a popular choice for neither battery or supercapacitor applications. Nevertheless, it is worth to note that the first rechargeable battery with a nominal cell voltage of 2.1 V, i.e. the lead-acid battery, was based on an aqueous electrolyte, thanks to the slow kinetics of side reactions. Because of the voltage limit in aqueous electrolytes, liquid organic electrolytes which possess a similar fast ionic conductivity have been widely used in Li-ion batteries instead.[3,4] On the other hand, liquid organic electrolytes, which usually contain flammable groups such as carbonates and ethers, are facing more serious safety issues in post-Li-ion batteries, which have much higher energy densities.[13] In this regard, inorganic crystalline electrolytes,[6,7] polymer electrolytes[14,15] and room-temperature ionic liquids[16,17] are promising alternatives.

Being ionic conductors, good ionic conductivity is a key requirement for electrolytes in order to reduce the concentration polarization and to enhance the diffusion rate. Different types of electrolytes achieve this through various types of ion transport mechanisms.[18] In aqueous electrolytes, the ion travels largely with its solvation shell and diffuses slower than the solvent molecules, which leads to the so-called vehicular mechanism; in polymer electrolytes, the ion hops among the solvent molecules and diffuses faster than the solvent, which is termed the structural mechanism. While the ion conduction in liquid organic electrolytes can vary between the vehicular mechanism and the structural mechanism due to the nature of the solvent, that in inorganic crystalline electrolytes occurs essentially by way of the structural mechanism via vacancies and interstitial sites. In addition, the concentration of electrolytes is another determining factor for the ionic conductivity. Increasing the electrolyte concentration does not only affect the ion transport because of the ion-ion correlation and the ion-pairing,[19,20] but it can also lead to a remarkable electrochemical stability, e.g. "water in salt" electrolytes which can stand up to 3 V operating voltage for Li-ion battery applications.[8,18]

Apart from desired bulk properties of the electrolyte such as a good ionic conductivity and a high electrochemical (thermal) stability, electrolyte interfaces are of paramount importance in electrochemical energy storage systems. There are two types of electrolyte interfaces: electrode-electrolyte interfaces and electrolyte-electrolyte interfaces. A famous example in the former case is the solid-electrolyte interphase (SEI) that grows at the carbon anode/liquid organic electrolyte interface due to the accompanying redox reactions.[21] What is astonishing is that in contrast to ethylene carbonate, propylene carbonate, which has a single methyl group difference, cannot form a stable SEI and leads to the exfoliation of the graphitic anode because of co-intercalation.[22] This highlights the necessity to understand the interfacial chemistry at the

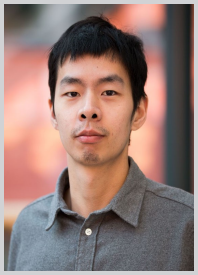
[a] *Y. Shao, L. Knijff, F. M. Dietrich, Prof. K. Hermansson, Dr. C. Zhang*
Department of Chemistry-Ångström Laboratory, Uppsala University, Lägerhyddsvägen 1, BOX 538, 75121, Uppsala, Sweden
E-mail: chao.zhang@kemi.uu.se

Special Collection *An invited contribution to a Special Collection on Artificial Intelligence in Electrochemical Energy Storage*

electrode-electrolyte interface.[23] The same applies to electrolyte-electrolyte interfaces, which seem less mentioned. Besides the obvious case in composite electrolytes,[24] where the interfacial resistance can be a critical issue,[7] one needs to realize that the SEI itself is an electrolyte as well and electrolyte-electrolyte interfaces are ubiquitous. In addition, it is worth mentioning that electrodes are usually in the form of porous structures. This means that the electrolyte is actually in (nano) confinement and possesses unique physico-chemical properties, which has been the key to the success of high-voltage supercapitors.[9,10]

The vast amount of insight about electrolytes and electrolyte interfaces in electrochemical energy storage systems, that has been collected so far and mentioned above briefly, would not have been possible without spectroscopic measurements and structural characterizations.[25] Compared to the top-down approach commonly used in experiments, atomistic modelling is usually built in a bottom-up fashion and provides complementary information. Based on the principles of quantum mechanics and statistical mechanics, density functional theory (DFT) calculations[26] and DFT based molecular dynamics (DFTMD) simulations[27] are eminently suitable techniques for modelling electrolytes and electrolyte interfaces where the distinction between reactive solutes and solvent has all but disappeared. Despite recent advances in the applications of DFT and DFTMD in this area,[28–37] many electrochemical properties of bulk electrolytes and electrolyte interfaces, and the corresponding processes, e.g. fully converged ionic conductivities and ion migrations across the grain boundary, cannot possibly be described with a few hundred atoms and tens of picoseconds which are typical for DFTMD.[38,39] Therefore, data-driven approaches such as atomistic machine learning may be a game changer because of the automated feature engineering, and the universal approximation ability.

In this review, we will focus on recent developments and applications of atomistic machine learning in modelling electrolytes and associated interfaces. The goal is to provide a timely snapshot of this emerging area and to identify gaps and new directions. Standing as computational (electro)chemists, we limit the scope of this review to the intersection between atomistic modelling and machine learning, where exciting progresses such as automated experiments with machine learning will not be included here.[40] In addition, our viewpoint is from electrolytes, where long-range interactions, external fields, fluctuations and dynamics are important. Therefore, atomistic machine learning of electrode materials will not be discussed and excellent reviews covering this topic can be found elsewhere.[41–44] In the following, we will first introduce the framework of atomistic machine learning and discuss technical aspects which are particularly relevant to modelling electrolytes. Then, selected examples on different types of electrolytes and electrolyte interfaces modelled with atomistic machine learning will be presented, where challenges and new opportunities will be discussed.

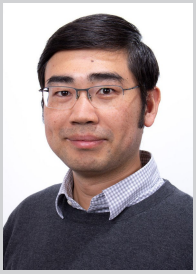


Yunqi Shao is a PhD student in the Department of Chemistry-Ångström Laboratory at Uppsala University. He obtained his B. Eng. in Chemical Engineering and Technology (2015) and M. Eng. in Chemical Technology (2018), from East China University of Science and Technology. His research interest includes the machine learning-aided modelling of bulk electrolytes and associated interfaces.

Lisanne Knijff is a PhD student in the Department of Chemistry-Ångström Laboratory at Uppsala University. She obtained her B.Sc. in Chemistry from Utrecht University in 2018, and her M.Sc. in Theoretical Chemistry from Uppsala University in 2020. Her research interests include the machine learning-aided simulations of aqueous electrolyte solutions, and metal oxide-electrolyte interfaces.

Florian M. Dietrich is a graduate student in the Structural Chemistry programme at Uppsala University. He obtained his B.Sc. in Chemistry at the Technische Universität Darmstadt (TUD) in 2019. His research interests include developing neural-network potentials and machine learning methods for multi-scale modelling.

Kersti Hermansson is Professor of Inorganic Chemistry at Uppsala University. She conducts research on multi-scale modelling of workflows to enable (adequately) accurate simulations of both fundamental problems and complex chemical applications, especially at the electronic and atomistic levels. Redox chemistry of nano-structured metal oxides, water in all sorts of molecular or condensed environments including water/solid interfaces, and computational vibrational spectroscopies are special interests. She is a Fellow of the Royal Swedish Academy of Sciences (KVA).



Chao Zhang is a tenure-track assistant professor in the Department of Chemistry-Ångström Laboratory at Uppsala University. He obtained Dr. rer. nat. from RWTH Aachen University in 2013 and Docent ("venia docendi") from Uppsala University in 2020. Before joining Uppsala in 2017, he was a postdoctoral researcher in the Department of Chemistry at Cambridge University. His group focuses on developing finite-field methods in computational electrochemistry as well as multi-scale modelling of electrolyte materials and electrified solid-liquid interfaces in energy storage/conversion. He is the recipient of ERC Starting Grant (2020), Junior Research Fellowship from Wolfson College (2015) and Jülich Excellence Prize for Young Scientists (2013).

2. Atomistic machine learning

2.1. Atomistic machine learning in a nutshell

According to Arthur Samuel who coined the term, machine learning is defined as a field where "Programming computers to learn from experience should eventually eliminate the need for much of this detailed programming effort".[45] In general, machine learning techniques become useful when i) We have data; ii) There is a pattern in the data; iii) The pattern is difficult to describe mathematically.[46] When we have data in both the input space $\mathcal{X}$ and the output space $\mathcal{Y}$, then the data are called labelled data and machine learning applied to this data is supervised learning. Otherwise, in unsupervised learning, we deal with data in the input space $\mathcal{X}$ only. While unsupervised learning is quite powerful in data clustering and dimensionality reduction, it is supervised learning which beats humans in image recognition.[47] In supervised learning, neural networks (NN) with regularization and kernel methods such as Gaussian process regression (GPR) and kernel ridge regression (KRR) are popular choices.[48] In spite of their different technical flavors and computational efficiencies, the goal to maximize the posterior probability $p\left(h(\mathbf{x})|\mathcal{D}=\{(\mathbf{x}_i,y_i)\}_{i=1}^{N}\right)$ for a chosen model h given data $\mathcal{D}$ is the same.[49]

The key aspect of current atomistic machine learning[50–53] is to establish a computationally efficient mapping $h:\{\mathbf{x}_i,Z_i,S,Q,\mathbf{D}\cdots\}\rightarrow y$ via machine learning techniques such as NN, GPR and KRR,[48] where y is the desired property, $\mathbf{x}_i$ is the position of atom i, Z_i is the atomic number of atom i, S is the multiplicity of the system, Q is the net charge of the system and $\mathbf{D}$ is the external electric field. One popular task in atomisitic machine learning is to approximate the potential energy surface (PES), i.e. machine learning potential, when y is the potential energy. A machine learning potential has the ability to provide quantum mechanical accuracy without explicitly including electrons,[54] thanks to the flexibility of the multi-layer high-dimensional atomic neural network developed by Behler and Parrinello.[55] Beyond machine learning potentials, atomistic machine learning has also been successful in predicting physico-chemical properties of molecules and materials in both compositional and conformational spaces.[56–58]

To successfully carry out atomistic machine learning, one would need representations of the data, i.e. so-called descriptors, and representative data, i.e. a subset of the data with the most information and a low redundancy. We will survey these aspects in the following sections, including considerations which are particularly important for modelling bulk electrolytes and electrolyte interfaces, e.g. long-range interactions and external fields.

2.2. Descriptors in atomistic machine learning

What makes atomistic machine learning unique is that the input data contains at least element types and structures of molecules or materials. This further requires descriptors in atomistic machine learning to satisfy two conditions: symmetry and uniqueness. A good descriptor should be invariant under symmetry operations such as translation and rotation of the entire structure, and the permutation of same-type atoms. This prevents nonphysical behavior of the resulting PES and reduces the dimension of the descriptor space. A unique descriptor means that there is a one-to-one mapping from each structure to each representation given by the descriptors. This guarantees that the learning algorithm always reaches a higher accuracy with more data.[59,60] Generally speaking, there are three types of descriptors in the current landscape of atomistic machine learning: global descriptors, local descriptors and graph based descriptors, which are summarized in Figure 1 and discussed as follows.

While the internal coordinates such as bond length are natural descriptors which satisfy the translational and rotational symmetry, additional procedures are needed in order to preserve the permutational invariance which was recognized in the early works[61,62] and led to the development of permutation invariant polynomials (PIPs) for an entire molecular cluster.[63] However, the major limitation of global descriptors $\mathbf{G}\left(\{\mathbf{x}_i\}_{i=1}^{N}\right)$ for a N-atom system is that they are hard to apply to condensed phase systems with periodic boundary conditions. Global descriptors either do not transfer to systems of different sizes (e.g. the Coulomb matrix:[64] a $N\times N$ matrix with scaled and inverse atomic distances) or degrade in terms of performance with an increasing number of atoms (e.g. the Bag of Bonds:[65] an element-pair histogram of the Coulomb matrix).

The issue with global descriptors was overcome by exploring the locality of chemical interactions.[66] Behler and Parrinello (BP) proposed the decomposition of total energy into atomic contributions and the usage of atom-centered descriptors.[55] In the same spirit, a spectrum of local descriptors $\mathbf{L}\left(\mathbf{x}_i,\{\mathbf{x}_j\}_{j\in O_i}\right)$, which contains a center atom i and surrounding atom j in its neighborhood O_i, was constructed by histograming many-body correlation functions, e.g. the BP symmetry functions (SFs), the many-body tensor representation (MBTR) and the Faber-Christensen-Huang-Lilienfeld (FCHL) descriptor,[55,67,68] employing a local version of the symmetry invariant polynomials, e.g. atomic PIPs and the moment tensor potential (MTP),[69,70] referring to the atomic local frame (ALF) with the center atom as the origin[71,72] and using symmetrized coefficients from the spherical harmonics expansions of the atomic density, e.g. the Smooth Overlap of Atomic Positions (SOAP) descriptor.[54,73,74] Interestingly, it has been shown recently that different ways of creating unique and symmetric descriptions are largely equivalent.[75,76]

Local descriptors enjoy convenience and efficiency, since their dimensions are pre-chosen and independent of the size of the physical system. However, many physical phenomena relevant to electrolytes and electrolyte interfaces, such as ion-ion correlations,[77] charge defects at the interface[78] and charge transfer in organic electrodes,[79] are long-range in nature. In DFT, the energy functional $E[\rho(\mathbf{x})]$ reaches its minimum value when the electron density ρ is the exact ground state density.[80] When distributing the electron density as point charges q_i, the corresponding total energy E_{tot} has to satisfy the

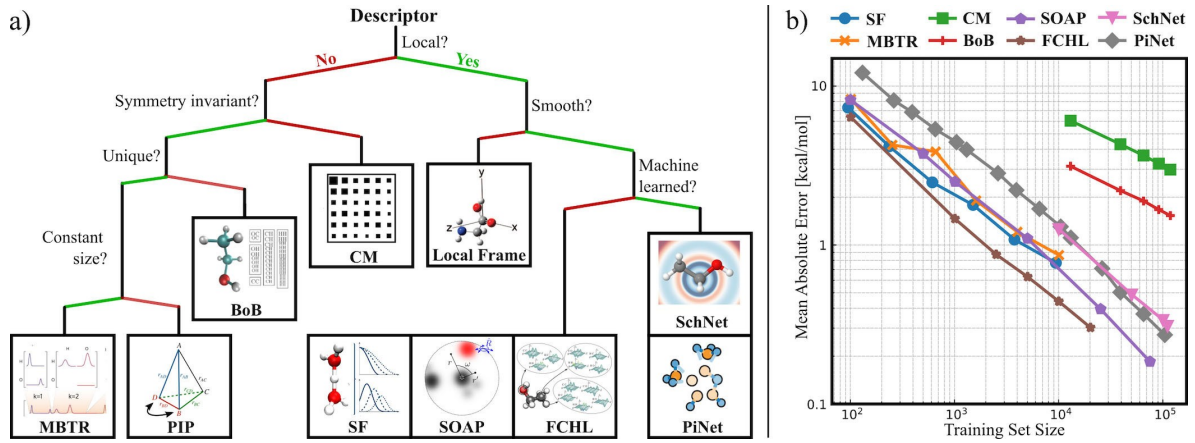


Figure 1. Representative descriptors used in atomistic machine learning. a) Illustrations of different descriptors categorized by their characteristic features. Adapted from Ref. [90] under the terms of CC-BY licence (Copyright 2020 Himanen *et al.*) for the many-body tensor representation (MBTR), Ref. [69] under the terms of CC-BY licence (Copyright 2020 van der Oord *et al.*) for the permutation invariant polynomials (PIP), Ref. [65] with permission (Copyright 2015 American Chemical Society) for the Bag of Bonds (BoB), Ref. [91] under the terms of CC-BY licence (Copyright 2013 Montavon *et al.*) for the Coulomb matrix (CM), Ref. [71] with permission (Copyright 2008 American Chemical Society) for the Atomic Local Frame (ALF), Ref. [92] with permission (Copyright 2019 American Chemical Society) for the Symmetry Function (SF), Ref. [75] with permission (Copyright 2019 AIP Publishing) for the Smooth Overlap of Atomic Positions (SOAP) descriptor, Ref. [68] with permission (Copyright 2018 AIP Publishing) for the Faber-Christensen-Huang-Lilienfeld (FCHL) descriptor, Ref. [93] with permission (All rights reserved by Klaus-Robert Müller) for SchNet and Ref. [88] under the terms of CC-BY licence (Copyright 2020 Shao *et al.*) for PiNet. b) Learning curve of selected descriptors for the QM9 dataset. Plotted with data from Ref. [94] for SF and MBTR, Ref. [53] for CM, BoB, SOAP, FCHL and SchNet. Data for PiNet was generated using the same method as detailed in Ref. [88] with a varying training set size.

same variational principle (Eq. 1) and this ensures the proper inclusion of induction and charge transfer effects. On the other hand, the Hartree potential $V_H(\mathbf{x})$ in DFT includes the $1/r$ decay over all the space and it transforms to the familiar summation when using the point charge representation (Eq. 2). This long-range feature in the Hartree potential also makes the computation rather expensive in condensed phase systems with periodic boundary conditions.

$$\frac{\delta E[\rho(\mathbf{r})]}{\delta\rho(\mathbf{r})} = \mu \Rightarrow \frac{\partial E_{\text{tot}}(\{q_i\}_i^N, \{\mathbf{x}_i\}_i^N)}{\partial q_i} = 0 \tag{1}$$

$$V_H(\mathbf{r}) = \int \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}' \Rightarrow \sum_i^N \frac{q_i}{|\mathbf{x}-\mathbf{x}_i|} \tag{2}$$

The simplest way of adding long-range interactions is to use a non-self-consistent Coulomb term as the baseline model and to generate the short-range atomic energy description on top of that, as first applied to different crystal structures of ZnO with BP neural networks.[81] While this *ad hoc* approach is possible, self-consistent schemes would be preferred. The first scheme appeared by introducing environmental-dependent atomic electronegativities χ_i and element-specific hardness J_{ii}.[82] Note that point charges q_i in this scheme were solved variationally following Eq. 1, and are therefore self-consistent. Recently, a different scheme was devised to achieve self-consistent point charges by including new types of reference data in the training, which contains the derivative of the total energy with respect to the charge $\partial E_{\text{tot}}/\partial q_i$ from constrained DFT.[83] Alternatively, when considering ρ in Eq. 2 as superimposed atomic densities instead of the electron density, the $1/r$ decay can be incorporated smoothly into the descriptor,[84] where the explicit inclusion of charges was bypassed.

Crafting global and local descriptors requires insight and domain knowledge in physics and chemistry, which is in contrast with end-to-end deep learning.[85] In deep learning, the representation of the input is learned from data rather than handcrafted. In particular, graph convolution neural networks (GCNNs), which consider the atoms as nodes and the pairwise interactions as weighted edges in the graph and explore the idea of local receptive field, are an emerging approach in atomistic machine learning of both molecules and materials.[86–88] In GCNN, one can view the graph convolution operation as an iterative generator of local descriptors by aggregating neighboring atomic information. This is carried out through two consecutive steps: a pairwise interaction operation (PI) and an interaction pooling operation (IP). First, the edge feature vector $\vec{l}_{ij}^{\,t} = \text{PI}\left(\vec{P}_i^{\,t}, \vec{P}_j^{\,t}, r_{ij}\right)$ is updated by the PI function using the node feature vectors $\vec{P}_i^{\,t}$ and $\vec{P}_j^{\,t}$ of two atoms and their distance r_{ij}, where t is an iterator which goes up to the number of graph convolution blocks N_{GC}. Then, the IP function creates an update of the node feature vector through $\vec{P}_i^{\,t+1} = \text{IP}\left(\sum_j \vec{l}_{ij}^{\,t}\right)$, where the summation ensures the permutation invariance. The actual form of the IP and PI functions varies from one GCNN to another: for instance, the edge feature vector $\vec{l}_{ij}$ depends on both node feature vectors $\vec{P}_i$ and $\vec{P}_j$ in PiNet,[88] so that each component of the edge feature vector $\vec{l}_{ij}$ has its own radial dependence. This differs from the commonly used single radial-dependent filter function (i.e. the attention mask) as in SchNet.[86]

Apart from the automated descriptor engineering, there are several advantages of GCNN over global and local descriptors. Because the generation of node feature vectors includes the

element specificity by construction, GCNN has exactly the same subnet for each element and the number of parameters in the network is constant with respect to the increase of elements. In addition, the interaction range in GCNNs is $N_{GC} \times R_c$, with R_c as the cutoff radius. In this way, long-range interactions can be effectively included. Moreover, compared to many-body or density based descriptors, the computational complexity of GCNN scales linearly with the number of neighboring atoms. This comes from the fact that $\vec{l}_{ij}$ is already a many-body function through the iteration of GC blocks. Therefore, only pairwise filters are normally needed. Nevertheless, it needs to be pointed out that the uniqueness problem of graph convolution's expressive power is still under active investigation,[89] although its representation is likely to be over-complete as evinced by a linear behavior of the corresponding learning curve even with hundreds of thousands of data (Figure 1).

2.3. Coupling to external fields in atomistic machine learning

When it comes to the modelling of electrolyte interfaces, the inclusion of the electric double layer (EDL) is inevitable. The meeting of two phases gives rise to charge separation at the boundary, which leads to the formation of EDL. Recently, it has been shown that finite field MD simulations can be quite useful for modelling EDLs at electrolyte interfaces with both metal oxides and metals.[95–99] In contrast to the standard Hamiltonian, the finite field Hamiltonian contains terms which couple polarization **P** to the Maxwell field.[100] Thus, this calls for new schemes to include the description of polarization **P** and its finite field coupling for both molecules and materials in atomistic machine learning (Figure 2).

Polarization **P**, the key quantity in finite field MD, is calculated differently for different types of systems. For isolated molecules or clusters, it is more convenient to express polarization **P** in terms of its counterpart dipole moment **M**. In this case, the dipole moment **M**, which contains both an ionic contribution $\mathbf{M}_{ion}$ and an electronic contribution $\mathbf{M}_e$ can be computed from either electron density $\rho(\mathbf{r})$ or atom-centered point charge q_i (Eq. 3). For condensed phase systems with periodic boundary conditions, **P** is formally defined through the Berry phase formula instead (Eq. 4),[101] where Ψ is the many-body wave function, $e^{i(2\pi/L)\sum_i \hat{r}_i}$ is the multiplicative operator and L is the cell length of a cubic lattice (for the sake of simplicity) with the volume Ω. Equivalently, polarization in condensed phase systems can also be expressed using the maximally localized Wannier centers $\bar{\mathbf{r}}_j$ which reflect the average position of occupied state j.[102] As a consequence, different representations of polarization mean different regression tasks in terms of atomic charges q_i, Wannier centers $\bar{\mathbf{r}}_j$ and electron density $\rho(\mathbf{r})$ in atomistic machine learning, as summarized below.

$$\mathbf{M} = \mathbf{M}_{ion} + \mathbf{M}_e = \sum_i^N Z_i \mathbf{x}_i + \int \mathbf{r}\rho(\mathbf{r})d\mathbf{r} = \Rightarrow \mathbf{M} = \sum_i^N q_i \mathbf{x}_i \quad (3)$$

$$\mathbf{P} = \frac{\mathbf{M}_{ion}}{\Omega} - \frac{e\gamma}{2\pi L^2} = \frac{\mathbf{M}_{ion}}{\Omega} - \frac{e}{2\pi L^2}\,\mathrm{Im}\,\ln\left\langle \Psi \middle| e^{i(2\pi/L)\sum_i \hat{r}_i} \middle| \Psi \right\rangle$$
$$\Rightarrow \mathbf{P} = \frac{1}{\Omega}\left(\mathbf{M}_{ion} - \sum_j^{occ} 2e\bar{\mathbf{r}}_j\right) \quad (4)$$

The best way to assign these atomic charges q_i is something chemists have struggled with for decades because there is no unique way to define the atomic charge. This is manifested by the co-existence of three main categories of charge analysis methods in computational chemistry:[103] wave function-based methods, methods based on the partitioning of the electron density, and methods which derive charges from the electrostatic potential (ESP). Depending on the type of reference charges used, atomistic machine learning has been applied to predict different types of atomic charges using local descriptors.[81,104,105] In this context, the Wannier centers $\bar{\mathbf{r}}_j$ may be seen as off-centered point charges in which their values are prescribed and their positions are to be determined. A local reference frame can be used to ensure the predicted Wannier centers rotate together with molecular entities in condensed phase systems and fluctuate near the most electronegative atoms during the time evolution of the system.[106]

Unlike atomic charge or the Wannier centers, electron density $\rho(\mathbf{r})$ is defined on a large number of grid points, making training and prediction a data-intensive task. This issue was first addressed by projecting the charge density to the Fourier basis, and describing the atomic configuration with a smooth nuclear potential.[107] However, this global descriptor approach limits its transferability to larger systems as mentioned in Section 2.2. Treating electron density as the superposition of atomic contributions $\rho(\mathbf{r}) = \sum_i^N \rho_i(\mathbf{r})$ transforms the problem into the description of local chemical environments. Similar to the case of predicting atomic multipoles, the anisotropic electron density benefits from local descriptors that

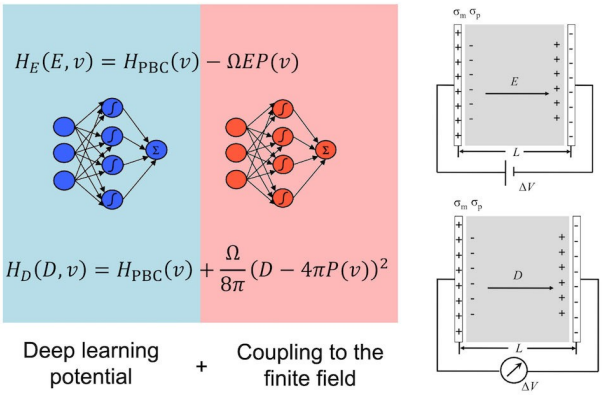


Figure 2. Combining finite field Hamiltonians with atomistic machine learning. $v = (\mathbf{r}^N, \mathbf{p}^N)$ are the classical degrees of freedom of the N atom system. E is the homogeneous Maxwell field and D is the displacement field. P is the polarization and Ω is the cell volume. Reproduced from Ref. [99] under the terms of CC-BY licence (Copyright 2020 Zhang *et al.*).

are co-variant to coordinate transformations such as the one used in the Symmetry Adapted (SA)-GPR method.[108] An alternative to this is to treat grid points as imaginary atoms, and to use local descriptors based on that.[109,110] Note that the latter approach circumvents the difficulty of regression on tensorial properties, however, it also requires a higher computational overhead because of the large amount of grid points.

2.4. Dataset generation and active learning

The purpose of sampling in atomistic machine learning is to generate a representative and labelled dataset. The prevailing set of techniques to fulfil this purpose is called active learning, in which the key ingredient is sending queries (i.e. a selection of unlabelled data) to the labelling source during the learning.[111] Depending on whether the query is synthesized by the learner *de novo* or not, there are two main scenarios in active learning: "Query Synthesis" and "Pool-based Sampling". However, "Query Synthesis" seems more appropriate in atomistic machine learning, in particular for developing machine learning potentials (Figure 3). It shares a similar spirit with the "learn on the fly" (LOTF) scheme used for accelerating MD simulations,[112,113] in which new quantum-mechanical calculations (i.e. labelling) used to fit a classical potential were performed only once in a while and MD trajectories were produced with the classical potential instead.

One way of performing active learning is directly using the descriptor. A widely adapted strategy is to propose queries based on the distance between data points in the descriptor space. The simplest way is to detect extrapolations according to the feature boundary of the training set.[114] Otherwise, the method of novelty detection based on the current density of points in the configurational space can be used, where new data that appear in the low density region are preferably included.[115] When it comes to kernel methods such as GPR and KRR, it is advantageous to find the optimal subset since the computational complexity with respect to the training data size n scales as $O(n^3)$. This includes maximizing the volume (the determinant) of the feature space following the D-optimality criterion,[116] finding the best columns of the data matrix with the leverage-score CUR algorithm[117] and farthest point sampling (FPS) which sequentially labels new data with the longest distance from the existing ones.[118] In addition, dimensionality reduction methods such as PCA can be used to cluster reference data and to perform grid based sampling with principal components.[119]

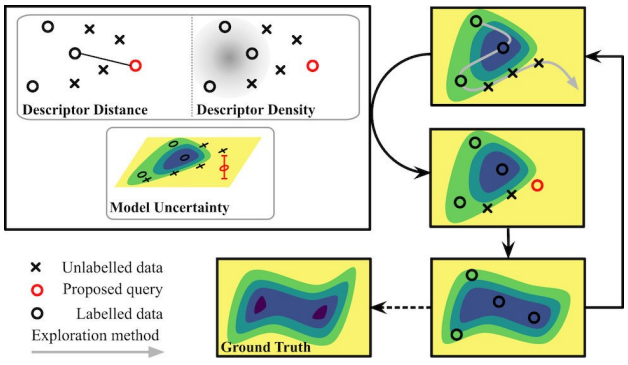


Figure 3. Illustration of a typical workflow in the active learning of the potential energy surface. Labelled data stands for atomic structures with energy and force information obtained from reference quantum-mechanical calculations. Unlabelled data means new atomic configurations generated by an exploration method (e.g. MD or MC simulation). Proposed query is an unlabelled data point which is selected based on certain metrics (e.g. descriptor distance, descriptor density and model uncertainty) and passed to reference quantum-mechanical calculations. See Text in Section 2.4 for details.

Another way of proposing queries is to use the model's uncertainty of the prediction. This is explicitly done as part of Bayesian models such as GPR[120–122] or by estimating the uncertainty with an ensemble of models, i.e. the query by committee (QbC) method.[123–125] Not surprisingly, the uncertainty estimation method affects the convergence of the learning algorithm. It has been shown that using the variance from GPR in active learning may lead to 1 order of magnitude larger prediction errors compared to that based on QbC.[120] Presumably, an overconfident estimation of the potential energy combined with a Boltzmann sampler, e.g. MD simulation, would potentially cause certain structures to be neglected, which cannot be signaled by evaluating the generated dataset. However, even QbC may still underestimate the uncertainty, in which case the descriptor based uncertainty measurement might be more reliable.[126]

3. Selected examples

3.1. Liquid electrolytes

The applications of atomistic machine learning to liquid electrolytes in the literature are mainly centered around predicting the electrochemical stability window and the ionic conductivity.

High-throughput computational screening of potential electrolyte materials has been routinely carried out for organic solvents and ionic liquids. For example, the ionization potential (IP) and the electron affinity (EA) were estimated as indicators of the electrochemical stability window of redox active species.[127–129] Obviously, the ranking accuracy depends on the iterative inclusion of structure relaxation, solvent effects and higher-level electronic structure calculations. To this end, machine learning potentials can accelerate screening tasks when the required high-level calculations become computationally prohibitive. For example, Wang *et al.*[130] combined active learning with GCNN to explore the configurational space for ether-lithium complexes. They showed that the configurational space visualized by a 2D projection of the SOAP descriptor was well covered after adding 22763 geometries from 40 distinct oligoethers to the training set.

At the same time, the transport properties of electrolytes such as ionic conductivity, transference number and viscosity requires simulating bulk liquid electrolytes at a time-scale

beyond what DFTMD can typically offer. Although those tasks can be performed with classical force fields, reactive species, polarization and charge transfer effects are often neglected. In this regard, machine learning potentials can serve as a more efficient alternative. Hellström and Behler[131] parameterized a reactive BPNN potential for an aqueous NaOH solution using 16113 reference structures from DFT calculations at different concentrations up to the solubility limit, which gave a root mean square error (RMSE) of energy and force as 1 meV/atom and 180 meV/Å respectively. This potential subsequently enabled the prediction of full ionic conductivities at different concentrations and temperatures in alkaline electrolyte solutions, which were in reasonable agreement with experiments.[132] Along the same line, Tovey *et al.* had used SOAP descriptors with GPR to generate a Gaussian approximation potential for molten NaCl with a RMSE of energy as 1.5 meV/atom.[133] They showed that the computed self-diffusion coefficients of Na^+ and Cl^- ions from MD simulations using this potential agree well with experimental results.

3.2. Solid electrolytes

There are a number of examples in the literature where atomistic machine learning was applied to study solid state electrolytes (SSEs), going from computing diffusion coefficients with machine learning potentials to mapping the composition-conductivity relationship.

Approximating the PES of ion-conducting crystalline materials with machine learning potentials is quite promising in the area of quantitative conductivity predictions for SSEs. A summary of these works is given in Table 1. Compared to experiments, the predicted diffusion activation energy has an error between 20 meV to 130 meV.[134–140]

As mentioned before, the training set constitutes the largest part of the computational cost and determines the applicability of the trained machine learning potentials. The simplest approach is to generate a large set of data that covers the widest possible variety of structures. Li *et al.*[134] emphasize the importance of tailoring the dataset to the area of application. In the case of studying diffusion, that means including structures from the region around the diffusion activation barrier. They did that by including transient structures obtained from nudged elastic band (NEB) sampling, which turned out to reduce the error of model predicted barrier energies from 73 meV to 48 meV. Alternately, attempts have been made to design automated active learning frameworks (See Section 2.4) in studying superionic conductors. The workflow of Wang *et al.*[136] is based on LOTF-MD where MTP was trained on an initial training set generated from high-temperature DFTMD. Whenever in production a new structure was encountered via the extrapolation grade criterion, then that structure was labelled and the model was retrained. By doing this, they achieve an impressive computational speedup of 10^7 with LOTF-MD, compared to DFTMD. Similarly, Marcelongo *et al.*[137] combined LOTF with DeepMD,[72] where evenly spaced trajectory frames sampled with an initially trained machine learning potential were labelled and added to the training set, then the network was retrained and the process was repeated periodically. In their scheme, the convergence of the workflow was controlled via monitoring the evolution of the target property, i.e. the cation diffusion coefficient. Huang *et al.*[138] used uncertainty based active learning to conduct exploration, labelling and training steps, where the maximum uncertainty of forces within a predefined window was used as the criteria to select queries for labelling. Guan *et al.*[140] explored a global optimization technique, i.e. Stochastic Surface Walking (SSW), in the data generation for training a neutral network potential to investigate the oxygen anion diffusion.

While these are certainly impressive feats, there is still a long way to go towards screening the entire chemical space for superionics, even with efficient bond-valence calculations to estimate the ion migration barrier.[141] One of the bottlenecks is to get sufficient amount of high-quality labelled data. In this regard, transfer learning (TL), in which the knowledge learned from the source task is used in the learning of the target task,[142] could be quite useful here. Cubuk *et al.*[143] first trained a linear support vector machine model with 40 experimental data and physically inspired descriptors such as the equilibrium Li-anion distance, then made predictions on 12176 lithium

Table 1. Summary of works on SSEs using machine learning potentials. The abbreviations used in this table are: neural network (NN), symmetry functions (SFs), moment tensor potential (MTP), linear regression (LR). See Section 2.2 and Section 2.4 for details of technical terms.

Ref.	System	Descriptor	Method	Training set	RMSE
[134]	Amorphous Li_3PO_4	SFs	NN	38592 structures from DFTMD up to 4000 K and NEB for diffusion paths	*E*: 6 meV/atom
[135]	Amorphous Ta_2O_5	SFs	NN	2000 structures from simulated annealing starting at 6000 K from DFTMD	*E*: 5 meV/atom
[136]	Eight crystalline Li oxides and sulfides	MTP	LR	LOTF with 7500 structures from 15 ps DFTMD at 1000 K	*E*: 6 meV/atom *F*: 84 meV/Å
[137]	LGPS, LLZO, NASICON	DeepMD	NN	LOTF with initial structures between 60 to 240 from DFTMD at 300 K, 600 K and 900 K	Not reported
[138]	Three LGPS-type superionic conductors	DeepMD	NN	Active learning with 590 randomly perturbed structures from DFT	*E*: 2 meV/atom *F*: 80 meV/Å
[139]	Tellurium compounds PbTe and CdTe	SFs	NN	4898 (PbTe) and 2866 (CdTe) defect-free and interstitial cation containing crystalline structures from DFT	*E*: 0.5 meV/atom *F*: 70 meV/Å
[140]	Yttria-stabilized zirconia $Y_{2x}Zr_{1-2x}O_{2-x}$	SFs	NN	28803 DFT calculations including Zr, ZrO_x, Y_2O_3, and Y/Zr mixed oxides	*E*: 7.7 meV/atom *F*: 165 meV/Å

containing structures from the Materials Project (materialsproject.org), and finally trained a model using this transferred knowledge and chemical compositions as the only inputs to screen 20 billion ternary and quaternary lithium-containing compounds. Similarly, Hatakeyama-Sato *et al.*[144] used GCNN to create a 32-dimensional feature vector by training on *de novo* generated polymer/monomer structures and their 2000 molecular properties such as the molecular polarizability, then transferred this learned feature vector together with the other system information such as the one-hot coding of inorganic additives to GPR in order to map the relationship between the composite polymer and the experimentally measured conductivity for 3000 data points. By screening different chemicals, they were able to identify dimethyl-substituted polyphenylene sulfide (PMPS) as a promising glass polymer electrolyte material.

In contrast to crystalline inorganic electrolytes, the dynamics in polymer electrolytes account for the long-distance ion transport.[19] Xie *et al.*[145] applied GCNN to coarse-grain the dynamics and to learn a linear Koopman model from MD data of a solid polymer electrolyte system – poly(ethylene oxide) (PEO) and lithium bis-trifluoromethane sulfonimide (LiTFSI). By training a four-state (i.e. local solvation environments) Koopman model with 5 independent 80 ns MD trajectories, they were able to predict the full ionic conductivity and to single out the contribution of each state.

3.3. Electrolyte interfaces

Compared to bulk electrolytes, the complexity of electrolyte interfaces is enormous. On the one hand, this imposes a great challenge for modelling electrolyte interfaces with atomistic modelling; on the other hand, it offers a fantastic opportunity for atomistic machine learning. Indeed, a number of examples have already emerged, focusing on the interfacial stability and dynamics (Figure 4).

Ahmad *et al.* used a GCNN to train and predict the mechanical properties of inorganic solid electrolytes, which is an important factor for stabilizing the interface and suppressing dendrite growth in lithium metal batteries.[146] 2041 crystal structures from the Materials Project database containing elastic properties were used to train the GCNN model, which achieved an RMSE of 0.1268 and 0.1013 log (GPa) for the shear and bulk moduli respectively. Then, using mechanical properties predicted from the GCNN model as inputs and computing the stability parameter $\chi(k)$ which indicates whether the electrodeposition is stable or not, they were able to screen 12950 Li-containing compounds and to confirm that none of the materials in the database can be stabilized without the help of surface tension at $k = 10^8$ m^{-1}. Instead, by taking into account the anisotropy of the elastic tensor of solid electrolytes and the orientation of the lithium metal, they found 20 mechanically stable interfaces in the electrolyte screening.

In addition to the mechanical stability, the thermodynamic stability is another factor which plays an important role in the screening of solid electrolytes for lithium metal

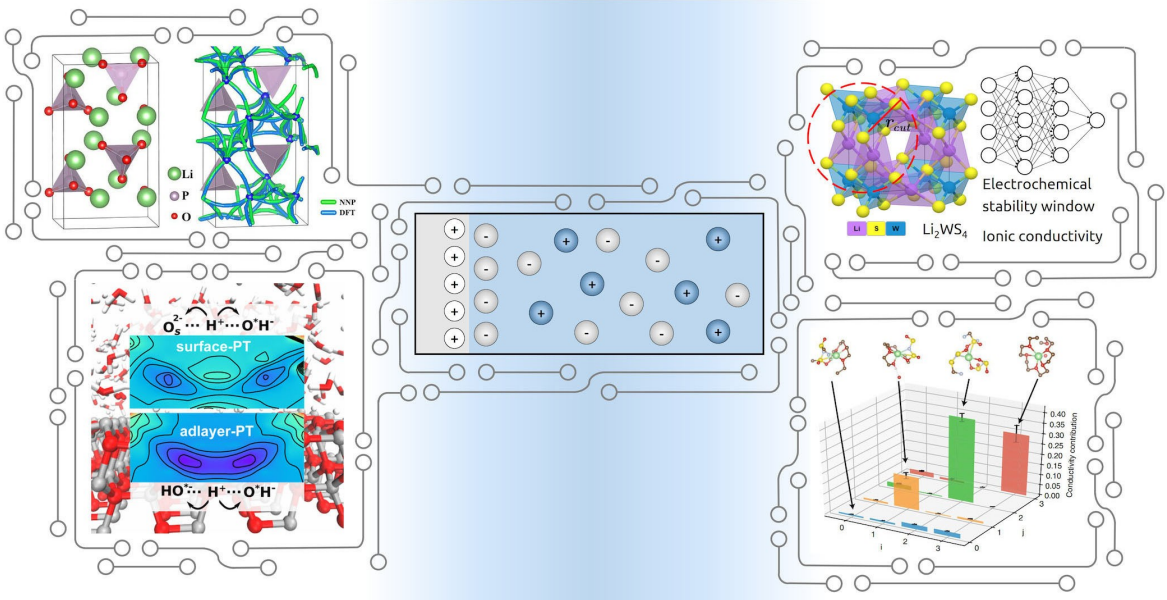


Figure 4. Applications of atomistic machine learning for modelling bulk electrolytes and electrolyte interfaces. Top left: the modelling of diffusion pathways of Li atoms in the amorphous $Li_{12}P_4O_{16}$ model, reproduced from Ref. [134] under the terms of CC-BY licence (Copyright 2017 Ando, Minamitani and Watanabe). Top right: the direct property predictions from neural nets. Bottom left: the modelling of proton-transfer mechanisms at the Water-ZnO interface, reproduced from Ref. [148] with permission (Copyright 2017 American Chemical Society). Bottom right: the prediction of contributions from local environments to the lithium ion conduction in polymer electrolytes, reproduced from Ref. [145] under the terms of CC-BY licence (Copyright 2019 Xie *et al.*).

anodes. Liu *et al.* established an automated workflow using DFT calculations to determine all possible reactions when $Li_7La_3Zr_2O_{12}$ (LLZOM, M = dopant) comes into contact with differing amounts of Lithium and trained a KRR model with 100 data to map 15 structural features (such as the coordination number) to the reaction energy for Li | LLZOM interfaces.[147] With an RMSE of 0.04 eV for the reaction energy, their KRR model predicted 18 unexplored LLZOM systems, which were then validated by DFT calculations.

In these two examples just mentioned, the interface was not modelled explicitly but rather introduced as either orientational or as compositional information. The representation of the interface under experimental conditions calls for the explicit modelling of solid-electrolyte interfaces with atomistic machine learning. In this regard, Natarajan and Behler attempted to develop a machine learning potential for modelling solid-liquid interfaces between water and copper surfaces.[149] In that study, the BPNN was used with SFs as descriptors and DFT reference data which contained 9258 structures of bulk copper, bulk liquid water, ice as well as cuprous oxide and several different ordered and disordered copper/water interfaces. The reference calculations were composed of periodic DFT calculations and DFTMD simulations at 300 to 800 K. Subsequently, this dataset was further refined by including missing structures extracted from machine learning potential based MD simulations. Additional data were also added following the QbC strategy in active learning, as discussed in Section 2.4. This self-consistent training led to a model accuracy with an energy RMSE of 1 meV per atom, and an RMSE for the forces of 125 meV/Å. Note that even earlier Artrith and Kolpak reported a study on bimetallic Au/Cu nanoparticles in aqueous solution with a BPNN-type machine learning potential, where training on energy alone allowed MC simulations of the Au/Cu (211) | water interface.[150] Compared to the clean metal electrode surface, the oxide surface is known to undergo hydroxylation in the presence of adsorbed water molecules. The work from the Quaranta *et al.* on the $ZnO(10\bar{1}0)$ | water interface showed that surface reactions such as proton transfer can also be well captured by a machine learning potential.[148] A similar work for the anatase TiO_2 (101) | water interface was done recently by Calegari Andrade *et al.* with DeepMD.[151]

4. Summary and outlook

In this review, we introduced and summarized the main ingredients of atomistic machine learning for modelling bulk electrolytes and electrolyte interfaces. By discussing recent examples according to their similarities and differences, we aimed to provide a timely snapshot of this emerging area and its connection to the past. The *status quo* focuses on approximating the PES and predicting physico-chemical properties (e.g. electrochemical stability, ionic conductivity, elastic tensor) of inorganic solid electrolytes with supervised learning (See Section 3). Obviously, there are many gaps to be filled on the method development side and lots of new applications to be explored.

Given that the problem of symmetry and uniqueness in descriptor engineering has now been satisfactorily understood in most cases (Section 2.2), further improvement could focus on the interpretability. For instance, the GCNN approach rationalized the model performance by visualizing the connectivity of edge features.[88] It can be expected that similar approaches would eventually help to improve both the physical understanding and the practical performance of atomistic machine learning.

Compared to inorganic solid electrolytes, studies on generating machine learning potentials for liquid electrolytes and electrolyte interfaces are just a handful. This is not really a surprise, since the conformational space of liquid electrolytes and the complexity of electrolyte interfaces are much larger. With the development of open-source codes and new initiatives such as the BIG-MAP (Battery Interface Genome) project,[152] significant progresses on developing machine learning potentials for modelling liquid electrolytes and electrolyte interfaces can be anticipated. However, as already stated in Section 2.2. and 2.3., new generations of machine learning potentials which are able to include self-consistent long-range interactions as well as surface charges at electrified interfaces would be needed to fulfill this purpose on a full-scale. Furthermore, applications to supercapacitors in addition to battery systems are yet to be seen.

The current paradigm of atomistic machine learning of electrolytes focuses on supervised learning. However, unsupervised learning can offer quite a different and complementary perspective in this area. Apart from clustering structures to identify fingerprints of superionics in inorganic solid electrolytes[153] and to distinguish different solvating states in ion conduction from MD trajectories,[154] the low-dimensional representation of coarse-grained dynamics can be found using the variational approach of Markov processes.[155] In addition, high-throughput computational screening of electrolytes will benefit from the development of generative machine learning models,[156] which could propose a broad range of likely candidates beyond human intuition.

Despite the fact that this review and the selected examples lie at the intersection between atomistic modelling and machine learning, it is clear that theoretical works need the company of experimental works in order to crack real-world problems in electrochemical energy storage systems. In fact, machine learning may become another common tongue (in addition to basic theories such as laws of thermodynamics) which both theoreticians and experimentalist speak, because it provides a unified approach for distilling computed and experimental data.

Acknowledgements

The Swedish Research Council (VR) is acknowledged for a starting grant (no. 2019-05012) to C.Z. and a project grant (2019-04824)

to K.H. Funding from the Swedish National Strategic e-Science program eSSENCE is also gratefully acknowledged. L.K. is partly supported by a PhD studentship from the Centre for Interdisciplinary Mathematics (CIM) at Uppsala University.

Conflict of Interest

The authors declare no conflict of interest.

Keywords: Materials Modelling · Machine Learning · Neural Network · Electrolyte · Interface

[1] R. E. Oesper, M. Speter, *Sci. Mon.* **1937**, *45*, 535–546.
[2] J. W. Servos, *Physical Chemistry from Ostwald to Pauling*, Princeton University Press, Princeton, **1996**.
[3] K. Xu, *Chem. Rev.* **2014**, *114*, 11503–11618.
[4] J. Lai, Y. Xing, N. Chen, L. Li, F. Wu, R. Chen, *Angew. Chem. Int. Ed.* **2020**, *59*, 2974–2997.
[5] H. Che, S. Chen, Y. Xie, H. Wang, K. Amine, X.-Z. Liao, Z.-F. Ma, *Energy Environ. Sci.* **2017**, *10*, 1075–1101.
[6] Z. Gao, H. Sun, L. Fu, F. Ye, Y. Zhang, W. Luo, Y. Huang, *Adv. Mater.* **2018**, *30*, 1705702–1705727.
[7] T. Famprikis, P. Canepa, J. A. Dawson, M. S. Islam, C. Masquelier, *Nat. Mater.* **2019**, *18*, 1278–1291.
[8] L. Suo, O. Borodin, T. Gao, M. Olguin, J. Ho, X. Fan, C. Luo, C. Wang, K. Xu, *Science* **2015**, *350*, 938–943.
[9] P. Simon, Y. Gogotsi, *Acc. Chem. Res.* **2013**, *46*, 1094–1103.
[10] C. Merlet, B. Rotenberg, P. A. Madden, P.-L. Taberna, P. Simon, Y. Gogotsi, M. Salanne, *Nat. Mater.* **2012**, *11*, 306–310.
[11] J. B. Goodenough, *Annu. Rev. Mater. Res.* **2003**, *33*, 91–128.
[12] B. E. Conway, *J. Electrochem. Soc.* **1991**, *138*, 1539–1548.
[13] J. Kalhoff, G. G. Eshetu, D. Bresser, S. Passerini, *ChemSusChem* **2015**, *8*, 2154–2175.
[14] J. Mindemark, M. J. Lacey, T. Bowden, D. Brandell, *Prog. Polym. Sci.* **2018**, *81*, 114–143.
[15] D. Zhou, D. Shanmukaraj, A. Tkacheva, M. Armand, G. Wang, *Chem* **2019**, *5*, 2326–2352.
[16] D. R. MacFarlane, N. Tachikawa, M. Forsyth, J. M. Pringle, P. C. Howlett, G. D. Elliott, J. H. Davis, M. Watanabe, P. Simon, A. Angell, *Energy Environ. Sci.* **2014**, *7*, 232–250.
[17] J. Krummacher, C. Schütter, L. H. Hess, A. Balducci, *Curr. Opin. Chem. Eng.* **2018**, *9*, 64–69.
[18] O. Borodin, J. Self, K. A. Persson, C. Wang, K. Xu, *Joule* **2020**, *4*, 69–100.
[19] V. Bocharova, A. P. Sokolov, *Macromolecules* **2020**, *53*, 4141–4157.
[20] Y. Shao, K. Shigenobu, M. Watanabe, C. Zhang, *J. Phys. Chem. B* **2020**, *124*, 4774–4780.
[21] E. Peled, *J. Electrochem. Soc.* **1979**, *126*, 2047.
[22] K. Xu, *Chem. Rev.* **2004**, *104*, 4303–4418.
[23] X. Yu, A. Manthiram, *Energy Environ. Sci.* **2018**, *11*, 527–543.
[24] M. Dirican, C. Yan, P. Zhu, X. Zhang, *Mater. Sci. Eng. R* **2019**, *136*, 27–46.
[25] D. Aurbach, *J. Power Sources* **2000**, *89*, 206–218.
[26] W. Kohn, *Rev. Mod. Phys.* **1999**, *71*, 1253–1266.
[27] R. Car, M. Parrinello, *Phys. Rev. Lett.* **1985**, *55*, 2471–2474.
[28] Y. Li, K. Leung, Y. Qi, *Acc. Chem. Res.* **2016**, *49*, 2363–2370.
[29] S. Shi, J. Gao, Y. Liu, Y. Zhao, Q. Wu, W. Ju, C. Ouyang, R. Xiao, *Chin. Phys. B* **2015**, *25*, 18212–18225.
[30] A. Urban, D.-H. Seo, G. Ceder, *npj Comput. Mater.* **2016**, *2*, 16002.
[31] G. Ceder, S. P. Ong, Y. Wang, *MRS Bull.* **2018**, *43*, 746–751.
[32] A. Groß, *Top. Curr. Chem.* **2018**, *376*, 17.
[33] G. Åvall, J. Mindemark, D. Brandell, P. Johansson, *Adv. Energy Mater.* **2018**, *8*, 1703036.
[34] Y. Tateyama, B. Gao, R. Jalem, J. Haruyama, *Curr. Opin. Chem. Eng.* **2019**, *17*, 149–157.
[35] A. A. Franco, A. Rucci, D. Brandell, C. Frayret, M. Gaberscek, P. Jankowski, P. Johansson, *Chem. Rev.* **2019**, *119*, 4569–4627.
[36] I. E. Castelli, M. Zorko, T. M. Østergaard, P. F. B. D. Martins, P. P. Lopes, B. K. Antonopoulos, F. Maglia, N. M. Markovic, D. Strmcnik, J. Rossmeisl, *Chem. Sci.* **2020**, *11*, 3914–3922.
[37] K. Leung, *Phys. Chem. Chem. Phys.* **2020**, *22*, 10412–10425.
[38] O. Borodin, X. Ren, J. Vatamanu, A. von Wald Cresce, J. Knap, K. Xu, *Acc. Chem. Res.* **2017**, *50*, 2886–2894.
[39] A. Wang, S. Kadam, H. Li, S. Shi, Y. Qi, *npj Comput. Mater.* **2018**, *4*, 15.
[40] J.-P. Correa-Baena, K. Hippalgaonkar, J. van Duren, S. Jaffer, V. R. Chandrasekhar, V. Stevanovic, C. Wadia, S. Guha, T. Buonassisi, *Joule* **2018**, *2*, 1410–1420.
[41] N. Artrith, *J. Phys. Energy* **2019**, *1*, 032002.
[42] A. Van der Ven, Z. Deng, S. Banerjee, S. P. Ong, *Chem. Rev.* **2020**, *120*, 6977–7019.
[43] V. L. Deringer, *J. Phys. Energy* **2020**, *2*, 041003.
[44] Y. Liu, B. Guo, X. Zou, Y. Li, S. Shi, *Energy Storage Mater.* **2020**, *31*, 434–450.
[45] A. L. Samuel, *IBM J. Res. Dev.* **1959**, *3*, 210–229.
[46] S. A.-M. Yaser, M.-I. Malik, L. Hsuan-Tien, *Learning from Data: A Short Course*, AMLbook, **2012**.
[47] K. He, X. Zhang, S. Ren, J. Sun, in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, **2016**, pp. 770–778.
[48] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, **2016**.
[49] D. J. C. MacKay, *Information theory, inference, and learning algorithms*, Cambridge University Press, Cambridge, **2003**.
[50] A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, M. Ceriotti, *Sci. Adv.* **2017**, *3*, e1701816.
[51] A. Aspuru-Guzik, R. Lindh, M. Reiher, *ACS Cent. Sci.* **2018**, *4*, 144–152.
[52] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, *Nature* **2018**, *559*, 547–555.
[53] O. A. von Lilienfeld, K.-R. Müller, A. Tkatchenko, *Nat. Chem. Rev.* **2020**, *4*, 347–358.
[54] A. P. Bartók, M. C. Payne, R. Kondor, G. Csányi, *Phys. Rev. Lett.* **2010**, *104*, 136403.
[55] J. Behler, M. Parrinello, *Phys. Rev. Lett.* **2007**, *98*, 146401.
[56] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, *npj Comput. Mater.* **2017**, *3*, 54.
[57] J. Schmidt, M. R. G. Marques, S. Botti, M. A. L. Marques, *npj Comput. Mater.* **2019**, *5*, 83.
[58] N. E. Jackson, M. A. Webb, J. J. de Pablo, *Curr. Opin. Chem. Eng.* **2019**, *23*, 106–114.
[59] G. Fasshauer, M. McCourt, *Kernel-based Approximation Methods using MATLAB*, Vol. 19 of *Interdisciplinary Mathematical Sciences*, World Scientific, **2015**.
[60] B. Huang, N. O. Symonds, O. A. von Lilienfeld in *Handbook of Materials Modeling*, Springer, Cham, **2018**.
[61] M. A. Collins, D. F. Parsons, *J. Chem. Phys.* **1993**, *99*, 6756–6772.
[62] H. Gassner, M. Probst, A. Lauenstein, K. Hermansson, *J. Phys. Chem. A* **1998**, *102*, 4596–4605.
[63] Z. Xie, J. M. Bowman, *J. Chem. Theory Comput.* **2010**, *6*, 26–34.
[64] M. Rupp, A. Tkatchenko, K.-R. Müller, O. A. von Lilienfeld, *Phys. Rev. Lett.* **2012**, *108*, 058301.
[65] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller, A. Tkatchenko, *J. Phys. Chem. Lett.* **2015**, *6*, 2326–2331.
[66] E. Prodans, W. Kohn, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 11635–11638.
[67] H. Huo, M. Rupp, *arXiv preprint* **2017**.
[68] F. A. Faber, A. S. Christensen, B. Huang, O. A. von Lilienfeld, *J. Chem. Phys.* **2018**, *148*, 241717.
[69] C. van der Oord, G. Dusson, G. Csányi, C. Ortner, *Mach. Learn.: Sci. Technol.* **2020**, *1*, 015004.
[70] A. V. Shapeev, *Multiscale Model. Simul.* **2016**, *14*, 1153–1173.
[71] M. G. Darley, C. M. Handley, P. L. A. Popelier, *J. Chem. Theory Comput.* **2008**, *4*, 1435–1448.
[72] L. Zhang, J. Han, H. Wang, R. Car, W. E, *Phys. Rev. Lett.* **2018**, *120*, 143001.
[73] A. P. Thompson, L. P. Swiler, C. R. Trott, S. M. Foiles, G. J. Tucker, *J. Comput. Phys.* **2015**, *285*, 316–330.
[74] A. P. Bartók, R. Kondor, G. Csányi, *Phys. Rev. B* **2013**, *87*, 184115.
[75] M. J. Willatt, F. Musil, M. Ceriotti, *J. Chem. Phys.* **2019**, *150*, 154110.
[76] R. Drautz, *Phys. Rev. B* **2019**, *99*, 014104.
[77] L. Belloni, D. Borgis, M. Levesque, *J. Phys. Chem. Lett.* **2018**, *9*, 1985–1989.

[78] H.-P. Komsa, T. T. Rantala, A. Pasquarello, *Phys. Rev. B* **2012**, *86*, 045112.
[79] H. Oberhofer, K. Reuter, J. Blumberger, *Chem. Rev.* **2017**, *117*, 10319–10357.
[80] M. Finnis, *Interatomic Forces in Condensed Matter*, Oxford University Press, Oxford, **2003**.
[81] N. Artrith, T. Morawietz, J. Behler, *Phys. Rev. B* **2011**, *83*, 153101.
[82] S. A. Ghasemi, A. Hofstetter, S. Saha, S. Goedecker, *Phys. Rev. B* **2015**, *92*, 045131.
[83] X. Xie, K. A. Persson, D. W. Small, *J. Chem. Theory Comput.* **2020**, *16*, 4256–4270.
[84] A. Grisafi, M. Ceriotti, *J. Chem. Phys.* **2019**, *151*, 204105.
[85] Y. Lecun, Y. Bengio, G. Hinton, *Nature* **2015**, *521*, 436–444.
[86] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, K.-R. Müller, *J. Chem. Phys.* **2018**, *148*, 241722.
[87] C. Chen, W. Ye, Y. Zuo, C. Zheng, S. P. Ong, *Chem. Mater.* **2019**, *31*, 3564–3572.
[88] Y. Shao, M. Hellström, P. D. Mitev, L. Knijff, C. Zhang, *J. Chem. Inf. Model.* **2020**, *60*, 1184–1193.
[89] K. Xu, W. Hu, J. Leskovec, S. Jegelka, *arXiv preprint* **2018**.
[90] L. Himanen, M. O. J. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, A. S. Foster, *Comput. Phys. Commun.* **2020**, *247*, 106949.
[91] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, O. A. von Lilienfeld, *New J. Phys.* **2013**, *15*, 095003.
[92] C. Schran, J. Behler, D. Marx, *J. Chem. Theory Comput.* **2020**, *16*, 88–99.
[93] K. T. Schütt, P.-J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko, K.-R. Müller, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, (Eds: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett), Curran Associates, Inc., **2017**, p. 9921002.
[94] M. F. Langer, A. Goeßmann, M. Rupp, *arXiv preprint* **2020**.
[95] C. Zhang, *J. Chem. Phys.* **2018**, *149*, 031103.
[96] C. Zhang, J. Hutter, M. Sprik, *J. Phys. Chem. Lett.* **2019**, *10*, 3871–3876.
[97] T. Sayer, M. Sprik, C. Zhang, *J. Chem. Phys.* **2019**, *150*, 041716.
[98] T. Dufils, G. Jeanmairet, B. Rotenberg, M. Sprik, M. Salanne, *Phys. Rev. Lett.* **2019**, *123*, 195501.
[99] C. Zhang, T. Sayer, J. Hutter, M. Sprik, *J. Phys. Energy* **2020**, *2*, 032005.
[100] M. Stengel, N. A. Spaldin, D. Vanderbilt, *Nat. Phys.* **2009**, *5*, 304–308.
[101] R. Resta, D. Vanderbilt in *Physics of Ferroelectrics: a Modern Perspective*, Springer, Berlin, **2007**, pp. 31–68.
[102] N. Marzari, A. A. Mostofi, J. R. Yates, I. Souza, D. Vanderbilt, *Rev. Mod. Phys.* **2012**, *84*, 1419–1475.
[103] F. Jensen, *Introduction to computational chemistry*, John Wiley & Sons, Chichester, **2007**.
[104] B. K. Rai, G. A. Bakken, *J. Comput. Chem.* **2013**, *34*, 1661–1671.
[105] P. Bleiziffer, K. Schaller, S. Riniker, *J. Chem. Inf. Model.* **2018**, *58*, 579–590.
[106] L. Zhang, M Chen, X. Wu, H. Wang, W. E. R. Car, *Phys. Rev. B* **2020**, *102*, 041121.
[107] F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke, K.-R. Müller, *Nat. Commun.* **2017**, *8*, 872.
[108] A. Grisafi, A. Fabrizio, B. Meyer, D. M. Wilkins, C. Corminboeuf, M. Ceriotti, *ACS Cent. Sci.* **2018**, *5*, 57–64.
[109] A. Chandrasekaran, D. Kamal, R. Batra, C. Kim, L. Chen, R. Ramprasad, *npj Comput. Mater.* **2019**, *5*, 22.
[110] S. Gong, T. Xie, T. Zhu, S. Wang, E. R. Fadel, Y. Li, J. C. Grossman, *Phys. Rev. B* **2019**, *100*, 184103.
[111] B. Settles, *Active Learning; volume 18 of Synthesis Lectures on Artiflcial Intelligence and Machine Learning*, Morgan & Clay- pool Publishers., San Rafael, **2012**.
[112] A. De Vita, R. Car, *MRS Online Proc. Libr.* **1997**, *491*, 473.
[113] G. Csányi, T. Albaret, M. C. Payne, A. De Vita, *Phys. Rev. Lett.* **2004**, *93*, 175503.
[114] V. Botu, R. Ramprasad, *Int. J. Quantum Chem.* **2015**, *115*, 1074–1083.
[115] L. M. Raff, M. Malshe, M. Hagan, D. I. Doughan, M. G. Rockley, R. Komanduri, *J. Chem. Phys.* **2005**, *122*, 84104.
[116] E. V. Podryabinkin, A. V. Shapeev, *Comput. Mater. Sci.* **2017**, *140*, 171–180.
[117] N. Bernstein, G. Csányi, V. L. Deringer, *npj Comput. Mater.* **2019**, *5*, 99.
[118] M. Ceriotti, G. A. Tribello, M. Parrinello, *J. Chem. Theory Comput.* **2013**, *9*, 1521–1532.
[119] V. Botu, R. Batra, J. Chapman, R. Ramprasad, *J. Phys. Chem. C* **2017**, *121*, 511–522.
[120] E. Uteva, R. S. Graham, R. D. Wilkinson, R. J. Wheatley, *J. Chem. Phys.* **2018**, *149*, 174114.
[121] R. Jinnouchi, F. Karsai, G. Kresse, *Phys. Rev. B* **2019**, *100*, 014105.
[122] J. Vandermause, S. B. Torrisi, S. Batzner, Y. Xie, L. Sun, A. M. Kolpak, B. Kozinsky, *npj Comput. Mater.* **2020**, *6*, 20.
[123] N. Artrith, J. Behler, *Phys. Rev. B* **2012**, *85*, 45413–45439.
[124] M. Gastegger, J. Behler, P. Marquetand, *Chem. Sci.* **2017**, *8*, 6924–6935.
[125] J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, A. E. Roitberg, *J. Chem. Phys.* **2018**, *148*, 241733.
[126] J. P. Janet, C. Duan, T. Yang, A. Nandy, H. J. Kulik, *Chem. Sci.* **2019**, *10*, 7913–7922.
[127] M. Korth, *Phys. Chem. Chem. Phys.* **2014**, *16*, 7919–7926.
[128] O. Borodin, M. Olguin, C. E. Spear, K. W. Leiter, J. Knap, *Nanotechnology* **2015**, *26*, 354003.
[129] C. Lian, H. Liu, C. Li, J. Wu, *AIChE J.* **2019**, *65*, 804–810.
[130] W. Wang, T. Yang, W. H. Harris, R. Gómez-Bombarelli, *Chem. Commun.* **2020**, *56*, 8920–8923.
[131] M. Hellström, J. Behler, *J. Phys. Chem. Lett.* **2016**, *7*, 3302–3306.
[132] Y. Shao, M. Hellström, A. Yllö, J. Mindemark, K. Hermansson, J. Behler, C. Zhang, *Phys. Chem. Chem. Phys.* **2020**, *22*, 10426–10430.
[133] S. Tovey, A. N. Krishnamoorthy, G. Sivaraman, J. Guo, C. Benmore, A. Heuer, C. Holm, *J. Phys. Chem. C* **2020**, *124*, 25760–25768.
[134] W. Li, Y. Ando, E. Minamitani, S. Watanabe, *J. Chem. Phys.* **2017**, *147*, 214106.
[135] W. Li, Y. Ando, S. Watanabe, *J. Phys. Soc. Jpn.* **2017**, *86*, 104004.
[136] C. Wang, K. Aoyagi, P. Wisesa, T. Mueller, *Chem. Mater.* **2020**, *32*, 3741–3752.
[137] A. Marcolongo, T. Binninger, F. Zipoli, T. Laino, *ChemSystemsChem* **2020**, *2*, e1900031.
[138] J. Huang, L. Zhang, H. Wang, J. Zhao, J. Cheng, W. E, *arXiv preprint* **2020**.
[139] M. Mińkowski, K. Hummer, C. Dellago, *J. Phys. Condens. Matter* **2020**, *33*, 015901.
[140] S.-H. Guan, C. Shang, Z.-P. Liu, *J. Phys. Chem. C* **2020**, *124*, 15085–15093.
[141] N. A. Katcho, J. Carrete, M. Reynaud, G. Rousse, M. Casas-Cabanas, N. Mingo, J. Rodríguez-Carvajal, J. Carrasco, *J. Appl. Crystallogr.* **2019**, *52*, 148–157.
[142] S. J. Pan, Q. Yang, *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359.
[143] E. D. Cubuk, A. D. Sendek, E. J. Reed, *J. Chem. Phys.* **2019**, *150*, 214701.
[144] K. Hatakeyama-Sato, T. Tezuka, M. Umeki, K. Oyaizu, *J. Am. Chem. Soc.* **2020**, *142*, 3301–3305.
[145] T. Xie, A. France-Lanord, Y. Wang, Y. Shao-Horn, J. C. Grossman, *Nat. Commun.* **2019**, *10*, 2667.
[146] Z. Ahmad, T. Xie, C. Maheshwari, J. C. Grossman, V. Viswanathan, *ACS Cent. Sci.* **2018**, *4*, 996–1006.
[147] B. Liu, J. Yang, H. Yang, C. Ye, Y. Mao, J. Wang, S. Shi, J. Yang, W. Zhang, *J. Mater. Chem. A* **2019**, *7*, 19961–19969.
[148] V. Quaranta, M. Hellström, J. Behler, *J. Phys. Chem. Lett.* **2017**, *8*, 1476–1483.
[149] S. K. Natarajan, J. Behler, *Phys. Chem. Chem. Phys.* **2016**, *18*, 28704–28725.
[150] N. Artrith, A. M. Kolpak, *Nano Lett.* **2014**, *14*, 2670–2676.
[151] M. F. Calegari Andrade, H.-P. Ko, L. Zhang, R. Car A Selloni, *Chem. Sci.* **2020**, *11*, 2335–2341.
[152] A. Bhowmik, I. E. Castelli, J. M. Garcia-Lastra, P. B. Jørgensen, O. Winther, T. Vegge, *Energy Storage Mater.* **2019**, *21*, 446–456.
[153] Y. Zhang, X. He, Z. Chen, Q. Bai, A. M. Nolan, C. A. Roberts, D. Banerjee, T. Matsunaga, Y. Mo, C. Ling, *Nat. Commun.* **2019**, *10*, 5260.
[154] I. B. Magdău, T. F. Miller, *ChemRxiv* **2020**.
[155] F. Noé, A. Tkatchenko, K.-R. Müller, C. Clementi, *Annu. Rev. Phys. Chem.* **2020**, *71*, 361–390.
[156] B. Sanchez-Lengeling, A. Aspuru-Guzik, *Science* **2018**, *361*, 360–365.

Manuscript received: November 5, 2020
Revised manuscript received: December 9, 2020
Accepted manuscript online: December 9, 2020
Version of record online: January 4, 2021