Applied Machine Learning for Developing Next-Generation Functional Materials

*Filip Dinic, Kamalpreet Singh, Tony Dong, Milad Rezazadeh, Zhibo Wang, Ali Khosrozadeh, Tiange Yuan, and Oleksandr Voznyy**

Machine learning (ML) is a versatile technique to rapidly and efficiently generate insights from multidimensional data. It offers a much-needed avenue to accelerate the exploration and investigation of new materials to address time-sensitive global challenges such as climate change. The availability of large datasets in recent years has enabled the development of ML algorithms for various applications including experimental/device optimization and material discovery. This perspective provides a summary of the recent applications of ML in material discovery in a range of fields, from optoelectronics to batteries and electrocatalysis, as well as an overview of the methods behind these advances. The paper also attempts to summarize some key challenges and trends in current research methodologies.

1. Introduction

We now live in the era of data. Our daily interactions, from the groceries we buy to the music we listen to, all generate massive amounts of data that can be collected and analyzed to extract insights, make predictions about future behaviors, and thereafter be used to make business decisions to capitalize on the data. The explosion in data availability has led to rapid development of machine learning (ML) techniques—a series of scalable statistical algorithms for multidimensional data analysis, such as interpolation and extrapolation of existing data to new examples.[1]

Now the experts in other domains can leverage the power of ML in their own fields too. One such exciting application of ML is the design and optimization of materials, particularly those enabling society's transition toward sustainable energy (e.g., batteries, fuel cells, photovoltaics, etc). In solving this time-sensitive design problem, it is becoming increasingly evident that traditional trial and error methods are simply too inefficient, given the vastness of the chemical combinatorial space, where human intuition is often inadequate to capture the trends in materials properties. In fact, today we have only explored $\approx 10^6$ crystalline materials and $\approx 10^9$ molecules while there are estimated to be at least 10^{60} possible small organic molecules alone.[2]

With the development of cheminformatics tools, it is now easy to represent any molecule or material uniquely by a row of numbers in a way computers can understand. ML operates on these representations and finds the attributes important to the behavior/properties/performance of known materials, and consequently enables exploration of new chemical spaces. Even imperfect predictions can inform scientists of where to dedicate computational or experimental resources to maximize the probability of making an advancement. The results of these explorations can be employed in further training and refining the model.

In this perspective, we describe recent applications of ML in the development of materials for energy-related applications. Primarily targeting an experimentalist reader, we provide an analysis on the effectiveness of current methods as well as their shortcomings. Both the "theoretical" approaches (based on computed data) and analysis/optimization of experimental data are discussed. We conclude by offering our perspective on ways to address the current challenges associated with applying ML in materials design and directions for future development.

2. How Machine Learning Works

ML, in essence, is a function fitting (regression) endeavor based on several known values of a function at some input points (training data). ML approaches have multiple flavors, depending on the type of data utilized and the reward system. This includes supervised learning, unsupervised learning, reinforcement learning, generative models, etc. In supervised learning, data needs to be labeled first, which then can be used to find the relationship between the inputs and the labels. Known relationship can then be used to predict the target property for new inputs. Examples include predicting properties of materials such as formation energy and bandgaps. Within supervised learning, there exist two broad types, regression and classification. Regression modeling involves predicting a variable value over a continuous spectrum, for example, bandgap, elastic modulus, formation energy, etc. Classification involves categorizing a variable, for example, classifying materials into metallic or non-metallic.

F. Dinic, K. Singh, T. Dong, M. Rezazadeh, Z. Wang, A. Khosrozadeh, T. Yuan, O. Voznyy
Department of Physical and Environmental Sciences
Department of Chemistry
University of Toronto
Toronto M1C 1A4, Canada
E-mail: o.voznyy@utoronto.ca

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/adfm.202104195.

DOI: 10.1002/adfm.202104195

Unsupervised learning analyzes unlabeled data and focuses on grouping and finding existing patterns or similarities in the data. Reinforcement learning involves optimizing a models ability to complete a task via a set of decisions. This is done by rewarding or penalizing the model's decisions based on the actions it has performed. Within material science at this time, supervised learning is most commonly used.

To investigate complex data, such as images or crystal structures, the data must be first encoded as a list of numbers to make them amenable to ML analysis. For example, an image may be encoded using a list of pixel colors and brightness values, yielding a multidimensional input that can be fed into the regression machinery to provide a (multidimensional) numerical output of a function, such as 0 for a cat, 1 for a human, 2 for a table, and so on.

The number of inputs and outputs (dimensionality of the data) is often significantly larger than one (sometimes thousands), and the available data points are not necessarily located on a regular grid, making it difficult to interpolate using conventional methods.

Given that the analytical form of the desired function is usually not known, the method resorts to fitting the data locally using a spline-like interpolation in between neighboring data points. Statistical comparison to a hidden subset of data (validation dataset) allows one to choose the stiffness and smoothness of these splines that minimizes the prediction error.

Such locality highlights the prescient problem of ML: its inability to generalize, that is, extrapolate into the regions of parameter space where no training data is available. A straightforward approach to deal with this problem is to either provide more data or to use some stiffer, longer-ranged splines (or splines whose functional form follows the known physical law for the problem at hand).

A more sophisticated approach to minimize ML's shortcomings is to re-encode the inputs by exploiting some domain knowledge and in such a way reduce the dimensionality of the parameter space. For example, fitting a parabola can require an infinite number of adjustable parameters (splines) between all pairs of points, or just two constants a and b for $f(x) = ax^2 + b$ if the functional form is known. In "symbolic learning" one can even learn which functional form to use.[3]

In another example, an infinitely large image can be split into smaller blocks of pixels, that are analyzed independently, but using the same shared set of parameters (so-called convolutions), thus reducing the overall number of free parameters. For each block, the presence of some features (e.g., vertical or horizontal lines) can be extracted, compressing the image into a lower-dimensional representation. In both examples, dimensionality reduction allows for the number of adjustable parameters to be decreased, thus requiring fewer data points to obtain a good fit. In the presence of infinite data, ML turns into a basic statistical analysis that allows compressing all data into a semi-analytical representation and interpolating between known data points. However, data is always scarce (compared to the dimensionality of a problem at hand). For example, even with high-throughput experiments or computations, data acquisition is too slow and cannot collect sufficiently rich and diverse training dataset. New developments in the field of ML are constantly being developed aiming to make more accurate predictions in the regime of low data. All of them, in one form or another, are variations of the above approaches—embedding the domain knowledge or applying the same transformation to different blocks of data.

One such technique is transfer learning, also known as few-shot learning, where domain knowledge (e.g., understanding of what an image is and what features it can have) is embedded by pretraining the model in the regime of big data, and then this model is fine-tuned to analyze some new features on a smaller dataset (e.g., images of chest X-rays). Similar in concept but with somewhat different implementation is data fusion,[4] where additional domain knowledge is embedded by combining different but complementary sources of data to create a larger or richer training dataset.

3. Parameter Space Optimization and Design of Experiments

Traditional materials research, much like other scientific domains, heavily relies on parameter exploration/optimization. In small parameter spaces, a comprehensive sweep (grid search) is typically affordable. For large parameter spaces, optimization generally involves modifying variables one at a time and following the local gradient until an optimum is achieved. This procedure quickly becomes prohibitive for a larger number of parameters and does not guarantee finding a global minimum or maximum. Noise in experimental measurements further undermines the reliability of such procedure.

The development of robust open-source libraries[5] and free online services[6] have increased the accessibility of techniques such as Bayesian optimization. These methods offer experimentalists the ability to "objectively" evaluate past data and efficiently explore the parameter space to optimize the desired objective, for example, maximize electrolyte conductivity, electrode voltage, solar cell performance, etc.

In Bayesian optimization, just like in conventional ML regression, all available data points are interpolated with the aid of some splines (often Gaussians), creating a map of the parameter space (**Figure 1**a).[7–10] A notable addition is that the confidence of the predictions is also estimated at each point. The parameter space is initially treated as a random function with a prior probability distribution, popular method being Gaussian process. This probability distribution acts as the confidence interval for any given interpolated point. An acquisition function then can be derived from this prior, aiming to minimize the variance for the model at large, by proposing new locations in which data can restrict the model the best. Regions of the parameter space far from the already explored data points have larger uncertainty, indicating that exploration there may be fruitful, as they might contain the desired optimum.

Overall, Bayesian optimization is mathematically proven to be a more efficient search strategy.[5] Oftentimes, less than 100 experiments are required for finding an optimum in 5D–6D spaces, as opposed to ≈100 000 data points in a rigorous grid search. Packages such as Phoenics[7] allow for relatively easy deployment and integration of Bayesian optimization into existing workflows making it ever more accessible to users.

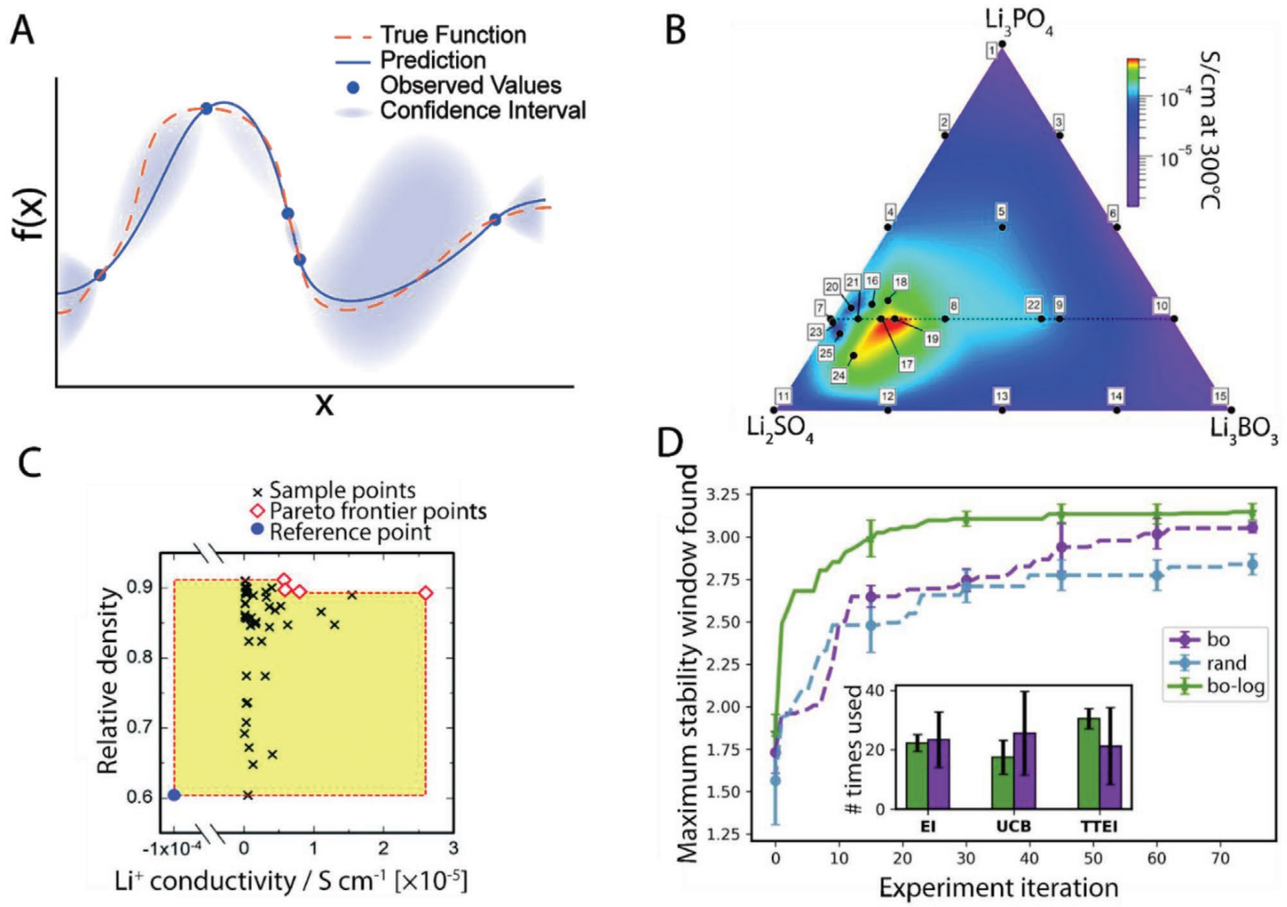


Figure 1. a) Overview of Bayesian optimization; a prediction for the true underlying function is made based on observed values along with the uncertainty of the current predictions. b) Ternary component contour map of Li-ion conductivity with labeled points representing proposed experiments by Bayesian optimization. Reproduced with permission.[11] Copyright 2020, American Chemical Society. c) A plot of relative density versus Li-ion conductivity with a distribution of experiments proposed by multi-objective Bayesian optimization and the associated pareto frontier points. Reproduced with permission.[12] Copyright 2020, Royal Society of Chemistry. d) Improvement to maximum electrolyte stability window as a function of experimental iterations; Comparisons between Bayesian optimization and random sampling. Reproduced with permission.[13] Copyright 2020, Elsevier.

3.1. Compositional Space

One of the possible applications of Bayesian optimization is the exploration of a compositional space to improve desired properties. For example, Homma, et al. have applied Bayesian optimization to tune the ionic conductivity of a ternary solid electrolyte system ($Li_3PO_4/Li_3BO_3/Li_2SO_4$). They explored the continuous 3D search space spanned by the ratios of each constituent (Figure 1b) and followed the experimental plan devised by the algorithm. With only 25 experiments (out of totally 5151 possible tests), 15 of which served as initial training data, they derived a composition that gave an ionic conductivity of 4.9×10^{-4} S cm^{-1} (300 °C), which was threefold higher than any binary forms of the same three materials.[11]

Another study by Harada, et al. sought to optimize the ionic conductivity of a known material system, the CaO- and Y_2O_3-co-doped NASICON-type $LiZr_2(PO_4)_3$.[12] Apart from demonstrating the benefit of Bayesian optimization, they showcased the promise of multi-objective Bayesian optimization (MOBO). In this co-doped system, the relationship between the desired mechanical and phase stabilities and the conductivity was shown to be difficult to navigate by intuition. To tackle this issue, MOBO was used to optimize the pareto frontier—a set of points where the improvement of a particular objective cannot be made without sacrificing the other—of two objectives (relative density and ionic conductivity) (Figure 1c). MOBO was demonstrated to be more effective than a random search even on a small experimental sample space (47 samples), and its efficiency was shown to greatly improve in larger spaces.

Bayesian optimization has also been applied to the development of aqueous electrolytes for batteries. In a recent example, a robotic platform coupled with a Bayesian optimizer was used to explore optimal combinations of mixed-anion aqueous electrolytes for use in Li- and Na-ion batteries.[13] The chemical space consisted of nitrates, sulfates, halides, and perchlorates of Li and Na. The conductivity, pH, and electrochemical stability of all different combinations were evaluated and optimized continuously without human intervention (Figure 1d). Using this approach, an optimal blend of two saturated aqueous electrolytes (6.7 mL $NaClO_4$ with 0.3 mL $NaNO_3$) was found, with an improvement in the electrochemical stability window relative to the $NaClO_4$ baseline.

Using Bayesian optimization and employing data fusion of experimental and density functional theory (DFT) data, Sun, et al. searched for a stable alloyed organic–inorganic perovskite composition.[4] The $Cs_xMA_yFA_{1-x-y}PbI_3$ (MA = methylammonium, FA = formamidinium) was subjected to humidity, heat, and light stress tests, with their degradation monitored via camera and processed. DFT Gibbs free energy data allowed to eliminate portions of the compositional space deemed unstable, decreasing the search space where experimental tests should be employed. This approach revealed a formamidinium-rich Cs-poor composition ($Cs_{0.17}MA_{0.03}FA_{0.8}PbI_3$) exhibiting >17*x* stability increase over pure $MAPbI_3$ while only sampling 1.8% of the total search space.

3.2. Device Fabrication

Device optimization is another domain where ML can be used to improve the performance, layer composition, layer thickness and even testing conditions. As with many cases of applied ML, the key challenge remains in the acquisition of a dataset.

For organic light-emitting diodes (OLEDs), modifications to a device layout, such as material interface and thickness, can drastically change the photon yields of the device. By ranking the structural features (band structures, layer thicknesses) by the impact on the device performance (current, power, and quantum efficiencies), the optimal device structure can be achieved with greater efficiency. In the work by Janai, et al., a multivariate regression model was built using random forests to correlate the device efficiency with input parameters.[14] The surrogate model was then used to sample new input combinations that maximize the device efficiency, allowing for a significantly reduced search space for further experiments.

In the realm of organic photovoltaics, David, et al. used ML to investigate factors affecting stability. They fabricated ≈1900 experimental devices with various materials for transport and active layers and investigated the devices under several testing conditions. The model was able to elucidate the most detrimental and beneficial device elements, an impressive feat as the device parameters (transport/active layer composition/thickness) are often convoluted and, therefore, the individual contributions are challenging to pinpoint. Among the different materials that were studied, the best transport layer to improve stability was found to be bathophenanthroline while the worst was PEDOT:PSS.[15]

ML has also been used to optimize the bulk heterojunction in the active layer of solar cells as demonstrated by Kirkley, et al.[10] Aiming to improve the power conversion efficiency (PCE), the authors optimized the donor fraction, solution concentration, annealing time, and temperature. With this model, they were able to more effectively optimize the active layer, through two rounds of experiments. After the first round, a rough map of the impact of various variables on the PCE was established. The second round was then used to fine-tune the peaks, increasing the data point density in these regions.[10]

Coupling ML with high-throughput experiments can be used to produce a more accurate and relevant dataset and allow for a more thorough exploration of the parameter subspace. Du, et al. created a fully automated system, capable of fabricating every aspect of the solar cell, from deposition, annealing, and modifying up to 100 various parameters. The automated system could also characterize the material by photographing the layers, obtaining UV–vis measurements, and current density versus voltage measurements. The trained ML model provided a better understanding of how various parameters influence each other with respect to device performance and stability. For instance, changing the amount of ordered phase of the donor material was linked to higher power conversion efficiency, while spin-coating speed and lower annealing temperature were found to have the biggest impact on device stability.[16]

Inkjet printing is a common method for easily depositing device layers for such high-throughput testing. The rapid deposition of small droplets allows one to rapidly populate a substrate using low volumes of reagents. However, in order to obtain consistent results, the printer must be able to supply similarly consistent sized droplets arranged in a way to maintain suitable yield. By tuning printer settings of jetting pressure, valve actuation rate, and the movement speed of the nozzle, the printer controls the volume, deposition rate, and spread of the droplets in question. Siemenn, et al. further applied image recognition to identify droplet fitness to gauge the optimal deposition parameters for testing.[17] To minimize convergence time for such methods, Bayesian optimization was found to proceed at twice the rate compared to stochastic gradient descent methods in finding the optimal printer settings for the experiment. By shortening this time, new compositions can thus be tested more rapidly and with fewer wasted materials during testing.

3.3. Materials Synthesis

Materials synthesis is another area amenable to ML-guided improvement since even the simplest synthetic pathway depends on many variables. Voznyy, et al. used Bayesian optimization to improve PbS quantum dots synthesis, focusing on monodispersity (represented by the full width at half maximum of the first exciton peak) as a target for each nanoparticle size.[18] The ML model was trained on 2000 experimental data points digitized from old lab notebooks (most being repeats) to create a continuous map of the parameter space. This enabled a reduction in experimental noise and highlighted the quantitative difference between the effect of precursor concentration versus reaction temperature, both of which have qualitatively similar effects on quantum dot size. The process also found that the addition of oleylamine amplified the effect of concentration and improved the monodispersity even further. The procedure yielded significant improvements in monodispersity for a range of quantum dot sizes despite this exact synthesis being systematically optimized by many research groups for the past 20 years.

The example above highlights the importance of failed experiments, which form the majority of data used for training ML algorithms. Synthetic conditions that result in poor monodispersity were never published but were duly recorded in lab notebooks. This impedes the utility of ML when using external data (for example, by scraping from existing publications) since the majority of data is missing.[19,20] This point is reinforced in the work of Raccuglia, et al. who created an ML model for predicting reaction success for vanadium selenites, heavily relying on the data from failed experiments.[20] One option to combat such bias is presented by Li, et al. by combining rapid, randomized, experimental testing via robotic acceleration.[19] As the combinations of reagents were randomized, some reactions were bound to fail but were instructive for training the algorithm. In contrast, a human researcher tries to avoid failed experiments at all costs. By following data-driven principles, human biases in experimental design can be avoided and random errors inherent to any experiment can be taken into account when evaluating the next appropriate trial.

While an ML guided retrospection of experimental work offers an avenue to missed insights, many research groups are now looking to leverage the combined efficiency of robotics and machine learning to automate future material discovery and synthesis.[4,19,21–25] For example, Chan, et al. have harnessed the power of robotics to synthesize 8172 different metal halide perovskites based on 45 different organic ammonium cations using inverse temperature crystallization (ITC).[19] This robot-accelerated perovskite investigation and discovery ("RAPID"), led to a fivefold increase in the number of metal halide perovskites synthesizable via ITC and was used to train an ML model capable of predicting likelihood of single crystal formation of such perovskites for future synthesis.

Similar works have also been conducted for other material classes/applications such as polyoxometalates,[21] nanoparticles,[22] thin films,[23] piezoelectrics,[24] and photocatalysts.[25] As highlighted above, this combination of robotics and machine learning bypasses human bias and labor limitations, leading to objective and rapid material discovery/synthesis. The major current limitation of this approach is the need for specialized equipment and personnel training to program such instruments, making accessibility a barrier to adoption. That being said, accessibility is expected to improve in the near future as equipment becomes more modular and widely available.

4. Automated Processing of Experimental Data

One of the emerging utilities of ML for materials research is the analysis of complex experimental data. Not only can labor-intensive data processing procedures be automated, valuable insights and patterns that are not intuitive to humans can also be potentially captured and exploited.

4.1. Image Recognition

In the work by Jiang, et al., ML served as a valuable tool to automate a labor-intensive image processing task, providing valuable insights about composite cathode design.[26] A Li-ion battery cathode is composed of the active material, binder, and conductive carbon additive. Understanding how the electrode microstructure evolves and potentially fails during the cycling of a battery is paramount to the development of more robust cathodes.

The 3D microstructure of a NiMnCo cathode was visualized through phase-contrast X-ray nano-tomography, and models were developed to calculate the degree of detachment, spatial heterogeneity, and electrical conductivity of active particles. To achieve results with statistical robustness, a large number of active particles needs to be analyzed within a single image. However, this is challenged by the labor intensiveness and failure of traditional algorithms in the identification and segmentation of a large number of particles (>650). As an effective alternative to solve the instance segmentation problem (i.e., to separate different objects in an image), a state-of-the-art mask convolutional neural network was used (Mask R-CNN).[26] By transfer-learning from the parameters of well-established ML models for computer vision, the neural network was efficiently optimized, and accurate active particle segmentation was achieved (**Figure 2**a). Thanks to the accuracy and efficiency of the ML model, meaningful relationship models between the detachment of active particles and cycling protocols (charge–discharge rates) were established.

Computer vision can also be applied for monitoring the progress of chemical reactions in standard glassware.[28–30] Exploiting developments in computer vision, it is possible to identify and monitor separate reaction vessels from a video feed.[29] From an image processing database, a model can learn to distinguish the visual boundaries between containers, their contents, and the phases within.[28,29] This potentially allows for autonomous reaction handling by monitoring for reaction completion via phase changes or separations. The ability to distinguish individual reaction vessels would also reduce the number of monitoring points needed.[29]

A similar approach has also been demonstrated for detecting anomalies and defects such as dust, scratches, circuit faults, etc. during the manufacturing of display panels. Using a simple webcam and image classification, Nguyen, et al. trained the model to separate severe defects from merely cleanable debris.[31]

ML image recognition has also been used to see whether crystals were formed in 96-well plates in a high-throughput setup for perovskite composition optimization.[27] Images of the wells were used to create a classifier, capable of differentiating between images of no crystals, a single crystal, or multiple crystals (Figure 2b). The system was also able to determine the emission wavelength and the relative brightness of the crystals. This data was then used with a second Bayesian ML model combined with the parameters from each experiment to guide a high-throughput search aiming to optimize a new perovskite material, eventually finding a new composition, $(3\text{-PLA})_2PbCl_4$, capable of blue emission.[27]

4.2. Analysis of Spectra

Several research applications rely on X-ray diffraction (XRD) as a quick technique to gather information about the material structure. However, data analysis often becomes a bottleneck in this process, especially for new materials not listed in existing databases. Sun, et al. proposed a classifier based on ML to determine the space group of the synthesized optoelectronic materials.[32] The ML model was able to classify the experimentally obtained XRD spectra into three categories of 0D, 1D,

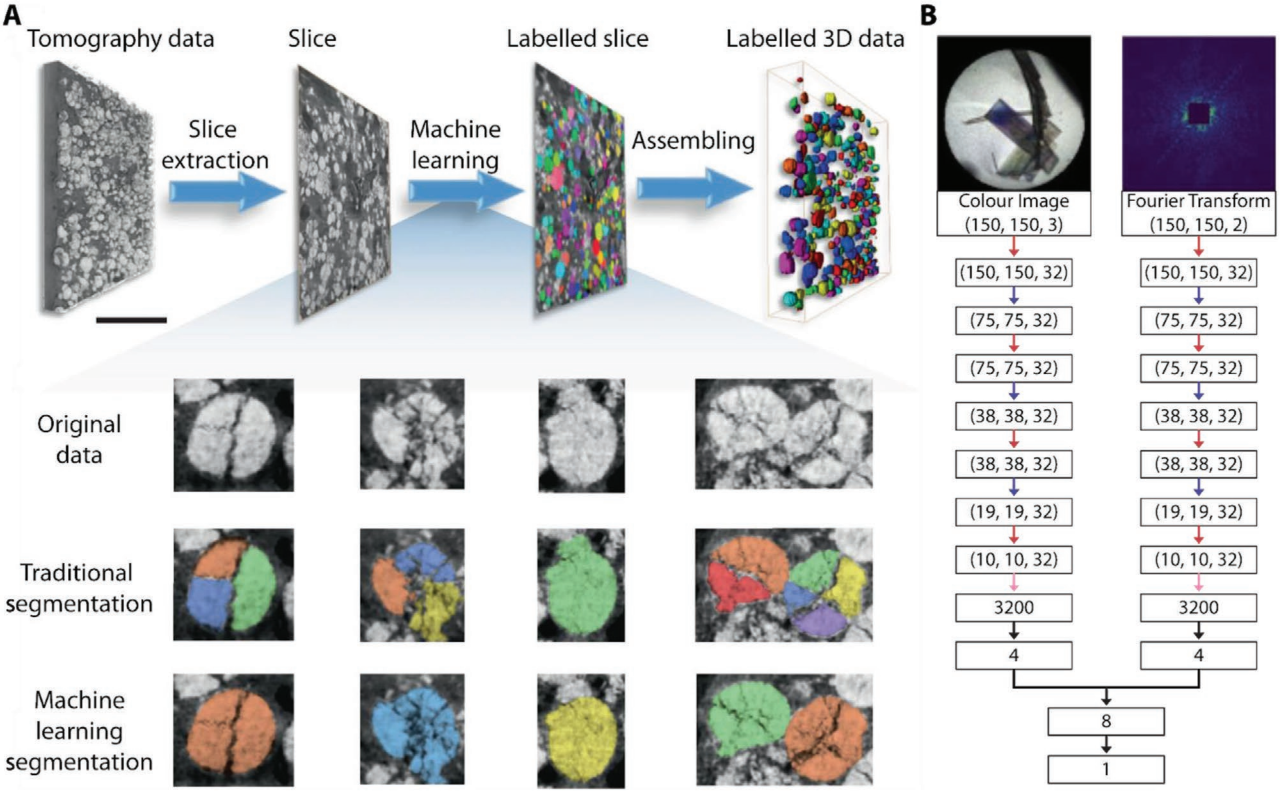


Figure 2. a) ML-enabled segmentation of individual active material particles. Reproduced with permission.[26] Copyright 2020, Springer Nature. b) Convolutional neural network architecture for classifying images of perovskite crystallization. Reproduced with permission.[27] Copyright 2020, Elsevier.

and 3D with an accuracy of 90% and significantly faster than a human.[32]

Analysis of spectra is also of great value in battery research, for example, to address the degradation of electrolytes at the electrode interface, a key factor in the failure of Li-ion batteries.[33] Nuclear magnetic resonance spectroscopy, high-performance liquid chromatography, gas chromatography, and inductively coupled plasma optical emission spectroscopy are among the most common characterization techniques used for quantitative analysis of battery electrolytes.[34] However, these techniques are temperature-dependent, expensive, and require the preparation of many solutions for calibration purposes. Dahn, et al. applied ML algorithms coupled with Fourier-transform infrared spectroscopy for qualitative analysis and solvation shell determination by predicting the features of an unknown electrolyte based on the features of known electrolytes.[35]

Another powerful spectral technique within battery research is electrochemical impedance spectroscopy (EIS). EIS is regularly employed to explore electrochemical properties, such as interfacial phenomena and reactions.[36] However, interpretation and fitting of the large data generated by EIS, even with a simplified equivalent circuit model, are time-consuming, inconsistent, and prone to human bias. Dahn, et al. have recently proposed an inverse model parameterized with deep neural networks to automate the fitting of such data to physical models and omit the initial human input.[37] The practical performance of this model was verified with around 100 000 EIS spectra with a failure rate of less than 1% for Li-ion cells.

4.3. Device Performance and Lifetime Prediction

Experimentally evaluating the lifetime of a device is a time-consuming task that can take months to years, acting as a severe bottleneck in the development of new materials. Data-driven models and ML algorithms can be used to predict the cycle life of batteries and, therefore, offer a swift route to reducing the time burden of such experiments. A similar approach is expected to be applicable for photovoltaic and light emitting diode (LED) devices.

Severson, et al., built data-driven ML models to quantitatively predict the cycle life of commercial Li-ion cells based on data generated from 124 lithium-ion batteries cycled to failure under fast-charging conditions. The cells were tested at high rates (3.6 C constant current–constant voltage) with a cycle life ranging from 150 to 2300 cycles. The uncertainty of the model was found to be 9.1% and 4.9% for the regressor and classifier models, respectively.[38]

Attia, et al. also implemented a two-step ML methodology to predict the cycle life of batteries. In the first step, an ML model estimated the cycle life of the cell based on the first few cycles, significantly reducing the time per experiment. The second stage used Bayesian optimization to explore the parameter space of charge–discharge protocols to reduce the number of experiments required to find the best candidates. Bayesian optimization stabilized the exploitation–exploration trade-off to find the next round of experiments. This approach was used to validate over 224 candidates in 16 days.[39]

5. Predicting New Materials

Perhaps the greatest value of ML for applied materials discovery lies in screening for new materials with properties desired for a particular application. ML can be trained to classify or predict some numerical parameter (e.g., voltage, stability, bandgap, conductivity) based on a limited experimental or computational dataset. It can then be applied to screen the materials in existing databases or to generate new alloyed, doped and even completely new crystal structures or molecules, bypassing expensive and time-consuming DFT calculations or experiments.

Preparing an application-specific training dataset requires a significant effort. However, the emergence of materials databases, such as the Materials Project,[40] OQMD,[41] and AFLOW[42] has eased this burden somewhat, and now serves as a popular starting point for model training.

Experimental datasets are rarer and often proprietary[43] or not machine readable.[44] To alleviate this bottleneck, researchers resort to ML-based natural language processing for scraping data from existing literature.[45,46] However, the lack of reported failed experiments, as discussed above, can be seen as a significant impediment to such efforts.

5.1. Representations

Current precomputed databases and repositories contain only a limited number of data points for any specific sub-class of materials (e.g., perovskites, or water-splitting catalysts, or semiconductors). Smaller datasets impede the accuracy of models, requiring either collecting expensive domain-specific training data or developing improvements to the models to make them more universal and generalizable in order to use the whole dataset for training, even if it is not directly relevant to the application of interest.

The starting point of ML-based screening efforts is generalizable representations of molecules or crystals (atomistic fingerprints) that can capture the underlying physics and key characteristics of the material without too many free parameters, so that reliable predictions of properties can be made using small training datasets (1000–10 000 samples).

There are many different techniques used for the representation and prediction of novel materials, such as sequence-, image-, and graph-based models. Image-based models, utilize pictures in order to represent data. In material science, image-based models can be used, for example, to encode the materials' unit cell (where every atom position is one pixel)[49,50] or to predict spectra.[51–54]

5.1.1. *Graph based Models*

One of the problems arising when working with an inhomogeneous dataset is that different materials contain different number of atoms per unit cell, which poses a challenge when trying to encode them as a fixed-length vector. Current solutions involve manual construction of a library of fragments (bonds, paths, polyhedrons, etc.)[55–57] or using pair distribution functions to encode the atomic neighbor shells.[58–60] Such methods, however, do not use to full extent the advantages of self-training/choosing the best features that ML could potentially bring.

The most recent approach to address these issues involves graph convolutional neural networks (GCNN)[61] (**Figure 3**). A graph-based model, in essence, is the representation of data as a collection of interconnected nodes. The 3D structure of a crystal is reduced into a set of vectors that encode each atom (graph nodes). Several rounds of convolutions are used to update the atomic node vectors based on the properties of the atom, its neighbor, and a bond to it (edges of the graph). These updated atomic vectors are then compressed into a fixed-length vector, independent of crystal structure, by using element-by-element (vector) summation (so called pooling). This allows GCNNs to be both flexible and relatively accurate while maintaining ease of use, only requiring a set of CIF files as inputs and any singular test value for training. Currently, graph based models are the most successful at predicting properties of inorganic crystals.

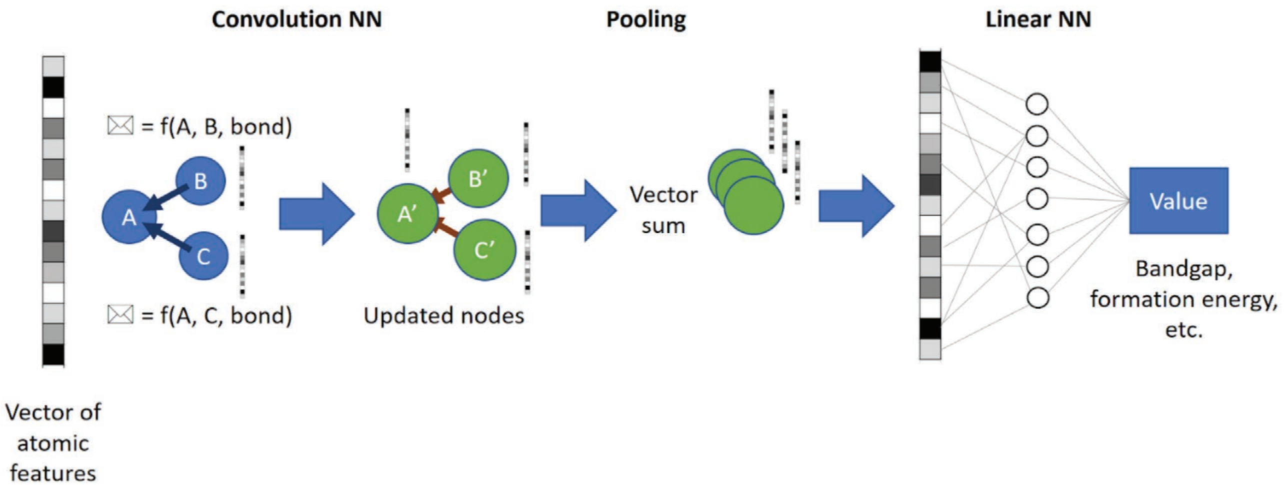


Figure 3. Schematic of a graph convolutional neural network for predicting the properties of crystalline materials. Each graph node (atom) is updated by collecting the "message" from its neighbors using the same rules for all atoms (convolution), then the atomic information is compressed (pooled) and converted into a macroscopic property using a linear neural network.

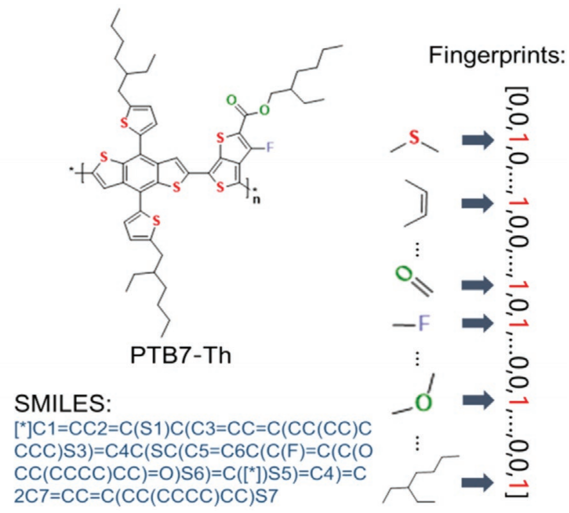


Figure 4. Schematic depicting various expressions of the molecule PTB7-Th and demonstrates the image, SMILES, and fingerprint method. Reproduced with permission.[64] Copyright 2020, American Association for the Advancement of Science.

This can be attributed to the model's enhanced ability in representing the connectivity of atoms within the crystal.

Xie and Grossman applied GCNN for the prediction of various properties of inorganic crystals. They report an MAEs of 0.039 eV per atom for formation energies, 0.388 eV for bandgaps, and 0.054 log(GPa) for bulk moduli from select materials in Materials Project.[61]

An improved version of CGCNN has been recently released,[62] in which bonds are allowed to be updated along with the atoms during the convolution stage, nearest neighbors are decided based on Voronoi tessellation, and 3-body interactions between nearest neighbors are included. Similar improvements were implemented in the MEGNET model,[63] which can be applied to both crystals and molecules, while also adding global state variables, such as pressure and temperature.

5.1.2. Representations for Organic Molecules

The material search space of organic molecules is considerably larger in comparison to inorganic crystalline materials. Furthermore, due to the lack of long-range order in organic materials, a different representation of the material is needed in order to capture the material features with ML.

Simplified molecular-input line-entry system (SMILES) strings are a popular way to encode molecules (**Figure 4**). The Aspuru-Guzik group pioneered the exploration of chemical spaces using SMILES strings with recurrent neural network models popular in natural language processing.[47] Converting the SMILES into a fingerprint vector allows for the most essential elements of the molecule to be captured, detailing how and if certain key features are present in the molecule.

More recently, autoencoders have been used for such conversion of SMILES into multidimensional continuous vector representations amenable to manipulation and optimization by ML algorithms. They are converted back to discrete molecular representations afterward. However, as SMILES were not designed with the grammatical flexibility required by such manipulations in mind, a large portion of the converted SMILES strings are invalid and do not correspond to any molecules. The recently proposed SELFIES[48] representation solves this problem by ensuring every SELFIE string can correspond to a molecule and vice versa. By using SELFIES instead of SMILES, the output of conversions from continuous vectors are guaranteed to be valid and orders of magnitude more diverse molecules can be successfully encoded. This improvement enables much greater utility for generative molecular ML models[65] using, for example, variational autoencoders and generative adversarial networks.

Another form of molecular representation is based on atomic environments, where local chemical environments are modeled by radial distribution functions for both two-body and three-body interactions. Examples include the Faber, Christensen, Huang, Lilienfeld representation introduced by Lilienfeld,[66,67] smooth overlap of atomic densities,[68] and atom-centered symmetry functions.[69,70] ML models based on these representations can enable predictions of atomic energies and forces with chemical accuracy with the speeds on the order of milliseconds.

5.1.3. Other Fingerprinting Methods

Many alternative fingerprinting methods, both generic and task-specific, are becoming increasingly accessible (Figure 4). Python packages such as RDKit[71] and Matminer[72] can vectorize almost any molecule or crystal uniquely based on chemical composition, electronic structure and molecular/crystal structure. The performance and interpretability of ML tasks based on these generic fingerprints may not be satisfactory because of excessive information (causing overfitting) and/or lack of properties that underlie the physics/chemistry of the target property (leading to underfitting). The development and selection of appropriate task-specific fingerprints is therefore of high interest.

As will be discussed below, ML practitioners in each material sub-field have often pioneered fingerprinting methods best suited for their respective material systems and property of interest, for example XRD patterns for exploration of ionic conductors.

5.2. Batteries

Advancements in battery technology will be key to the world's transition to electric vehicles and renewable but intermittent energy sources such as solar and wind. Toward this goal, the development of electrodes with high energy density and solid electrolytes with high ionic conductivity is imperative but remains a challenging task. The difficulty in developing these materials can be best illustrated by the fact that the energy density of current commercial Li-ion cells is only ten times greater than the cell chemistries developed two centuries ago.[73]

5.2.1. Solid-State Electrolytes

Zhang, et al. applied unsupervised learning to guide the discovery of solid-state Li-ion conductors[74] Unsupervised learning is a branch of ML that seeks to identify patterns in unlabeled data,

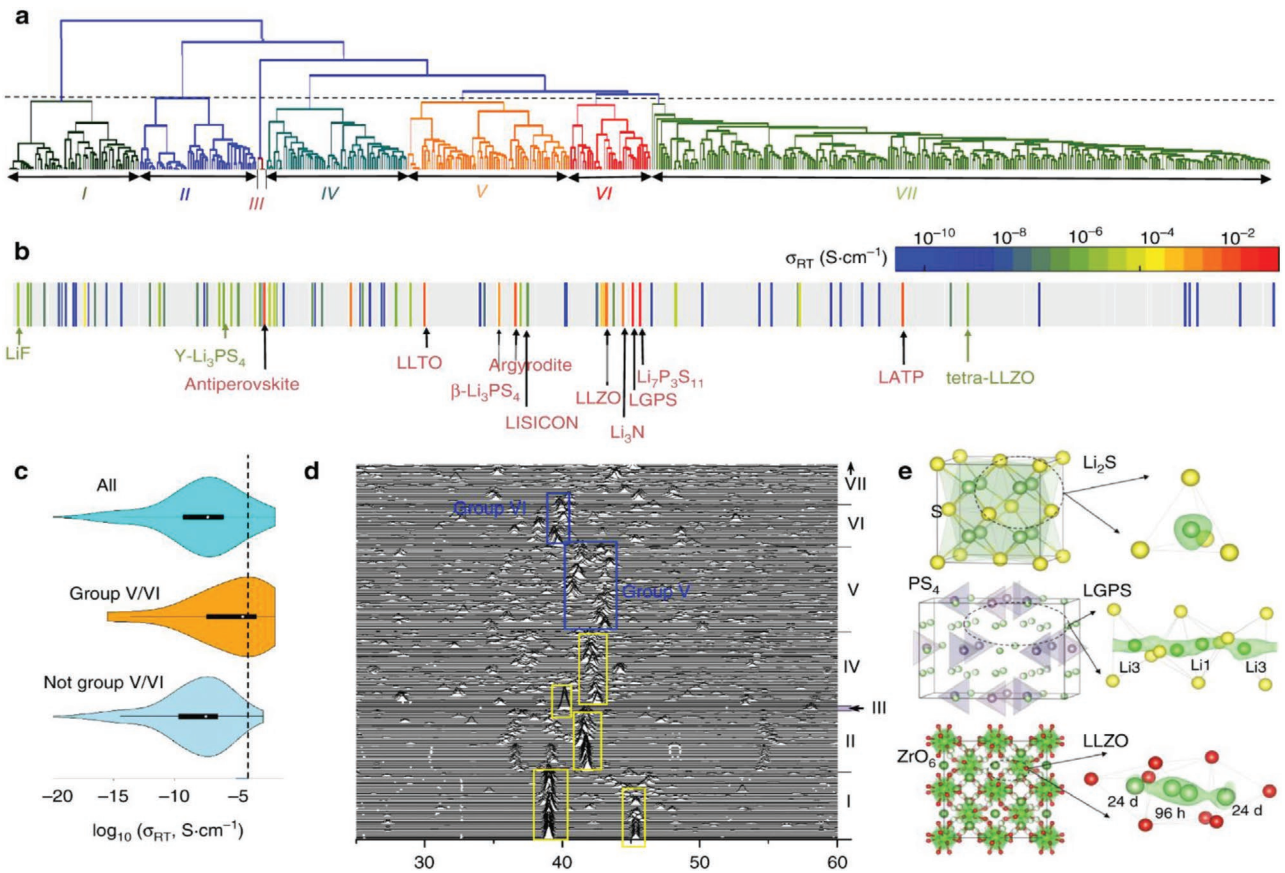


Figure 5. a) Bottom-up tree diagram (dendrogram) generated using the agglomerative hierarchical clustering method. b) Mapping the dendrogram to the conductivity reveals the grouping of known solid-state Li-ion conductors in group V and VI. c) Violin plots of σ_{RT} data grouped in the grouping. The outer shells of the violins bound all data, narrow horizontal lines bound 95% of the data, thick horizontal lines bound 50% of the data, and white dots represent medians. d) mXRD of all materials in group I–VI and a part of group VII. e) Crystal structures (left) and (right) Li sites (green sphere) determined by local anion (yellow/red sphere) configuration, corresponding to isosurfaces (green) of Li probability density from AIMD simulations. Reproduced with permission.[74] Copyright 2019, Springer Nature.

that is, without human guidance; for example, grouping of similar animals. In this particular case, similarities in crystal structures (which are known) are supposed to be correlated with similarities in ionic conductivities (knowledge about which is scarce) and thus can be used as a proxy for predicting the conductivity.

The authors demonstrated that the trained model was able to cluster XRD patterns of anion sub-lattices into distinct groups (**Figure 5**). Most impressively, previously known good ion conductors were all found to fall into the same two groups that had significantly better conductivity than the other five remaining groups. By analyzing the relationship between the clustering and conductivity, the importance of anion lattice distortion and disordered Li sublattices in enabling fast Li-ion conduction was established.

From an initial dataset of ≈3000 materials, the clustering technique was able to narrow down the candidate list to just 82 materials. More computationally expensive ab initio molecular dynamics simulations were then conducted to verify the conductivity of these materials. 16 materials with room temperature conductivity over 10^{-4} S cm^{-1}, of which three had conductivities over 10^{-2} S cm^{-1}, which is better than the best known Li ion conductors, were found. The results are yet to be experimentally validated.

5.2.2. Cathodes

ML models have also been applied for the exploration of high-voltage intercalation cathodes. Using training data from the Materials Project, Joshi, et al. demonstrated that ML models are capable of predicting the DFT-calculated voltages of intercalation cathode materials for various metal ion types (Li, Na, K, Mg, Ca, Al, Zn) to a reasonable accuracy (mean absolute error: ≈0.4 V).[75] Such models can enable accelerated qualitative screening of cathode materials prior to selecting candidates for more accurate DFT calculations and experimental synthesis. The authors demonstrated this ability by proposing nearly 5000 new promising cathode candidates in terms of ML-predicted voltage for Na-ion and K-ion batteries. The compounding of errors from ML predictions to DFT calculations and finally to experimental voltages leaves the utility of such models yet to be tested experimentally.

5.3. Electrocatalysis

Oxygen evolution reaction (OER) is the oxidative counterpart of many important cathodic reactions such as hydrogen evolution reaction, nitrogen reduction reaction, and carbon dioxide reduction reaction.[76–78] The half-reaction is notorious for its sluggish kinetics, requiring an additional energy input (overpotential) to drive the reaction.[76–81] Iridium and ruthenium oxide catalysts are the gold standard of OER in acid: however, their cost is prohibitive, and non-noble metal catalysts are widely sought for.[77–79] Unfortunately, the pathway to such economical catalysts remains challenging, as the acidic environment coupled with the oxidative bias adds an additional constraint, requiring the material to be stable under such harsh conditions.[78,82] ML offers a potential route for rapid screening of materials to resolve this challenge.

To find economical catalysts for acidic OER, Ulissi, et al. screened 2600 different bimetallic combinations of 26 transition metals in 8 crystal space groups using high-throughput DFT.[83] The authors narrowed the search down to 3 promising candidates (Co-Ir, Fe-Ir, and Mo-Ir) that contained 50% less iridium, all while surpassing performance and maintaining stability as determined via theoretical calculations. Similar work has also been conducted by the Norskov group, screening 47 814 metal oxides from the Materials Project database to find 68 promising acid-stable candidates, 15 of which were proposed to be OER-active and synthesizable (**Figure 6**).[84] Another DFT search among 2D materials found 3 stable candidates for acidic OER among a pool of 11 000 prospects.[85]

Although such high-throughput DFT explorations allow for the screening of new materials, the approach is bottlenecked by its computational expense. For example, in their study of IrO_2 and IrO_3 polymorphs, Ulissi, et al., reported an expenditure of 3 million CPU hours to conduct 2000 surface coverage and activity calculations.[86]

Another key limitation of DFT models is the simplifications used in calculations, for example, idealized non-defective surfaces in vacuum instead of electrolyte.[83] Such assumptions are often inadequate to match experimental conditions, where surface restructuring (and even amorphization) is common[83,87] and solvent interactions are crucial.[88] Even though more realistic models are now achievable, the sheer volume of possible configurations prevents a comprehensive exploration of all materials. The computational expense can be potentially reduced by calculating with DFT a small fraction of configurations for the particular system of choice, and using the resultant data to train an ML model that can predict the energetics associated with the remaining configurations of interest.[88] This approach was explored by Ulissi, et al. in their work with IrO_2 and IrO_3. Even with a relatively small training dataset—300 and 500 for surface coverage and OER overpotential calculations, respectively—the authors were able to predict the surface coverage energies and the overpotentials with a high degree of accuracy: 0.10 eV RMSE for coverage calculations and 0.18 eV RMSE for OER overpotentials.[86]

Another ML-based screening of IrO_2 and IrO_3 polymorphs was recently conducted in a collaboration between the Bajdich and Bligaard groups.[89] The authors employed an active learning approach (where newly generated data was added to the model and the model was retrained after each iteration) to screen a set of hypothetical structures (candidate space) to determine the most stable candidates. The iterative approach used a Gaussian process regression (GPR) (a sub-class of Bayesian optimization methods) to sample a small subset of structures from the candidate space based on the formation energies and the associated uncertainty. The structures were then optimized via DFT and subsequently encoded into a vector composed of Voronoi tessellations. The encoded information was then applied to augment the GPR model and predict the formation energies and the associated uncertainties for a test set of unseen structures. The process was repeated until the desired criterion was met. The authors found that the most stable polymorphs could be found with as few as 30 DFT optimizations without the need for prior DFT data, reducing the computational burden of DFT.[89]

A GPR-based ML model was also built in recent work by Xin, et al. to find novel perovskite catalysts for OER. Training on

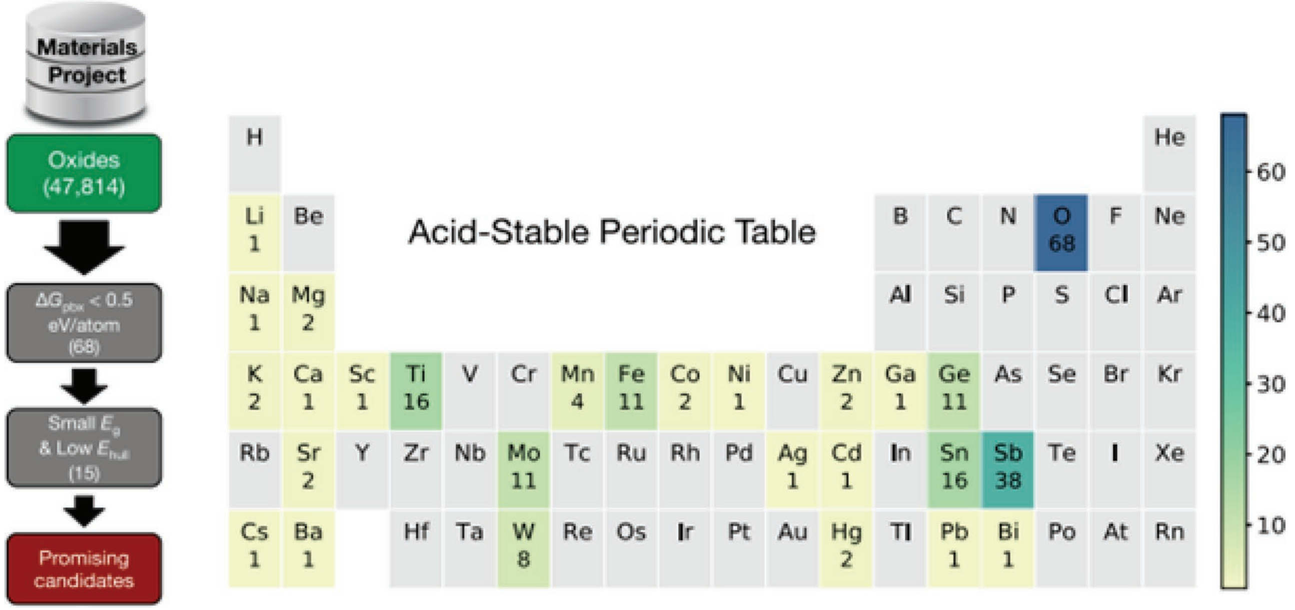


Figure 6. The high-throughput DFT material screening workflow employed by the Norskov group and the elemental distribution of the proposed acid-stable metal oxide candidates. Reproduced with permission.[84] Copyright 2020, American Chemical Society.

250 oxide perovskites, the authors employed an array of structural and electronic descriptors (such as A-site electronegativity, structure tolerance factor, and d-orbital electrons for B-site) and DFT-generated *O–OH and *OH adsorption energies to screen 4000 double perovskites.[90] Among the large set of structures, 10 candidates ($KRbCo_2O_6$, $BaSrCo_2O_6$, $KBaCo_2O_6$, $KCaCo_2O_6$, $BaPbTi_2O_6$, $BaRbTi_2O_6$, $BaSnTi_2O_6$, $BaTlTi_2O_6$, $LaTlTi_2O_6$, $RbEuTi_2O_6$) with an OER performance better than the perovskite gold standard, $LaCoO_3$, were found.[90]

The studies above demonstrate the potential of ML to eventually replace DFT and help accelerate catalytic material discovery. However, the development of such ML tools requires the availability of DFT data, which is scarce in the catalysis space, necessitating the use of high throughput DFT, at least in the short term.

In addition, given the DFT origin of training data, one can only expect to replicate the DFT level of accuracy with ML. Experimental validation of the predictions becomes crucial. In fact, one advantage of ML-based approaches is the possibility of using the experimental performance matrices to augment the accuracy of the ML model toward better capturing the experimental reality. It is an encouraging sign to see such validation practices starting to emerge in literature.[91,92] Using this hybrid approach, the development of new, previously elusive, non-precious metal-based catalysts may finally be in reach.

5.4. Photovoltaics and Light Emitters

For optoelectronic materials, the key metrics for material performance are the bandgap (or wavelength) and whether the material can be prepared with a sufficiently low concentration of defects that otherwise would annihilate the generated electron–hole pairs.

Existing DFT-based materials databases[40–42] offer a valuable starting point for predicting the above properties. With up to 20 000 semiconducting materials available in the databases, it remains unclear whether there remain any completely novel, previously unknown materials, or if all the advances will be made based on alloying and/or doping of existing materials.[93–95] Ongoing ML research in this field thus focuses on improving the accuracy of the bandgap predictions (given the relatively small training dataset), as well as on finding ways to predict properties of alloyed, distorted, and defected structures based solely on data available for perfect crystals.

Even before the accurate quantitative descriptors that take into account the exact atomic structure of the crystal were developed, qualitative models had shown promise in filtering and narrowing down the large search space. Jin, et al., created an ML-based classifier of a material's ability to succeed in photovoltaic (PV) applications. The model was trained on 196 data points of known and previously reported PV and non-PV materials with bandgaps ranging between 1 and 2 eV. The model implicitly included not only the bandgap, but also other important factors such as experimental defect densities, mobilities, etc. Application of the trained classifier to 187 000 experimentally known materials from ICSD identified 3011 promising candidates. After a structural screening along with DFT simulations, the list was further reduced to 26 candidates, among which Sb_2Te_3 and Bi_2Se_3 were identified as promising photovoltaic materials.[96]

Another approach to bypass the need to encode the atomistic crystal structure is to limit the search to a single crystal group where cations and anions all occupy the same lattice sites, and thus interactions between them are implicitly encoded and taken into account. Lu, et al. applied this approach to predict the bandgaps of hybrid organic–inorganic perovskites (HOIP). The ML model was trained on DFT data from 212 HOIPs, using a tolerance factor, octahedral factors, ionic charge, electronegativity, and orbital radii. The model was then applied to screen the space of 5158 potential HOIPs created by combining 32 different A-site cations, 43 B-site cations, and 4 halides. After an initial ML screening, 6 candidate materials were selected for validation based on the ML results and their anticipated ease of synthesis. Upon further DFT analysis, two materials, C2H5OSnBr3 and C2H6NSnBr3, were confirmed as having direct band gaps.[97]

While the above examples do allow for filtering a large list of candidates and obtaining a manageable shortlist for further verification, the lack of precision is likely to produce a lot of false negatives and thus eliminate a lot of potentially promising candidates too. Furthermore, the initial exploration space could have been reduced by at least 85% by simply applying the desired filters prior to the ML screening and exploring the remaining candidates with DFT.[98]

An example of applying ML to predict new materials for white-light phosphors was demonstrated by Zhuo, et al. The property of interest, quantum efficiency, was replaced by an easier to compute proxy, structural rigidity, which in turn is reliably estimated with the material's Debye temperature.[99,100] DFT calculation of the Debye temperature is reliable but prohibitively expensive, making it a great candidate to be predicted with ML. Using data from the Materials Project, 2071 materials of interest with more than three phases and common starting reagents were screened.[40,100] By maximizing the Debye temperature for given bandgaps, a phosphor host of $NaBaB_9O_{15}$ was identified to be optimal for synthesis. Subsequent synthesis of the phosphor host resulted in a streamlined preparation of Eu^{2+} doped, green-emitting material showing high efficiency (95% quantum yield).[99]

Analysis of trained ML models provides an avenue to better understanding the underlying physics and factors affecting the property of interest and, consequently, tailor those factors to tune the material. Im, et al. created an ML protocol that creates an importance score for each feature based on its effect on the model accuracy.[101] For instance, the most crucial features to predict the bandgap of a double perovskite are the space group, the electronegativity of the halogen, highest occupied atomic level, and lowest unoccupied atomic level of B^{1+} atom. For the formation energy, the main determining factor was the electronegativity of halogens and the B^{3+} to halogen bond length. Pilania, et al. had a similar approach, where their model had the ability to create correlations between the feature values and material bandgap. Using this approach, they were able to find the most useful features, and map out trends.[102]

5.5. Other Applications

In addition to the applications highlighted above, the diversity of material science and the rapid uptake of machine learning has meant that many other areas of research have also benefited, especially those of modular/combinatorial nature such as metal organic frameworks (MOFs),[103–107] covalent organic frameworks (COFs),[103,108] and 2D materials,[109–113] to name a few. For example, Aspuru-Guzik, et al. recently harnessed a variational autoencoder to design new MOFs for the CO_2 separation.[107] The top candidate among the generated MOFs exhibited a CO_2 capacity of 7.55 mol kg^{-1} with a selectivity value of 16 (CO_2 to CH_4).

Other application-based use-cases of ML within material science include superconductors,[57,114,115] topological insulators,[112,116] ferroelectric materials,[113,117–119] piezoelectric materials,[120–122] supercapacitors,[123–125] and 3D bioprinting.[126,127] This list is expected to grow in the near future, as access to trained personnel and high-throughput robotics increases.

6. Outlook

Given the progress in ML for material discovery in recent years, research in the field is now pivoting toward developing ML models with DFT-level accuracy, that are generalizable and transferable between different materials (including between organics and inorganics). In this section, we outline some of the emerging efforts in these directions along with a perspective on challenges around dataset availability and size.

6.1. Formation Energies, Defects, and Surfaces

The total energy is one of the primary outputs of any DFT calculation and can be used to estimate the synthesizability/stability of materials by comparing their total energy to the sum of total energies of the constituents. Databases with DFT formation energies of more than 100 000 crystalline materials (metals and non-metals) provide sufficient data for training the ML models to predict materials stability.[55,61,63] However, the current accuracy (≈40 meV per atom) of said models remains insufficient to differentiate the stability of competing same-stoichiometry phases. This inability to discriminate between phases can be detrimental for some applications such as batteries and optoelectronics, where the stabilization of the wrong phase (10–30 meV per atom) may mean that a completely irrelevant inactive phase is obtained instead of the desired active phase, derailing device performance.

If a DFT-level accuracy is desired from an ML model, the model needs to be provided with similar information to its DFT counterpart, for example, atomic orbital energies, shapes, and degeneracies; such rich descriptors are yet to be included. As such, further improvements of ML models toward DFT level accuracies are expected from changes to neural network architectures,[62,128] improved pooling procedures,[129,130] and more importantly the inclusion of physics-based information-rich descriptors.

As ML becomes proficient in material discovery within the domain of perfect crystals, more complicated systems focusing on defects, alloys, distorted structures, and surfaces will become the next research frontier. These imperfections play a crucial role in the material's performance but are less tangible and harder to measure experimentally and calculate theoretically. While some attempts have been made to weed out high-defect materials implicitly through macroscopic material performance, the most common way of calculating defect formation is by DFT, which is extremely laborious and computationally expensive. Public datasets containing defects are missing but are under active development.[131]

Defect formation energies, in principle, could be predicted from ML models that are trained on the bonding energies derived from the crystal formation energies in the available databases. The situation is most promising for dopants and alloys, where the bonding and atomic configuration remains the same as in the original crystals. In the case of vacancies and surfaces, where clipped bonds are formed, as well as for various adsorbents on the surface, the lack of similar examples in the training dataset will pose an interesting challenge.

Given the aforementioned limitations, further progress requires either generating more data (which is unfeasible given the practically infinite number of possible configurations) or developing more agnostic and generalizable ML methods that approach the problem on a more fundamental level, for example, by treating each atom as a set of orbitals of a given energy, just like DFT does, instead of associating an element label to each atom.

6.2. Machine-Learned Interatomic Potentials

A major limitation of current DFT-based methods is the high computational cost of modeling complex and long-time behavior especially in systems with a large number of atoms. For example, in the design of solid and liquid electrolytes, current ab initio molecular dynamics have been proven to be sufficient in distinguishing promising ionic conductors. However, it is still intractable to model accurately the long-timescale behavior of electrolyte–electrode interfaces, which play a critical role in the device failure of current materials.[132,133]

Classical molecular mechanics simulations are much faster but typically cannot simulate bond breaking, that is, chemical reactions. In rare instances, for example, in ReaxFF,[134] bond breaking can be included but has to be meticulously parametrized and is not yet available for the majority of elements. In addition, existing force fields are not sufficiently accurate in describing polarization effects.[135]

ML-based potentials/force-fields can provide a viable solution to enable accurate molecular dynamics at a fraction of the cost of DFT. Instead of devising a universal force field, "learning on the fly" (LOTF) from a system of interest could be used to learn the atomic interactions and forces relevant to the system of interest. To achieve this, a DFT molecular dynamics simulation must first be performed for a sufficient number of steps such that enough knowledge is collected to train the ML model. Having been trained on the local atomic environment of interest, the ML model can bypass the DFT and predict the atomic forces. Some available software packages that explore this feature include MLIP, SchNetPack, and VASP (to be implemented in 2021).[89,136,137] This LOTF strategy has been recently employed to calculate Li-ion diffusivities in solid-state electrolytes, opening the possibility to identify protective coating candidates for cathode materials.[138]

6.3. Dealing with Small Datasets

Experimental performance of materials often depends on complex parameters such as defect densities (PV), photoluminescence quantum yield (LEDs), ionic conductivity (batteries), and electronic transfer kinetics (catalysis). These parameters are hard to compute or experimentally collect in large volumes, resulting in small datasets that are inadequate to train the conventional atomistic ML models, yet the number of free parameters in such problems is too large for Bayesian optimization.

Transfer learning is a powerful tool that can address dataset limitations by taking an existing pretrained model and fine-tuning it to predict a new property. This approach is widely used in image recognition, where state-of-the-art image recognition models trained on millions of images are applied to recognize and classify new types of images, for example, X-ray scans. The pretrained model has effectively learned what an image is and how to extract features like lines or shades from any image. It thus only needs an add-on (a few last layers of a neural network) to analyze the extracted image features in order to differentiate a bad versus good X-ray scan.

Similarly, models pretrained to predict chemical properties (formation energy, bandgap, etc.) on large datasets of ≈100 000 materials from public databases have essentially learned the basics of chemistry and can be easily augmented to predict a more complex parameter of interest with a smaller dataset.

A typical neural network architecture—using a funnel of nodes—allows for the compression of the long material representation vector into a smaller dimensional space (called the latent space) amenable to Bayesian optimization[47,65] (**Figure 7**). It is, however, important to ensure that the property of interest correlates well with the complementary property on which the original model was trained, otherwise one risks finding that the property of interest is not continuously distributed in the latent space.[47] One approach to deal with such situations is to employ proxies—computationally cheaper properties that are strongly correlated with the desired but costly target parameters. For example, the empty volume available for Li diffusion within a crystal structure can be correlated to ionic conductivity in solid-state electrolytes.[139] In catalysis, the oxygen d-band position has long been used as a proxy to OH binding energy. These proxies can either be used as an output property on which the initial model is trained or can be added as an additional input feature for the model, helping it to find the exact functional form of correlation to the property of interest.

A similar approach to leveraging small datasets is through multi-fidelity prediction. By augmenting the small, high-fidelity dataset with a more expansive but low-fidelity set, one can improve the accuracy of extrapolation beyond the boundaries of the smaller dataset. This tiered system can help to fine-tune the predictions from the coarser set and achieve the accuracy of the higher-fidelity one. Chen, et al. employed such a multi-fidelity approach to improve bandgap predictions using 52 348 samples of the low-accuracy Perdew–Burke–Ernzerhof bandgaps, 6030 high-accuracy Heyd-Scuseria-Ernzerhof bandgaps, and 2981 experimental bandgaps, allowing to lower the prediction errors by up to 45%.[140]

6.4. Data Generation and Availability

Scientific research is a collaborative enterprise, the progression of which is heavily reliant on the availability of past knowledge. Open communication of experimental details is, therefore, not only important for the sake of replication/verification but also for further progress. In the realm of ML, a unified effort to publish all data could help alleviate some of the issues stemming from data shortages. Unfortunately, far too often, data used to build a given ML model is not published. Even when established databases such as the Materials Project are utilized, researchers often fail to specify the exact data that was employed for the development and testing of the ML model. This serves as a significant hurdle in the reproduction and development of improved ML models.

Data sharing is vitally important and should be enforced for all ML reports. This includes sharing the datasets used for the training and validation, along with the ML code and the weights of the model.

Utilizing experimental data for ML model training is a promising approach to capture parameters not easily described by DFT simulations. To alleviate data scarcity, the scientific community should look to transitioning to digital laboratory record keeping (including failed experiments) amenable to rapid data analysis and ML.

Published pretrained models offer experimentalists the capability to quickly assess, prioritize new ideas, and screen new materials. For example, Shields, et al. have demonstrated that Bayesian optimization can be used to find synthetic pathways for targeting organic molecules that are more efficient than counterparts proposed by a human.[141] When properly integrated, such models can be a useful tool to reduce the optimization time required for an experimentalist, increasing the productivity for experimentalists.

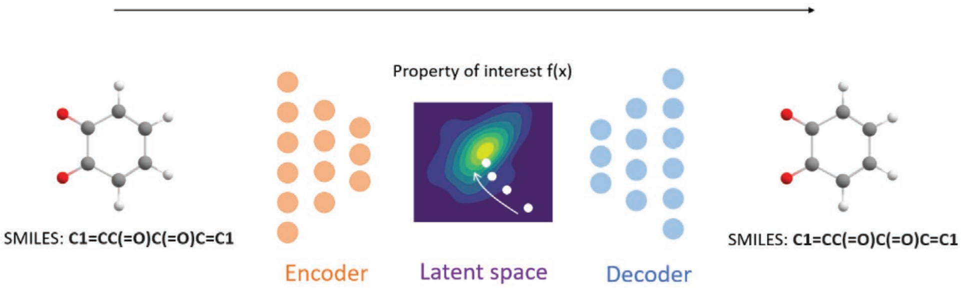

Property of interest f(x)
SMILES: C1=CC(=O)C(=O)C=C1
SMILES: C1=CC(=O)C(=O)C=C1
Encoder
Latent space
Decoder

Figure 7. Autoencoder architecture for chemical space exploration and optimization.

7. Conclusion

The resurgence of machine learning has occurred at a pivotal point in time and is expected to be an important tool in combatting global issues, such as climate change, the resolution of which necessitates rapid material discovery. In addition to the reduction in time cost, the use of such algorithms bypasses human biases that otherwise may hinder reproducibility and innovation. Further improvements are expected from interfacing the ML infrastructure with high-throughput robotics to expedite material discovery via partial or complete elimination of human labor.

ML has shown great promise in screening databases to find new materials, predicting basic material properties, optimizing experimental parameter spaces, and automating data analysis. These tools have now become freely available and easy to use by experimentalists and theorists alike and promise significant time savings in the search for new materials.

While it is encouraging to see the progress toward replacing DFT-based screening of large datasets, the ultimate goal of such screening—finding new functional materials—should not be forgotten. Given that simulations are often idealized replications of nature, it is of great importance to evaluate the proposed candidates beyond the in silico screening, that is, experimentally preparing the materials and assessing their functionality. While such datasets are inherently small, techniques that can leverage smaller data are being actively developed. Such experimental tests, if made publicly available, will serve as next-generation datasets for even more accurate models.

Acknowledgements

F.D., K.S., and T.D. contributed equally to this work.

Conflict of Interest

The authors declare no conflict of interest.

Keywords

batteries, electrocatalysis, machine learning, materials discovery, optoelectronics, solid-state electrolytes

Received: May 3, 2021
Revised: July 20, 2021
Published online: September 13, 2021

[1] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, *Nature* **2018**, *559*, 547.

[2] J.-L. Reymond, *Acc. Chem. Res.* **2015**, *48*, 722.

[3] S.-M. Udrescu, M. Tegmark, *Sci. Adv.* **2020**, *6*, eaay2631.

[4] S. Sun, A. Tiihonen, F. Oviedo, Z. Liu, J. Thapa, Y. Zhao, N. T. P. Hartono, A. Goyal, T. Heumueller, C. Batali, A. Encinas, J. J. Yoo, R. Li, Z. Ren, I. M. Peters, C. J. Brabec, M. G. Bawendi, V. Stevanovic, J. Fisher, T. Buonassisi, *Matter* **2021**, *4*, 1305.

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, *arXiv:1201.0490* **2018**.

[6] R. Martinez-Cantin, K. Tee, M. McCourt, *arXiv:1712.04567* **2017**.

[7] F. Häse, L. M. Roch, C. Kreisbeck, A. Aspuru-Guzik, *ACS Cent. Sci.* **2018**, *4*, 1134.

[8] J. T. Springenberg, A. Klein, S. Falkner, F. Hutter, in *Advances in Neural Information Processing Systems* (Eds: D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett), Curran Associates, Red Hook, NY **2016**.

[9] J. Snoek, O. Rippel, K. Swersky, R. Kiros, N. Satish, N. Sundaram, M. Patwary, M. Prabhat, R. Adams, In Proc. of the 32nd Int. Conf. on Machine Learning (Eds: F. Bach, D. Blei), PMLR, Lille **2015**, pp. 2171–2180.

[10] A. Kirkey, E. J. Luber, B. Cao, B. C. Olsen, J. M. Buriak, *ACS Appl. Mater. Interfaces* **2020**, *12*, 54596.

[11] K. Homma, Y. Liu, M. Sumita, R. Tamura, N. Fushimi, J. Iwata, K. Tsuda, C. Kaneta, *J. Phys. Chem. C* **2020**, *124*, 12865.

[12] M. Harada, H. Takeda, S. Suzuki, K. Nakano, N. Tanibata, M. Nakayama, M. Karasuyama, I. Takeuchi, *J. Mater. Chem. A* **2020**, *8*, 15103.

[13] A. Dave, J. Mitchell, K. Kandasamy, H. Wang, S. Burke, B. Paria, B. Póczos, J. Whitacre, V. Viswanathan, *Cell Rep. Phys. Sci.* **2020**, *1*, 100264.

[14] M. A. B. Janai, K. L. Woon, C. S. Chan, *Org. Electron.* **2018**, *63*, 257.

[15] T. W. David, H. Anizelli, T. J. Jacobsson, C. Gray, W. Teahan, J. Kettle, *Nano Energy* **2020**, *78*, 105342.

[16] X. Du, L. Lüer, T. Heumueller, J. Wagner, C. Berger, T. Osterrieder, J. Wortmann, S. Langner, U. Vongsaysy, M. Bertrand, N. Li, T. Stubhan, J. Hauch, C. J. Brabec, *Joule* **2021**, *5*, 495.

[17] A. E. Siemenn, M. Beveridge, T. Buonassisi, I. Drori, *arXiv:2105.02858* **2021**.

[18] O. Voznyy, L. Levina, J. Z. Fan, M. Askerka, A. Jain, M.-J. Choi, O. Ouellette, P. Todorović, L. K. Sagar, E. H. Sargent, *ACS Nano* **2019**, *13*, 11122.

[19] Z. Li, M. A. Najeeb, L. Alves, A. Z. Sherman, V. Shekar, P. Cruz Parrilla, I. M. Pendleton, W. Wang, P. W. Nega, M. Zeller, J. Schrier, A. J. Norquist, E. M. Chan, *Chem. Mater.* **2020**, *32*, 5650.

[20] P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, A. J. Norquist, *Nature* **2016**, *533*, 73.

[21] D. S. Salley, G. A. Keenan, D.-L. Long, N. L. Bell, L. Cronin, *ACS Cent. Sci.* **2020**, *6*, 1587.

[22] F. Mekki-Berrada, Z. Ren, T. Huang, W. K. Wong, F. Zheng, J. Xie, I. P. S. Tian, S. Jayavelu, Z. Mahfoud, D. Bash, K. Hippalgaonkar, S. Khan, T. Buonassisi, Q. Li, X. Wang, *npj Comput. Mater.* **2021**, *7*, 55.

[23] 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*, eaaz8867.

[24] D. Xue, P. V. Balachandran, R. Yuan, T. Hu, X. Qian, E. R. Dougherty, T. Lookman, *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 13301.

[25] B. Burger, P. M. Maffettone, V. V. Gusev, C. M. Aitchison, Y. Bai, X. Wang, X. Li, B. M. Alston, B. Li, R. Clowes, N. Rankin, B. Harris, R. S. Sprick, A. I. Cooper, *Nature* **2020**, *583*, 237.

[26] Z. Jiang, J. Li, Y. Yang, L. Mu, C. Wei, X. Yu, P. Pianetta, K. Zhao, P. Cloetens, F. Lin, Y. Liu, *Nat. Commun.* **2020**, *11*, 2310.

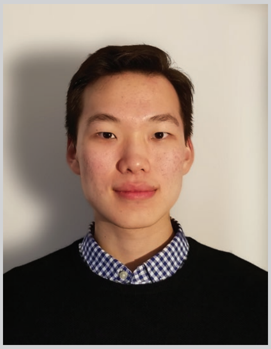
[27] J. Kirman, A. Johnston, D. A. Kuntz, M. Askerka, Y. Gao, P. Todorović, D. Ma, G. G. Privé, E. H. Sargent, *Matter* **2020**, *2*, 938.

[28] S. Eppel, T. Kachman, *arXiv:1404.7174* **2014**.
[29] S. Eppel, H. Xu, M. Bismuth, A. Aspuru-Guzik, *ACS Cent. Sci.* **2020**, *6*, 1743.
[30] C. L. Phillips, G. A. Voth, *Soft Matter* **2013**, *9*, 8552.
[31] V. H. Nguyen, V. H. Pham, X. Cui, M. Ma, H. Kim, *J. Inf. Telecommun.* **2017**, *1*, 334.
[32] S. Sun, N. T. P. Hartono, Z. D. Ren, F. Oviedo, A. M. Buscemi, M. Layurova, D. X. Chen, T. Ogunfunmi, J. Thapa, S. Ramasamy, C. Settens, B. L. DeCost, A. G. Kusne, Z. Liu, S. I. P. Tian, I. M. Peters, J.-P. Correa-Baena, T. Buonassisi, *Joule* **2019**, *3*, 1437.
[33] E. Peled, *J. Electrochem. Soc.* **1979**, *126*, 2047.
[34] G. Gachot, P. Ribière, D. Mathiron, S. Grugeon, M. Armand, J.-B. Leriche, S. Pilard, S. Laruelle, *Anal. Chem.* **2011**, *83*, 478.
[35] L. D. Ellis, S. Buteau, S. G. Hames, L. M. Thompson, D. S. Hall, J. R. Dahn, *J. Electrochem. Soc.* **2018**, *165*, A256.
[36] A. Lasia, in *Electrochemical Impedance Spectroscopy and its Applications* (Ed: A. Lasia), Springer, New York **2014**, pp. 301–321.
[37] S. Buteau, J. R. Dahn, *J. Electrochem. Soc.* **2019**, *166*, A1611.
[38] K. A. Severson, P. M. Attia, N. Jin, N. Perkins, B. Jiang, Z. Yang, M. H. Chen, M. Aykol, P. K. Herring, D. Fraggedakis, M. Z. Bazant, S. J. Harris, W. C. Chueh, R. D. Braatz, *Nat. Energy* **2019**, *4*, 383.
[39] P. M. Attia, A. Grover, N. Jin, K. A. Severson, T. M. Markov, Y.-H. Liao, M. H. Chen, B. Cheong, N. Perkins, Z. Yang, P. K. Herring, M. Aykol, S. J. Harris, R. D. Braatz, S. Ermon, W. C. Chueh, *Nature* **2020**, *578*, 397.
[40] 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 Mater.* **2013**, *1*, 011002.
[41] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, C. Wolverton, *JOM* **2013**, *65*, 1501.
[42] S. Curtarolo, W. Setyawan, G. L. W. Hart, M. Jahnatek, R. V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M. J. Mehl, H. T. Stokes, D. O. Demchenko, D. Morgan, *Comput. Mater. Sci.* **2012**, *58*, 218.
[43] J. E. Saal, A. O. Oliynyk, B. Meredig, *Annu. Rev. Mater. Res.* **2020**, *50*, 49.
[44] Springer Materials [Database] SpringerNature, https://materials.springer.com/ (accessed: August 2021).
[45] E. Kim, Z. Jensen, A. van Grootel, K. Huang, M. Staib, S. Mysore, H.-S. Chang, E. Strubell, A. McCallum, S. Jegelka, E. Olivetti, *J. Chem. Inf. Model.* **2020**, *60*, 1194.
[46] E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder, E. Olivetti, *Chem. Mater.* **2017**, *29*, 9436.
[47] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, A. Aspuru-Guzik, *ACS Cent. Sci.* **2018**, *4*, 268.
[48] M. Krenn, F. Häse, A. Nigam, P. Friederich, A. Aspuru-Guzik, *Mach. Learn.: Sci. Technol.* **2020**, *1*, 045024.
[49] M. Askerka, Z. Li, M. Lempen, Y. Liu, A. Johnston, M. I. Saidaminov, Z. Zajacz, E. H. Sargent, *J. Am. Chem. Soc.* **2019**, *141*, 3682.
[50] H. Choubisa, M. Askerka, K. Ryczko, O. Voznyy, K. Mills, I. Tamblyn, E. H. Sargent, *Matter* **2020**, *3*, 433.
[51] H. Wang, Y. Xie, D. Li, H. Deng, Y. Zhao, M. Xin, J. Lin, *J. Chem. Inf. Model.* **2020**, *60*, 2004.
[52] M. Ziatdinov, O. Dyck, A. Maksov, X. Li, X. Sang, K. Xiao, R. R. Unocic, R. Vasudevan, S. Jesse, S. V. Kalinin, *ACS Nano* **2017**, *11*, 12742.
[53] F. Oviedo, Z. Ren, S. Sun, C. Settens, Z. Liu, N. T. P. Hartono, S. Ramasamy, B. L. DeCost, S. I. P. Tian, G. Romano, A. Gilad Kusne, T. Buonassisi, *npj Comput. Mater.* **2019**, *5*, 60.
[54] T.-W. Ke, A. S. Brewster, S. X. Yu, D. Ushizima, C. Yang, N. K. Sauter, *J. Synchrotron Radiat.* **2018**, *25*, 655.
[55] O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo, A. Tropsha, *Nat. Commun.* **2017**, *8*, 15679.
[56] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, M. Scheffler, *Phys. Rev. Lett.* **2015**, *114*, 105503.
[57] O. Isayev, D. Fourches, E. N. Muratov, C. Oses, K. Rasch, A. Tropsha, S. Curtarolo, *Chem. Mater.* **2015**, *27*, 735.
[58] K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. R. Müller, E. K. U. Gross, *Phys. Rev. B* **2014**, *89*, 205118.
[59] F. Faber, A. Lindmaa, O. A. von Lilienfeld, R. Armiento, *Int. J. Quantum Chem.* **2015**, *115*, 1094.
[60] A. Seko, H. Hayashi, K. Nakayama, A. Takahashi, I. Tanaka, *Phys. Rev. B* **2017**, *95*, 144110.
[61] T. Xie, J. C. Grossman, *Phys. Rev. Lett.* **2018**, *120*, 145301.
[62] C. W. Park, C. Wolverton, *Physical Review Materials* **2020**, *4*, 063801.
[63] C. Chen, W. Ye, Y. Zuo, C. Zheng, S. P. Ong, *Chem. Mater.* **2019**, *31*, 3564.
[64] W. Sun, Y. Zheng, K. Yang, Q. Zhang, A. A. Shah, Z. Wu, Y. Sun, L. Feng, D. Chen, Z. Xiao, S. Lu, Y. Li, K. Sun, *Sci. Adv.* **2019**, *5*, eaay4275.
[65] B. Sanchez-Lengeling, A. Aspuru-Guzik, *Science* **2018**, *361*, 360.
[66] F. A. Faber, A. S. Christensen, B. Huang, O. A. von Lilienfeld, *J. Chem. Phys.* **2018**, *148*, 241717.
[67] A. S. Christensen, L. A. Bratholm, F. A. Faber, O. Anatole von Lilienfeld, *J. Chem. Phys.* **2020**, *152*, 044107.
[68] A. P. Bartók, R. Kondor, G. Csányi, *Phys. Rev. B* **2013**, *87*, 184115.
[69] M. Gastegger, L. Schwiedrzik, M. Bittermann, F. Berzsenyi, P. Marquetand, *J. Chem. Phys.* **2018**, *148*, 241709.
[70] J. Behler, *J. Chem. Phys.* **2011**, *134*, 074106.
[71] G. Landrum, RDKit: Open-Source Cheminformatics Software, **2016**, https://www.rdkit.org/ (accessed: August 2021).
[72] L. Ward, A. Dunn, A. Faghaninia, N. E. R. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K. A. Persson, G. J. Snyder, I. Foster, A. Jain, *Comput. Mater. Sci.* **2018**, *152*, 60.
[73] J.-M. Tarascon, *Philos. Trans. R. Soc., A* **2010**, *368*, 3227.
[74] 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.
[75] R. P. Joshi, J. Eickholt, L. Li, M. Fornari, V. Barone, J. E. Peralta, *ACS Appl. Mater. Interfaces* **2019**, *11*, 18494.
[76] J. S. Kim, B. Kim, H. Kim, K. Kang, *Adv. Energy Mater.* **2018**, *8*, 1702774.
[77] N.-T. Suen, S.-F. Hung, Q. Quan, N. Zhang, Y.-J. Xu, H. M. Chen, *Chem. Soc. Rev.* **2017**, *46*, 337.
[78] Z. Shi, X. Wang, J. Ge, C. Liu, W. Xing, *Nanoscale* **2020**, *12*, 13249.
[79] Q. Shi, C. Zhu, D. Du, Y. Lin, *Chem. Soc. Rev.* **2019**, *48*, 3181.
[80] D. Yang, B. Ni, X. Wang, *Adv. Energy Mater.* **2020**, *10*, 2001142.
[81] M. Luo, S. Guo, *Nat. Rev. Mater.* **2017**, *2*, 17059.
[82] X.-K. Gu, J. C. A. Camayang, S. Samira, E. Nikolla, *J. Catal.* **2020**, *388*, 130.
[83] S. Back, K. Tran, Z. W. Ulissi, *ACS Appl. Mater. Interfaces* **2020**, *12*, 38256.
[84] Z. Wang, Y.-R. Zheng, I. Chorkendorff, J. K. Nørskov, *ACS Energy Lett.* **2020**, *5*, 2905.
[85] A. Jain, Z. Wang, J. K. Nørskov, *ACS Energy Lett.* **2019**, *4*, 1410.
[86] S. Back, K. Tran, Z. W. Ulissi, *ACS Catal.* **2019**, *9*, 7651.
[87] G. Wan, J. W. Freeland, J. Kloppenburg, G. Petretto, J. N. Nelson, D.-Y. Kuo, C.-J. Sun, J. Wen, J. T. Diulus, G. S. Herman, Y. Dong, R. Kou, J. Sun, S. Chen, K. M. Shen, D. G. Schlom, G.-M. Rignanese, G. Hautier, D. D. Fong, Z. Feng, H. Zhou, J. Suntivich, *Sci. Adv.* **2021**, *7*, eabc7323.
[88] Y. Chen, Y. Huang, T. Cheng, W. A. Goddard, *J. Am. Chem. Soc.* **2019**, *141*, 11651.

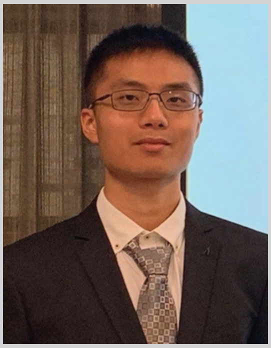
[89] R. A. Flores, C. Paolucci, K. T. Winther, A. Jain, J. A. G. Torres, M. Aykol, J. Montoya, J. K. Nørskov, M. Bajdich, T. Bligaard, *Chem. Mater.* **2020**, *32*, 5854.
[90] Z. Li, L. E. K. Achenie, H. Xin, *ACS Catal.* **2020**, *10*, 4377.
[91] B. Weng, Z. Song, R. Zhu, Q. Yan, Q. Sun, C. G. Grice, Y. Yan, W.-J. Yin, *Nat. Commun.* **2020**, *11*, 3513.
[92] B. Rohr, H. S. Stein, D. Guevarra, Y. Wang, J. A. Haber, M. Aykol, S. K. Suram, J. M. Gregoire, *Chem. Sci.* **2020**, *11*, 2696.
[93] Z. Wang, J. Ha, Y. H. Kim, W. B. Im, J. McKittrick, S. P. Ong, *Joule* **2018**, *2*, 914.
[94] Z. Wang, I.-H. Chu, F. Zhou, S. P. Ong, *Chem. Mater.* **2016**, *28*, 4024.
[95] S. Li, Y. Xia, M. Amachraa, N. T. Hung, Z. Wang, S. P. Ong, R.-J. Xie, *Chem. Mater.* **2019**, *31*, 6286.
[96] H. Jin, H. Zhang, J. Li, T. Wang, L. Wan, H. Guo, Y. Wei, *J. Phys. Chem. Lett.* **2020**, *11*, 3075.
[97] S. Lu, Q. Zhou, Y. Ouyang, Y. Guo, Q. Li, J. Wang, *Nat. Commun.* **2018**, *9*, 3405.
[98] A. Jain, O. Voznyy, E. H. Sargent, *J. Phys. Chem. C* **2017**, *121*, 7183.
[99] Y. Zhuo, J. Zhong, J. Brgoch, *ChemRxiv* **2019**, 10.26434/chem-rxiv.10110773.v1.
[100] Y. Zhuo, A. Mansouri Tehrani, A. O. Oliynyk, A. C. Duke, J. Brgoch, *Nat. Commun.* **2018**, *9*, 4377.
[101] J. Im, S. Lee, T.-W. Ko, H. W. Kim, Y. Hyon, H. Chang, *npj Comput. Mater.* **2019**, *5*, 37.
[102] G. Pilania, A. Mannodi-Kanakkithodi, B. P. Uberuaga, R. Ramprasad, J. E. Gubernatis, T. Lookman, *Sci. Rep.* **2016**, *6*, 19375.
[103] K. M. Jablonka, D. Ongari, S. M. Moosavi, B. Smit, *Chem. Rev.* **2020**, *120*, 8066.
[104] C. Altintas, O. F. Altundal, S. Keskin, R. Yildirim, *J. Chem. Inf. Model.* **2021**, *61*, 2131.
[105] M. Pardakhti, E. Moharreri, D. Wanik, S. L. Suib, R. Srivastava, *ACS Comb. Sci.* **2017**, *19*, 640.
[106] S. Kancharlapalli, A. Gopalan, M. Haranczyk, R. Q. Snurr, *J. Chem. Theory Comput.* **2021**, *17*, 3052.
[107] Z. Yao, B. Sánchez-Lengeling, N. S. Bobbitt, B. J. Bucior, S. G. H. Kumar, S. P. Collins, T. Burns, T. K. Woo, O. K. Farha, R. Q. Snurr, A. Aspuru-Guzik, *Nat. Mach. Intell.* **2021**, *3*, 76.
[108] P. Yang, H. Zhang, X. Lai, K. Wang, Q. Yang, D. Yu, *ACS Omega* **2021**, *6*, 17149.
[109] X. Dong, H. Li, Z. Jiang, T. Grünleitner, İ. Güler, J. Dong, K. Wang, M. H. Köhler, M. Jakobi, B. H. Menze, A. K. Yetisen, I. D. Sharp, A. V. Stier, J. J. Finley, A. W. Koch, *ACS Nano* **2021**, *15*, 3139.
[110] S. A. Tawfik, O. Isayev, C. Stampfl, J. Shapter, D. A. Winkler, M. J. Ford, *Adv. Theory Simul.* **2019**, *2*, 1800128.
[111] N. C. Frey, D. Akinwande, D. Jariwala, V. B. Shenoy, *ACS Nano* **2020**, *14*, 13406.
[112] G. R. Schleder, B. Focassio, A. Fazzio, *Appl. Phys. Rev.* **2021**, *8*, 031409.
[113] X.-Y. Ma, H.-Y. Lyu, K.-R. Hao, Y.-M. Zhao, X. Qian, Q.-B. Yan, G. Su, *Sci. Bull.* **2021**, *66*, 233.
[114] V. Stanev, C. Oses, A. G. Kusne, E. Rodriguez, J. Paglione, S. Curtarolo, I. Takeuchi, *npj Comput. Mater.* **2018**, *4*, 29.
[115] T. Konno, H. Kurokawa, F. Nabeshima, Y. Sakishita, R. Ogawa, I. Hosako, A. Maeda, *Phys. Rev. B* **2021**, *103*, 014509.
[116] N. Claussen, B. A. Bernevig, N. Regnault, *Phys. Rev. B* **2020**, *101*, 245117.
[117] S. M. Neumayer, S. Jesse, G. Velarde, A. L. Kholkin, I. Kravchenko, L. W. Martin, N. Balke, P. Maksymovych, *Nanoscale Adv.* **2020**, *2*, 2063.
[118] N. Liu, A. Ihalage, H. Zhang, H. Giddens, H. Yan, Y. Hao, *J. Mater. Chem. C* **2020**, *8*, 10352.
[119] P. V. Balachandran, B. Kowalski, A. Sehirlioglu, T. Lookman, *Nat. Commun.* **2018**, *9*, 1668.
[120] P. Jiao, *Nano Energy* **2021**, *88*, 106227.
[121] K. Choudhary, K. F. Garrity, V. Sharma, A. J. Biacchi, A. R. Hight Walker, F. Tavazza, *npj Comput. Mater.* **2020**, *6*, 64.
[122] Y. H. Jung, S. K. Hong, H. S. Wang, J. H. Han, T. X. Pham, H. Park, J. Kim, S. Kang, C. D. Yoo, K. J. Lee, *Adv. Mater.* **2020**, *32*, 1904020.
[123] H. Su, S. Lin, S. Deng, C. Lian, Y. Shang, H. Liu, *Nanoscale Adv.* **2019**, *1*, 2162.
[124] J. Ren, X. Lin, J. Liu, T. Han, Z. Wang, H. Zhang, J. Li, *Mater. Today Energy* **2020**, *18*, 100537.
[125] S. Zhu, J. Li, L. Ma, C. He, E. Liu, F. He, C. Shi, N. Zhao, *Mater. Lett.* **2018**, *233*, 294.
[126] Z. Jin, Z. Zhang, X. Shao, G. X. Gu, *ACS Biomater. Sci. Eng.* **2021**.
[127] K. Ruberu, M. Senadeera, S. Rana, S. Gupta, J. Chung, Z. Yue, S. Venkatesh, G. Wallace, *Appl. Mater. Today* **2021**, *22*, 100914.
[128] K. Das, B. Samanta, P. Goyal, S.-C. Lee, S. Bhattacharjee, N. Ganguly, *arXiv:2104.10869* **2021**.
[129] J. Lee, I. Lee, J. Kang, *arXiv:1904.08082* **2019**.
[130] E. Ranjan, S. Sanyal, P. P. Talukdar, *arXiv:1911.07979* **2020**.
[131] B. Medasani, A. Gamst, H. Ding, W. Chen, K. A. Persson, M. Asta, A. Canning, M. Haranczyk, *npj Comput. Mater.* **2016**, *2*, 1.
[132] Q. Zhao, S. Stalin, C.-Z. Zhao, L. A. Archer, *Nat. Rev. Mater.* **2020**, *5*, 229.
[133] G. Ceder, S. P. Ong, Y. Wang, *MRS Bull.* **2018**, *43*, 746.
[134] T. P. Senftle, S. Hong, M. M. Islam, S. B. Kylasa, Y. Zheng, Y. K. Shin, C. Junkermeier, R. Engel-Herbert, M. J. Janik, H. M. Aktulga, T. Verstraelen, A. Grama, A. C. T. van Duin, *npj Comput. Mater.* **2016**, *2*, 15011.
[135] D. Bedrov, J.-P. Piquemal, O. Borodin, A. D. MacKerell, B. Roux, C. Schröder, *Chem. Rev.* **2019**, *119*, 7940.
[136] R. Jinnouchi, F. Karsai, G. Kresse, *Phys. Rev. B* **2019**, *100*, 014105.
[137] I. S. Novikov, K. Gubaev, E. V. Podryabinkin, A. V. Shapeev, *Mach. Learn.: Sci. Technol.* **2021**, *2*, 025002.
[138] C. Wang, K. Aoyagi, P. Wisesa, T. Mueller, *Chem. Mater.* **2020**, *32*, 3741.
[139] X. He, Q. Bai, Y. Liu, A. M. Nolan, C. Ling, Y. Mo, *Adv. Energy Mater.* **2019**, *9*, 1902078.
[140] C. Chen, Y. Zuo, W. Ye, X. Li, S. P. Ong, *Nat. Comput. Sci.* **2021**, *1*, 46.
[141] B. J. Shields, J. Stevens, J. Li, M. Parasram, F. Damani, J. I. M. Alvarado, J. M. Janey, R. P. Adams, A. G. Doyle, *Nature* **2021**, *590*, 89.

Filip Dinic received his B.S. degree in material chemistry from the University of Toronto in 2018. Currently he is pursuing his Ph.D. in chemistry under the supervision of Prof. O. Voznyy at the University of Toronto. His research interests focus on energy storage and novel materials discovery using machine learning.

Kamalpreet Singh earned his M.Sc. in Chemistry from the University of Toronto. He is currently a Ph.D. candidate at the University of Toronto working under the supervision of Oleksandr Voznyy, where he is using density functional theory and machine learning to augment material discovery in clean energy. His primary research interests include machine learning models for fundamental material properties, optoelectronic materials, and economical catalysts for water-splitting.



Tony Dong is currently a senior undergraduate student studying materials science and engineering at the University of Toronto. He is interested in developing materials for energy storage and accelerating his research efforts using machine learning.



Zhibo Wang is a Ph.D. student at the University of Toronto under the supervision of Oleksandr Voznyy. He has received his M.Sc. from the University of Alberta in theoretical/computational chemistry. Preferring digital chemistry to the lab, he currently is invested in improving materials prediction and property analysis via machine learning.

Oleksandr Voznyy Alex earned his Ph.D. in 2004 in physics of semiconductors from Ukraine. He then worked as a postdoc at the University of Sherbrooke, NRC Canada, and University of Toronto on a broad range of topics, from computational modeling of photoexcitations and surface properties of quantum dots to experimental implementations of solar cells, lasers, and catalysts for water splitting and CO_2 conversion. In 2018, Alex joined the Department of Physical and Environmental Sciences at the University of Toronto Scarborough as an Assistant Professor in Clean Energy. His topics of interest are materials for energy storage and novel materials discovery using high-throughput experiments and machine learning.