Batteries & Supercaps
Concept
doi.org/10.1002/batt.202200309
Chemistry Europe
European Chemical Societies Publishing
www.batteries-supercaps.org
Special Collection

Battery Materials Discovery and Smart Grid Management using Machine Learning

Andrew Jun Yao Wong,+,[a] Xin Zhou,+,[b] Yanwei Lum,[a] Zhenpeng Yao,[c] Yang Choo Chua,[d] Yonggang Wen,*[b] and Zhi Wei Seh*[a]

The transition from fossil fuels to renewable energy represents a grand challenge for humankind. For this vision to come to pass, significant advances in energy storage technologies such as batteries, which solve the intermittency of renewable energy, need to be achieved. Developing new battery materials with higher capacities and longer lifetimes is thus of paramount importance. Moreover, the intermittency of renewable energy presents a significant challenge to smart grid management. To this end, researchers have begun turning to machine learning (ML) techniques: algorithms that learn from datasets and automatically improve through experience. These can be used to make predictions and informed decisions, which can accelerate the process of materials discovery and systems management. Here we discuss key ML concepts that have guided important developments in battery materials discovery and smart grid management. In the process, we also examine critical challenges, future opportunities, and how ML can make a significant impact.

1. Introduction

The depletion of fossil fuels and environmental concerns have led to an increase in demand for renewable and sustainable energy sources.[1,2] However, effective utilization of these sources requires efficient and reliable energy storage methods. For example, lithium-ion batteries, which are a form of electrochemical energy storage technology, are a crucial part of everyday applications from powering electric vehicles to grid storage.[3–5] Therefore, developing new battery materials that have higher energy/power densities and longer cycle lives is of utmost importance. However, the large chemical space of possible electrode and electrolyte materials is intractable using traditional trial and error, necessitating a more efficient materials discovery approach.

At the same time, there have been recent movements to restructure energy systems to integrate more distributed generation and renewables.[6] Variable yields and power fluctuations caused by shifting insolation and regional diversity, however, pose major barriers to integrating sustainable energy sources such as solar and wind power. Such fluctuations often present stochastic difficulties for energy system management, giving rise to grid instability due to mismatched supply and demand.[7] Therefore, batteries are often used in smart grids to store energy during peak hours for later use. However, the increase in scale and introduction of hybrid renewable energy sources have made modeling using traditional techniques more complicated.[8,9] As a result, there is a critical need for novel smart grid management approaches from both the supply and demand sides.

Machine learning (ML), which is inherently data-driven, represents a possible approach to tackle these challenges. ML algorithms can learn from datasets and automatically improve through experience to predict new battery materials with promising properties, which can accelerate the process of materials discovery and development.[10,11] Besides the predictive capabilities of ML using discriminative models, the use of generative models can also enable the autonomous inverse design of battery materials in a closed loop.[12,13] In addition, ML can also model renewable energy systems so that their power outputs can be more accurately predicted. Compared to traditional model-based techniques, ML-based forecasting/prediction of the state of renewable energy systems has become increasingly more accurate, owing to considerable advancements in the field of artificial intelligence (AI) in recent years.[14–18]

In this Concept article, we will discuss key ML concepts that have guided important developments in battery materials discovery and smart grid management. First, we examine the most representative works, critical challenges, and future opportunities of ML in battery materials research. Key ML concepts including discriminative ML, generative inverse design, explainable AI, data management, machine-learned potentials, and ML integrated robotic platforms will be

[a] A. J. Y. Wong,+ Prof. Y. Lum, Prof. Z. W. Seh
Institute of Materials Research and Engineering, Agency for Science, Technology and Research (A*STAR), 2 Fusionopolis Way, Innovis, Singapore 138634, Singapore
E-mail: sehzw@imre.a-star.edu.sg
[b] Dr. X. Zhou,+ Prof. Y. Wen
School of Computer Science and Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798, Singapore
E-mail: ygwen@ntu.edu.sg
[c] Prof. Z. Yao
The State Key Laboratory of Metal Matrix Composites, School of Materials Science and Engineering, and Center of Hydrogen Science, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, China
[d] Dr. Y. C. Chua
College of Engineering, Nanyang Technological University, 70 Nanyang Avenue, Singapore 637457, Singapore
[+] These authors contributed equally to this work

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

discussed. Next, we elaborate on current ML efforts in managing smart grids from both the supply and demand sides, as well as existing challenges of data scarcity and risk-averse mindset that hinder practical real-world deployment. We offer some perspectives on overcoming these challenges using a generalized framework of systems-level optimization based on an industry-grade digital twin, together with energy-intensive data centers as an "interruptible load" to stabilize and regulate these grids. We note that this Concept article is not meant to be a comprehensive literature review; instead, the aim is to highlight important ML concepts in materials- and systems-level research to spur future progress in sustainability.

2. Machine Learning in Battery Materials Discovery

The traditional approach in battery materials research is typically Edisonian in nature (Figure 1). Possible material candidates are selected from a large chemical space, then synthesized, characterized, and finally, tested to determine their electrochemical performance. This process of trial and error is slow and tedious, typically occurring over a 15 to 25 year time horizon.[19] High-throughput computations can enable faster turnaround by using theoretical calculations to screen potential battery materials, followed by experimental validation of the most promising candidates.[20,21] However, this approach is computationally costly and only explores a limited chemical space.

Since ML is inherently data-driven, it has emerged as a compelling approach, to facing these challenges.[22,23] ML is capable of finding patterns in large datasets, which correspond to structure-property relationships in materials.[24–26] After discriminative ML models are trained using data, they can be used to predict other potential candidates in the large chemical space. An even more promising approach is known as automated virtual screening using generative ML models (Figure 1). This method is capable of predicting an ideal structure when provided with just the specified materials property, thus enabling the inverse design of materials (i.e. property-to-structure approach).[27,28]

3. Discriminative Machine Learning

Discriminative ML models can determine conditional probabilities from training datasets, allowing them to predict a property *y* from a data point *x*.[24,25] As mentioned in the earlier section, the standard process of battery research is incredibly slow and as such, discriminative ML models can be used to speed up the pace of battery research. These ML models do this by improving the rate at which screening of materials is done and thus, the rate at which predictions regarding new battery materials are made. These ML models have been able to predict

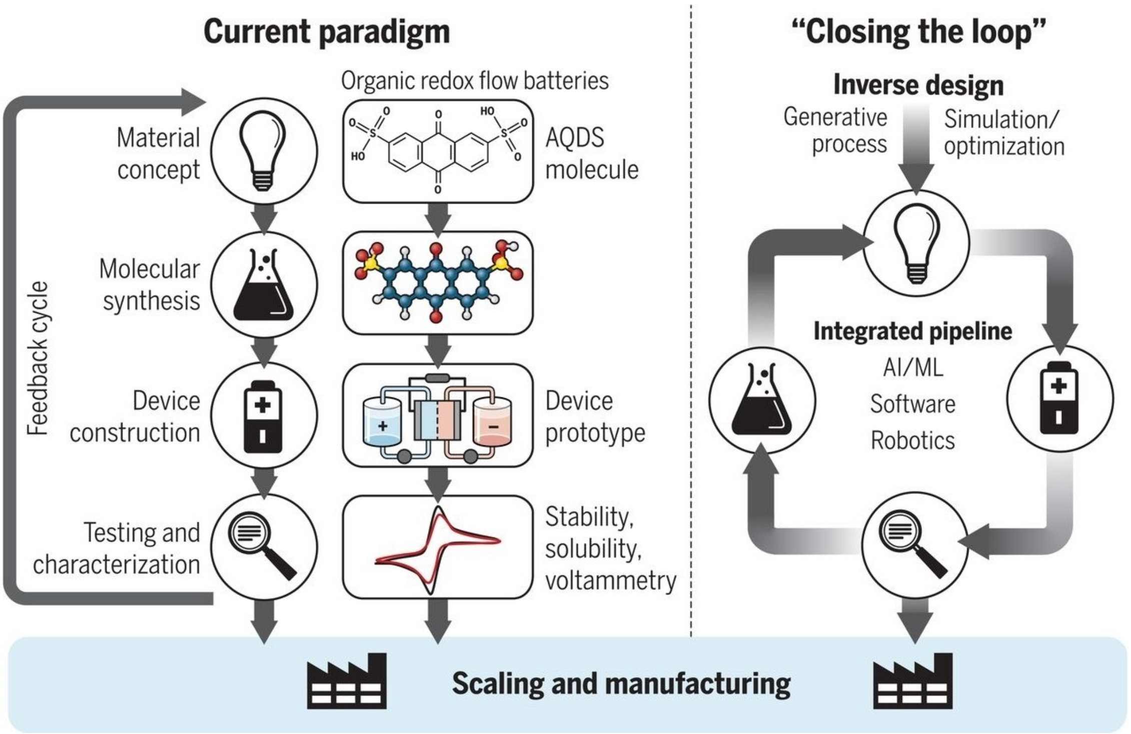

Figure 1. Schematic comparison of battery materials discovery paradigms. The current paradigm is outlined at left and exemplified in the center with organic redox flow batteries. A closed-loop paradigm is outlined at right. Closing the loop requires incorporating inverse design, smart software, AI/ML, embedded systems, and robotics into an integrated ecosystem. Reproduced with permission from Ref. [27]. Copyright (2018) The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science.

new materials for both the electrodes and the electrolytes to great success.[29,30]

For example, a work from 2019 employed several ML algorithms including deep neural networks (DNNs), support vector machine (SVM), and kernel ridge regression (KRR), to predict the voltage profiles of electrodes using feature vectors as their representations, where predicted profiles show good agreement with computational results[31] (Figure 2). The number of density functional theory (DFT) calculations required to explore the chemical space was reduced, and up to 5,000 candidate electrodes for Na- and K-ion batteries were shortlisted. An artificial neural network (ANN) was also combined with DFT to design metal-free molecular electrodes for lithium-ion batteries using input descriptors such as highest occupied molecular orbitals, lowest unoccupied molecular orbitals, and atom numbers.[32] The electronic properties of candidate materials were calculated using DFT, which was then used to train the ANN to predict redox potentials. Another example is by Kim et al. who utilized ML models along with 4401 experimental datasets obtained from the Materials Project to design Ni-rich layered cathode materials with multivalent dopants that maximize energy density and stability.[33] They considered a total of 33 elements, which made up 1617 potential cathode materials. Each potential material had 47 different battery material features (charge and discharge formulae, energy density, capacity, etc.), 145 chemical descriptor features (electronic structure, elemental properties, ionic compound attributes, etc.), 126 structural descriptor features (crystal structure attributes, bond length, cell volume, etc.), and 3 additional features (space group number, degree of delithiation, and gravimetric capacity). The ML model utilized was a machine learning regression model using LightGBM. This model was able to obtain respective R^2 scores and mean absolute errors of 0.873 and 0.323 V for voltage, and 0.562 and 2.890 % for volume change. They were also able to narrow the field down to 107 possible materials with gravimetric energy densities of more than 875 Wh/kg, average voltages of more than 3.5 V, and volume changes of less than 7 %. All of these results were also verified by DFT calculations to ensure their accuracy.

Besides electrode materials, ML has also been used to aid the design of solid-state ion conductors which can be used as

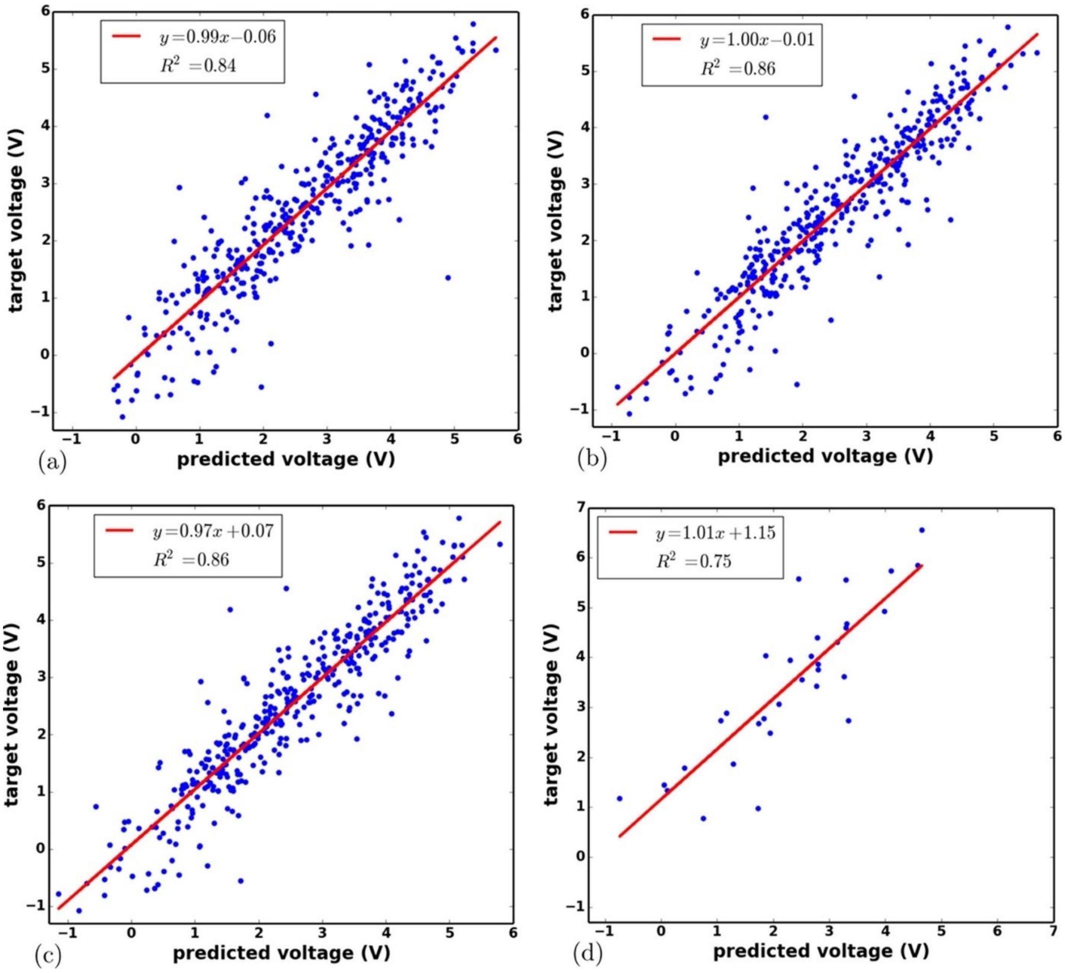

Figure 2. Architecture of the DNN used to predict the voltage of electrode materials. Scatter plot showing target vs. predicted voltages with different ML algorithms used in this work. a) DNN, b) SVM, and c) KRR, each on H-set. d) DNN on Na-set. The best-fit equation ($y = mx + c$) and R^2 values for linear fit between target and ML-predicted values are provided as an inset. Reproduced with permission from Ref. [31]. Copyright (2019) American Chemical Society.

solid electrolytes. For lithium-ion conductors, a transfer learning model was constructed to learn physical insights from existing crystal structures and then predict new compositions.[34] In a follow-up work, an ML-based predictive model was used to screen through 12,000 candidates with diverse structures and compositions, leading to the discovery of 10 new lithium-ion conductors.[35] To overcome the scarcity of available materials conductivity data, an unsupervised model was used to cluster the lithium-containing materials into high and low conductivity groups using descriptor inputs, after which ab-initio molecular dynamics simulations were then used to validate the classification, discovering 16 new promising lithium-ion conductors.[36] In addition, ANN models were also created for the optimization of electrode mesoscale structures by predicting their charge/discharge specific resistances, which agreed well with the simulated values.[37] A graph convolutional network was also shown to be able to predict inorganic solid electrolytes with high lithium-ion conductivity and dendrite-suppressing capability, by using computational data on the mechanical properties of over 12,000 inorganic solids.[38] Another showcase of ML being used in electrolyte design is by Zhang et al. who utilized Gaussian process regression models to predict the redox potentials of electrolyte additives for lithium-ion batteries from the molecular structural features of the electrolyte additives.[39] The dataset consisted of 149 electrolyte additives, with a total of 21 reported features that describe the molecule from the coordination number of each element in the molecule, to the number of times the element occurs in the molecule to the structure that the molecule is in (five-membered rings vs. six-membered rings). The model was able to obtain low RMSEs scores of 0.08 and 0.14 for V_{red} and V_{ox} respectively. These values were not only verified by DFT but it was also found that these Gaussian process regression models were possible efficient alternatives to DFT, due to being faster and requiring lesser computing power. Therefore, these examples showcase how ML can be an integral part of battery research.

4. Generative Inverse Design

The ultimate goal of battery materials research is generative inverse design, which can enable the automated virtual screening approach.[27,28] In conventional 'direct' materials design, the starting point is a specified region within chemical space. The properties of materials within this space are first predicted and experimentally verified. On the other hand, inverse design operates in reverse, whereby the desired functionality is first declared and from there, this information is used to predict materials that exhibit the required properties.[28] This inverse design methodology has a distinct advantage over conventional direct design. In conventional direct design, it is unfeasible to search the entire chemical space, due to it being intractably large. The search space is thus typically narrowed down using intuition and domain expertise. Unfortunately, this bias means that only the chemical neighborhood of known materials is explored, which inadvertently leads to the exclusion of unexpected candidates with potentially impressive properties.[28] Inverse design circumvents this issue by starting from the desired functionality and uses this to predict promising candidates within the entire chemical space.

The challenge is to construct robust approaches for effective inverse design, which can be thought of as an optimization approach for exploring the 'functionality manifold', without having to test all possible candidates in the chemical space.[27] Inverse design using ML employs deep generative models, which can produce large numbers of candidates by learning the joint probability distribution of x and y, the materials descriptor and property, respectively. Examples of such generative models are variational autoencoders (VAEs), reinforcement learning, and generative adversarial networks.[27]

The key to the application of inverse design is the use of appropriate 'representations' for materials, which need to be invertible from representation to material.[40] Inorganic solid-state materials are crucial in battery research. However, these materials lack suitable invertible representations, hindering inverse design. Recently, an image-based invertible input representation for materials crystal structure was developed, which contains both cell and basis information.[40] This allowed for the identification of more than 40 stable V_xO_y structures that have never been reported in the literature. Notably, this approach was found to have superior prediction efficiency compared to a genetic algorithm. Another recent work created an automated materials discovery platform for reticular frameworks, such as metal-organic and covalent-organic frameworks, using a supramolecular VAE[41] (Figure 3). Crucial to this was the development of RFcode: an invertible representation that captures all-important structural information efficiently without redundancies. RFcode is also unique such that each representation encodes a unique framework. This enabled the design of complicated reticular frameworks, with top candidates confirmed to be exceptional CO_2 absorbent materials. In the future, extending this promising approach to battery applications will likely require the development of generalized representations for more solid-state materials.

5. Explainable Artificial Intelligence

ML can aid battery materials design and discovery by enhancing property prediction and generating new molecules and crystals. However, due to the complexity of the algorithms being used, it is often not known how the algorithms employed arrive at their decisions, even to the designers of the algorithms themselves. Using solid-state ionic conductor design as an example, after proper training, the ML model successfully yields multiple new candidate materials with fast Li-ion transport kinetics.[36,42] However, it is not known how the various elemental combinations affect local transport in the crystal structure in the top candidates, allowing for high Li-ion mobility. Clearly, a better understanding of the underlying latent factors responsible would greatly benefit efforts in rational materials design.

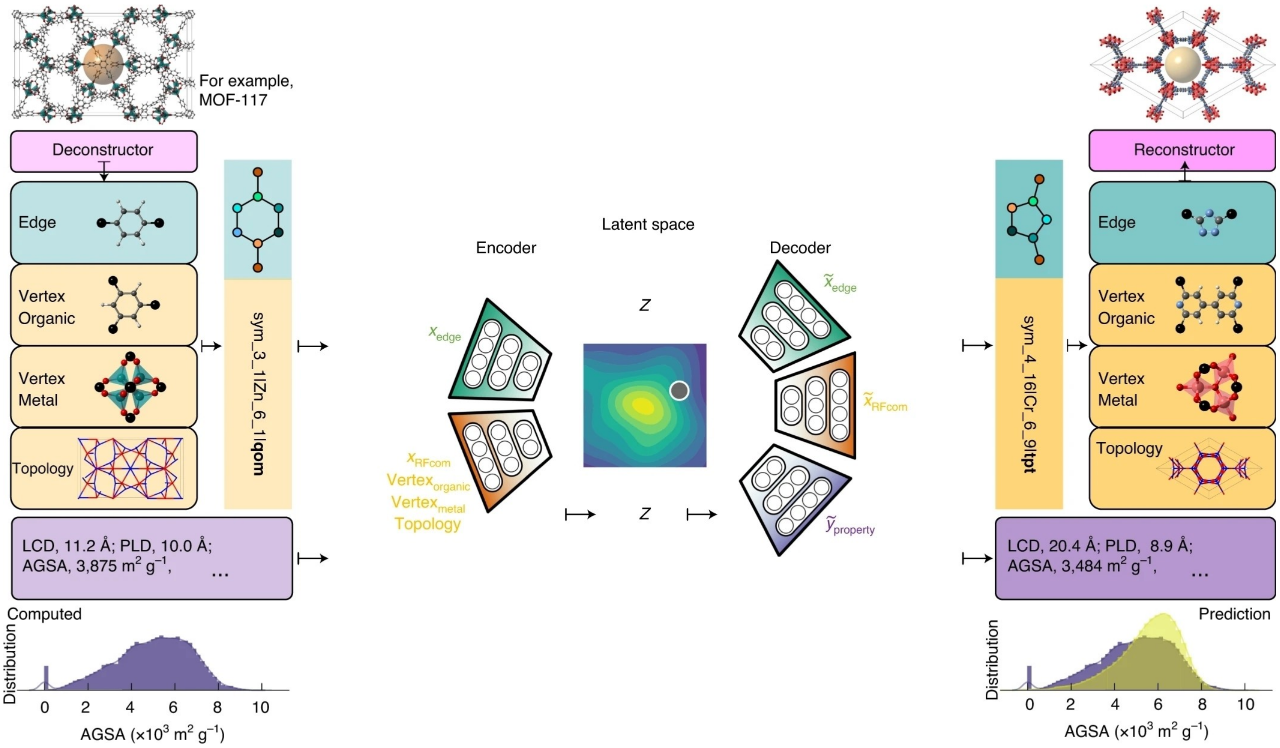

Figure 3. Schematic of the automated reticular framework discovery platform empowered by the SmVAE. The SmVAE is a multi-component variational autoencoder with modules that are in charge of encoding and decoding each part of the RFcode ($x_{\mathrm{edge}} \rightarrow \tilde{x}_{\mathrm{edge}}, x_{\mathrm{RFcom}} \rightarrow \tilde{x}_{\mathrm{RFcom}}$). Reticular frameworks are mapped with discrete RFcodes, transferred into continuous vectors (z) and then transferred back. To have the latent space organized around properties of interest, we add an extra component to the model that uses labelled data (y). This process is realized with the additional model that learns to predict properties ($\tilde{y}_{\mathrm{property}}$) from the latent space. RFcom: components of RFcode except edge. Topology: qom, tpt. Reproduced with permission from Ref. [41]. Copyright (2021) The Author(s) under exclusive licence to Springer Nature Limited.

This is what explainable AI (XAI) is trying to achieve. XAI is an emerging field in ML that desires to enhance the interpretability of ML algorithms, which are often "black boxes".[43] XAI's primary objective is to fade these "black boxes" by deciphering the factors surrounding the decision-making process of these ML algorithms, which in turn would not only make them easier to use but also make their results more trustworthy to the users.[44–46] XAI does this by deriving feature relevance and score importance from ML models to identify the level of significance of each feature to the result and its dependence to other features.

One such method is the random forest method. In the random forest method, it consists of the combined usage of many decision trees where each individual tree depends on the values of a random vector that is sampled independently and with the same distribution for each tree. This vector determines the features that determine how the data is split at each node in the tree. Due to the strong law of large numbers, although there would be some trees that are more accurate and other trees which are inaccurate, the trees are generally able to protect each other from individual mistakes, thus the resulting combined performance of all the trees would be more accurate as compared to the individual performance of any individual tree.[47] The feature importance can be ascertained by the change in prediction error between trees, where the rationale is that important variables tend to be split more often and as such, affect the prediction more significantly as opposed to other less important variables.[48] These feature importance scores thus provide a relative ranking of the various features, shedding light on which features are more important and which are less so.[49]

Another example is symbolic regression, which aims to capture the performance of the data with an algebraic expression. Due to there being an infinite number of expressions when all the input expressions are considered, the Edisonian approach to determining the ideal expression is not realistic. As such, symbolic regression utilized genetic programming to search for the ideal expression more efficiently. Initially, the population is comprised of expressions that are determined entirely at random, and the performance of each individual is determined based on their prediction error. After each round, individual expressions are given the chance to undergo either crossover, where two parent individuals are used to create two new child individuals using the subtrees of both parents, or mutation, where the expression would undergo random changes. There needs to be a balance to the rate of occurrence of the two as too much crossover would prevent the population from finding new combinations that might perform better than the current population, and on the other hand, too much mutation can lead to the process becoming more Edisonian-like, negatively affecting the convergence rate. However, once a proper ratio is determined, allowing for convergence to occur, it would be possible to determine the relative importance of each feature, based on the expression alone (Figure 4).[43,48]

Beyond these examples, there are other promising XAI methods including local interpretable model-agnostic explanations, shapley additive explanations, layer-wise relevance propagation, and deep Taylor decomposition.[50–53] If available, domain knowledge can also be used in a hybrid approach with ML to introduce constraints on the system, reducing the likelihood of overfitting and making models more explainable and interpretable.[54] For example, regularized linear framework models incorporating domain knowledge have exhibited notable accuracy for battery lifetime early prediction using only data from the first several cycles.[55,56]

6. Data Management

Before discussing the various fields in which ML has been used, an important caveat is that the data utilized when employing ML is of adequate quality. Data that contains either too few data points or too many low-quality data points, which are data points that are either irreproducible or affected by significant errors, can lead to inaccurate ML predictions, ultimately skewing the interpretation of the result. The first step in ensuring good quality data is by having proper data collection procedures. For example, in the case of experimental design, experiments need to be planned and described in such a way that they are replicable, such that the results obtained can be reproduced by other researchers. This would allow for the results of different research teams to be compared to one another with minimal confounding. Another thing that can be done is in terms of variable reporting. In the world of ML, it is a good practice to consider as many variables as possible to ensure that the issue is considered from as many angles as possible.

For example, in order to accurately predict battery materials properties, ML should be trained with large amounts of diverse and high-quality datasets, which can be obtained from various sources.[24] One source is the experimental data that is available in the literature, which can be collated and used to predict the properties of known compounds or for the discovery of new ones.[19] Text and natural language processing can be used to extract relevant data from the literature, which may even reveal 'hidden' relationships between different material properties.[57] Structure-property relationship databases such as Materials Project[58] and Harvard Clean Energy Project[59] also contain useful datasets for ML-driven materials discovery. Alternatively, high-throughput DFT calculations can be carried out in a targeted fashion within the desired chemical space, to generate data for learning.[27] While the increase in the number of features can lead to complexity issues when it comes to utilizing unsupervised techniques, feature selection can be done beforehand to remove redundant and unimportant features before testing is done. The better the dataset, the more accurate the ML model would be. After the data has been

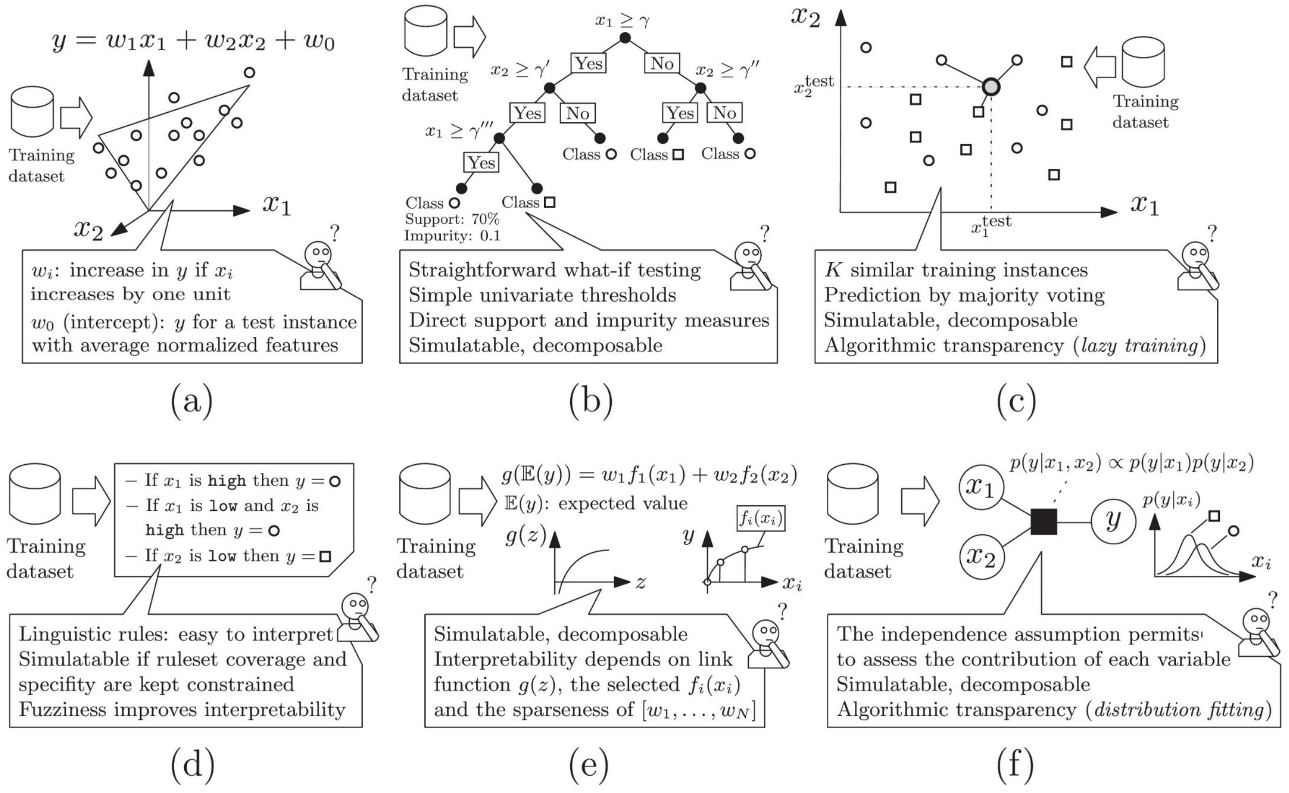
$y = w_1x_1 + w_2x_2 + w_0$
Training dataset
x_1
x_2
?
w_i: increase in y if x_i increases by one unit
w_0 (intercept): y for a test instance with average normalized features
(a)
Training dataset
$x_1 \geq \gamma$
Yes
No
$x_2 \geq \gamma'$
$x_2 \geq \gamma''$
Yes
No
Yes
No
$x_1 \geq \gamma'''$
Yes
Class ○
Class □
Class ○
Class ○
Class □
Support: 70%
Impurity: 0.1
?
Straightforward what-if testing
Simple univariate thresholds
Direct support and impurity measures
Simulatable, decomposable
(b)
x_2
x_2^{test}
Training dataset
x_1
x_1^{test}
?
K similar training instances
Prediction by majority voting
Simulatable, decomposable
Algorithmic transparency (*lazy training*)
(c)
Training dataset
– If x_1 is high then $y = $ ○
– If x_1 is low and x_2 is high then $y = $ ○
– If x_2 is low then $y = $ □
?
Linguistic rules: easy to interpret
Simulatable if ruleset coverage and specifity are kept constrained
Fuzziness improves interpretability
(d)
Training dataset
$g(\mathbb{E}(y)) = w_1f_1(x_1) + w_2f_2(x_2)$
$\mathbb{E}(y)$: expected value
$g(z)$
z
y
$f_i(x_i)$
x_i
?
Simulatable, decomposable
Interpretability depends on link function $g(z)$, the selected $f_i(x_i)$ and the sparseness of $[w_1, \ldots, w_N]$
(e)
Training dataset
$p(y|x_1, x_2) \propto p(y|x_1)p(y|x_2)$
x_1
x_2
y
$p(y|x_i)$
x_i
?
The independence assumption permits to assess the contribution of each variable
Simulatable, decomposable
Algorithmic transparency (*distribution fitting*)
(f)

Figure 4. Graphical illustration of the levels of transparency of different ML models. a) Linear regression; b) Decision trees; c) K-nearest neighbors; d) Rule-based learners; e) Generalized additive models; f) Bayesian models. Reproduced with permission from Ref. [43]. Copyright (2019) Elsevier B.V.

collected, the data also needs to undergo preprocessing, where these preprocessing steps ensure that the collected data is reliable and valid for use.[60]

There are 5 main preprocessing steps that ensure good data, namely, data cleaning, data reduction, data scaling, data transformation, and data partitioning.[61–63] Firstly, data cleaning refers to missing value imputation and the removal of outliers. There are 2 approaches to missing data, either simply removing the data point or filling in the missing values by inferring them from the rest of the dataset. While the former should only be done if the number of data points with missing values is insignificant to the total number of data points, the latter can always be done. Data imputation can be done via univariate methods such as replacing the missing values with median or mean values, or via multivariate methods such as *k*-nearest neighbour and regression model-based methods.[63–65] As for outlier detection, there are 2 main approaches, statistical methods, and clustering-based methods. Statistical methods are typically used for numerical data and assumes that the data follows a normal distribution, whereas clustering-based methods are able to more accurately detect outliers from non-uniformly distributed data, however, it has the drawback of being much slower and requires more computing power.[66,67]

Secondly, data reduction refers to feature selection and feature extraction. This is done to prevent the data from having an excessive number of features which can in turn lead to the models becoming overfitted. The two main methods in feature selection include the filter method and the embedded method.[68] The filter method involves evaluating each feature based on statistical theory and information theory to determine the importance of the feature to the data, keeping only the most important features in the final dataset. As for the embedded method, the model iteratively removes/adds features based on the search strategy that is being employed, and at every step, checks if the performance of the model has been improved. If the model is improved, the step is accepted and the process continues, else, the step is rejected, and the next feature is evaluated. As for feature extraction, it combines features together based on the linear or non-linear relationships between pre-existing features to ultimately reduce the total number of features, while still ensuring that they are taken into account.

One promising feature selection method was suggested by Liu et al. who suggested utilizing a multi-layer feature selection method that incorporated a weighted score-based system that depended on expert knowledge.[68] Although this is a promising method as this greatly increases the speed of the feature selection process, it is however heavily reliant on the assumption that the experts are correct. While many of the experts throughout all the different fields of research may have some differing views, most of them would agree with the more established ideas. Therefore, if experts who already believed the same thing are used to do weighted score-based feature selection, chances are, the features that they already know are important are the features that would be deemed as important by the method. However, not only have there been instances in the past where experts have been wrong, but this also defeats the primary purpose of ML where ML is meant to scour the features that we might have overlooked and draw links that we might never have drawn due to our existing preconceptions on the topic. Therefore, by doing weighted score-based feature selection, we might end up inputting our own bias into the ML model, which can ultimately lead to inaccurate results. However, in spite of this, the increase in computing speed is not something that should be dismissed. As such, if time is a limitation, this method would be ideal as it would be able to quickly provide an approximate feature set that is on the right track, but if more time is available, it would be better to do feature selection without the weighted scoring of the experts. One possible way to improve this could be by adding a "redemption" step at the end, where all the features which have been excluded are given a second chance to be included in the final feature set. If any of the features whereupon being reincluded results in an improvement in the model's accuracy, it is kept in instead of being excluded. This additional step would be able to help in preventing the exclusion of features that are important but were not accurately weighted by the experts. Since this step is also not computationally expensive, it should be able to be added in without a significant increase in computing time.

Thirdly, data scaling refers to transforming the data within each column to be in similar ranges to prevent inherent bias. This can be done either by min-max normalization, where all the values are scaled down to fit within a specified range, or by z-score standardization, which takes into account the mean and the standard deviation of the column.[69,70] Fourthly, data transformation refers to the transformation between numerical data and categorical data depending on the algorithm in use. This is done to ensure that the data is compatible with the algorithms in use as some algorithms are only able to accept numerical inputs whereas others are only able to accept categorical inputs. Lastly, data partitioning refers to splitting up the data into different groups to allow for more in-depth data analysis. This step segments the data based on characteristics specific to the situation which would allow for a more accurate understanding of the particular situation in question.[71]

However, despite having good data management procedures in place, data scarcity is still a problem and can occur due to three typical issues. Firstly, data scarcity is unavoidable when exploring novel applications. This is exacerbated when neural networks such as DNNs are used, which demand a larger amount of training data for accuracy. Secondly, negative data, such as a failed synthesis experiment, is usually not reported even though these are as equally important in ML as positive data. This scarcity of negative data affects the training quality, especially for classification models. To underscore the importance of negative data, researchers used an SVM trained with historical data consisting of both successful and failed synthesis experiments.[72] Interestingly, they were able to predict conditions for crystal synthesis with an 89% success rate, compared to only 78% when relying purely on human intuition. Thirdly, for ML-based multi-application materials design with a large number of targets, the amount of data required is larger than what might typically be available.

Data scarcity calls for robust neural networks to maximize learning from the available data, as well as improved data augmentation methods; examples are recent efforts in meta-learning.[73–75] In addition, active learning strategies,[76,77] which although require some amount of human intervention, could be one approach to reduce the amount of training data required. Transfer learning is also another promising approach to overcoming data scarcity, by applying ML models trained on datasets of related properties to the problem at hand.[78,79]

Increasingly, there have been calls for more open and effective data sharing, according to the 'FAIR' data principles.[80] This means that data should be 'Findable' and easily 'Accessible' for interested parties. Data must also be represented according to established standards, making it 'Interoperable' and thus 'Reusable' for other purposes beyond their original research intent. It is anticipated that widespread adoption of these 'FAIR' principles can greatly boost ML-driven research for materials discovery.

7. Machine-Learned Potentials

In this section, we want to discuss machine-learned potentials and their applications in material design. This emerging field of ML potentials promises to be able to match the accuracy of quantum mechanical computations without the massive computational cost that typically comes with such computations.[81] By utilizing ML models such as neural networks and kernel methods, they can construct ML potentials that are trained from ab initio data, that can be used to predict energies and forces in atomically resolved systems. Additionally, due to ML not distinguishing between bonded and non-bonded interactions, it can also simulate chemical reactions.

For example, Bereau et al. utilized a combination of physics-based potentials with an ML model, kernel ridge regression, to provide predictions for environment-dependent local atomic properties such as electrostatic multipole coefficients, the population and decay rate of valence atomic densities, and the polarizabilities across conformations and chemical compositions of H, C, N, and O atoms.[82] These predictions allowed for the accurate calculation of intermolecular contributions like electrostatics, charge penetration, repulsion, induction/polarization, and many-body dispersion. Uniquely, this model was also able to process new molecules in new conformations without any additional parametrization. They tested this model by comparing its ability to determine the energies between an array of molecules from dimers in DNA-base and amino-acid pairs to water clusters with up to 10 molecules, with the model obtaining a mean absolute error of 1.4 kcal/mol for the dimer test and 8.1 kcal/mol for the water cluster test.

Another example is by Hajibabaei et al. Who applied ML potentials using a sparse Gaussian process regression algorithm to sift through about 300 different ternary crystals to determine which among them had the potential to be a viable solid electrolyte.[83] They narrowed down the 300 candidates to 22 based on their Li-ion conductivity, which is an important characteristic to possess for a solid electrolyte. Of the 22, it consisted of 14 sulfides, 5 halides, 2 oxides and 1 selenide, where many of the ternary crystals narrowed down by the program like Li_3PS_4, Li_7PS_6, and $Li_7P_3S_{11}$, were already well documented and reported to have impressive ionic conductivities, thus, showcasing the viability of the method.[84,85]

8. Machine Learning Integrated Robotic Platforms

It is anticipated that with the rise of ML, robotics, and smart automation could herald a new era of autonomous battery materials discovery in 'self-driving' laboratories.[86,87] The key to realizing this vision is the development of robotic platforms for automated high-throughput materials synthesis and characterization. With advancements in combinatorial synthesis, battery materials of various phases, compositions, and dopants can be prepared and optimized. This is made possible with robotic platforms that use a variety of high-throughput synthesis techniques such as thin-film sputtering, jet dispensing, and pulsed laser deposition.[88,89] Characterization of battery materials can also be carried out in an automated high-throughput manner,[90,91] for instance, by coupling X-ray diffraction and X-ray absorption spectroscopy with an automated rotating sample changer (Figure 5). Electrode and cell production can be designed into a fully automated and integrated process as well, using emerging robotic technologies such as the cut-and-place module, continuous Z-folding, and stack winding process.[92] Importantly, these robotic platforms can be integrated with ML, which enables automated synthesis and characterization planning, as well as continual optimizations for materials discovery and development.

However, developing robotic platforms might be challenging, for example, different materials generally require distinct processing conditions for optimal performance. Optimal compositions may also exhibit poor performance if poorly synthesized, which could happen if experimental conditions were not ideal, leading to a high defect density. This could result in wrong predictions by the ML model in subsequent iterations. Additionally, materials that are identified with ML could be metastable,[93] which could be difficult to construct from high-throughput or simple combinatorial synthesis methods and thus motivates further research. Depending on the synthesis procedure, nominally identical samples could also have very different nanostructures, morphologies, and phases; factors that can dictate the properties of a material.[28] Therefore, this likely requires the development of comprehensive high-throughput characterization systems to uncover as many material descriptors as possible for training the ML model.

If one considers how a scientist performs materials design, a starting point would be a thorough literature review and a promising pathway will be decided upon based on chemical intuition. The cycles of synthesis→measurement→adjustment may be repeated multiple times until the target material (of correct phase) is synthesized. Based on this, one can consider how a robot scientist can be designed to fulfill the same

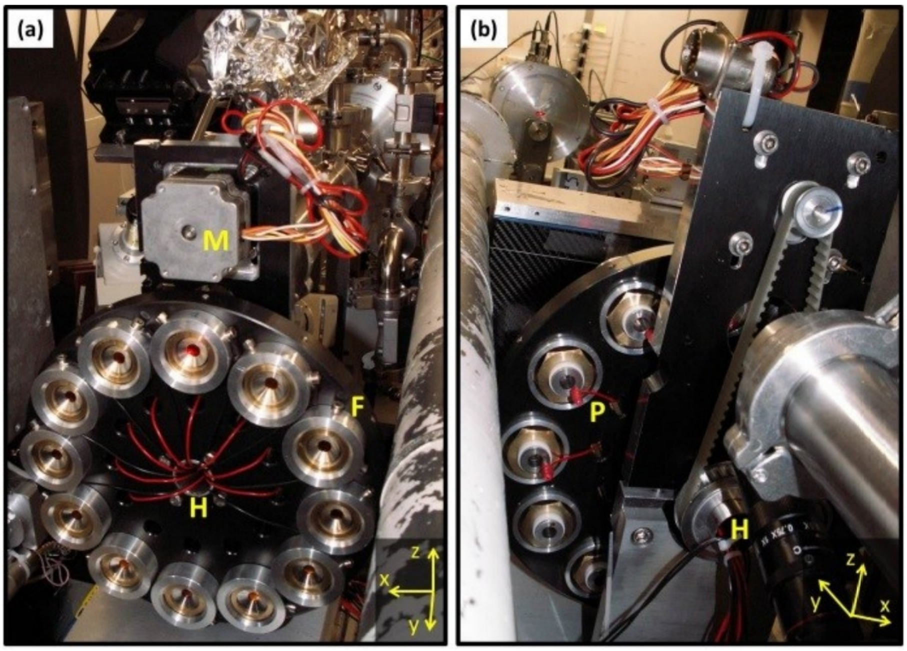

Figure 5. High-throughput setup for operando X-ray characterization of electrode materials. Photographs of a) front and b) rear of the of the automated sample changer. Stepper motor **M**, fixing screw **F**, hollow hub of the wheel **H** and banana socket **P**. Reproduced with permission from Ref. [91]. Copyright (2016) Int Union Crystallography.

goals,[94,95] with the aid of ML. First, the robot scientist can perform a literature search via text mining and obtain a detailed database on molecules with similar properties that have already been reported.

Based on this, it can choose suitable ML algorithms for training, which allows it to suggest promising pathways for initial experiments. From these, material synthesis and testing will be conducted in fully automated labs, with results fed back to the inverse design model, which then suggests the next round of synthesis experiments. It will be likely that the robot scientist will encounter unexpected situations and as such, occasional human interference will likely be required in the initial stages. Once the target material is synthesized, the robot scientist can report the best synthesis routes and provide explanations for why these routes are better.[43] Over time, the robot scientist will become more experienced, just like human scientists. We anticipate that due to the ability to perform a multitude of experiments in parallel and optimize designs on the fly, the fully automated android scientist will significantly accelerate the pace of scientific discovery.

In summary, an important framework on the battery materials discovery front is the ML-directed exploration of large chemical spaces. Enabling this requires designing unique and invertible representations that encode materials information efficiently, enhancing model interpretability, developing methodologies that can work with limited data, and engineering automated high-throughput experimental platforms.

9. Machine Learning in Smart Grid Management

In the following sections, we will discuss the application of ML algorithms in smart grid management. Sustainable energy resources are rapidly being introduced to the power system as carbon neutrality becomes a critical criterion for monitoring the grid. Fluctuations in intermittent renewable energy sources, however, pose a detrimental influence on stability and cause the energy system to be sensitive to weather/climate change.[96] Energy storage technologies, such as grid-scale battery systems, have been prominently used to address this problem since they enable the load to be adapted with time to meet supply and demand, to smooth out power output swings.[97] Current energy systems have also been designed[98] to forecast the quantity of electricity generated from renewable sources for energy management and distribution. As previously mentioned, substantial research has established ML methods to be more accurate and scalable for energy system power/fluctuation prediction than traditional model-based approaches.

From the supply management side, ML algorithms have been successfully implemented in renewable energy prediction. Traditional ML techniques, for example, offer to predict hourly solar irradiance and estimate energy generation with an accuracy of 95%.[99] The research has further adopted deep learning algorithms (e.g., long short-term memory and AutoEncoder) to forecast the energy output of 21 solar plants, with an average prediction error of less than 10%.[100] More recent research has focused on adopting ML-based algorithms to optimize automatic generation control of smart grids. Relaxed deep learning, for instance, has been applied as a real-time economic generation dispatch and control framework to decrease operating costs relative to conventional generation control frameworks.[101] A new algorithm based on reinforcement learning called correlated Q(λ) learning has also been used for automatic generation control, showing enhanced performance relative to other conventional algorithms.[102]

On the demand side of management, energy loads can be classified as rigid or interruptible[103] based on their urgency. Rigid loads are those that need to be met (e.g., lighting), whereas interruptible loads (e.g., vacuum cleaners) have the flexibility to be combined, divided, shifted, or rescheduled to achieve a smoothed demand profile and reduce peak energy demand.[104] Several methods have been proposed in this regard to precisely predict the demand load by using ML technologies. For example, DNNs have been used for building energy load forecasting that reduces the prediction error to 10%,[105] as well as load prediction of individual consumers' electricity consumption that exhibits an average error of 3.2%.[106] A deep belief network is also capable of modelling building energy consumption and using transfer learning to perform cross-building energy predictions with ~90% improvement in accuracy.[107] Accurate demand/load prediction has resulted in the empowerment of more complex decision-making operations in energy systems.[108] For example, a microgrid system can utilize reinforcement learning for demand scheduling while considering the dynamic pricing of the smart grid to reduce system costs by 25% compared to the conventional Q-learning algorithm.[109] As residential buildings play a critical role in the power grid, decentralized demand response has been used in smart grids to maximize renewable wind energy usage.[110] Multi-agent load forecasting can further assist in demand scheduling by using reinforcement learning to control the electricity needs of devices.[111]

10. Data Scarcity and Risk-Averse Mindset

Current research recognizes the critical role played by AI for a wide range of applications in the sustainable energy economy (e.g., quick wind speed prediction for siting wind farms,[112] accurate renewable energy availability forecasting,[113] and residential load forecasting[114]), yet AI deployment on real-world scenarios remains one of the most difficult challenges to tackle. Data and models are two crucial components of existing AI algorithms. Data-driven approaches can extract complex and non-linear patterns in the training dataset and convert the raw data into statistical models, which are then implemented to apply to various applications such as prediction, classification, or optimization. In real-world applications on physical systems, however, AI deployment faces two major inherent challenges, namely data scarcity and risk-averse mindset.

To ensure model quality, AI algorithms require a large amount of training data to enable a vast number of parameters to be updated precisely. In physical sustainable energy systems, however, data collection remains difficult.[115] In one case, as the cost of expensive measuring instruments (e.g., pyrheliometers for solar radiation) and expert manpower impedes system monitoring and data collection, researchers have resorted to estimating energy generation based on the weather and/or physical data.[116] Models are also known to deteriorate if there is only quantity but no diversity of datasets. Lengthy data collections over years are thus frequently performed to account for intermittency in solar and wind energy.[117] Furthermore, gathering data in certain situations (e.g., emergency faults, anomalies) can be extremely difficult and often result in system shutdowns and/or facility damages, both of which are costly and dangerous.[118] All these data acquisition challenges have adversely impacted AI deployment in real-world scenarios.

Decision-making is risky and a critical issue in sustainable energy systems. AI algorithms work on the assumption that all uncertainties in energy systems have been taken into account. However, due to limitations in historical data and the need for safety compliance, energy system operation continues to rely heavily on the decisions of human experts who are armed with domain knowledge and experience. This has resulted in a risk-averse mindset in the industry,[119] which has further impeded the industry's deployment of AI techniques. According to a report for the energy-intensive data center industry, the risk-averse mindset of operators necessitates a highly reliable operating environment with adequate redundancy before the energy efficiency of data center operations can be improved.[120] The above challenge warrants the need for novel approaches for the incorporation of AI solutions into the sustainable energy economy.

11. Digital Twin for Smart Grid Optimization

ML algorithms can be classified into three categories based on their practical functionality:[15] a) Descriptive AI aims to extract essential characteristics from operational data to identify system dynamics, which is potentially advantageous for physical system modelling; b) Prescriptive AI offers optimized operating policies to increase system efficiency according to monitored system states; c) Predictive AI attempts to figure out the causes of system state changes and simulates system behaviours to forecast future states based on hypothetical inputs. The above classification provides a more practical perspective compared to traditional AI classification based on supervised, unsupervised, and reinforcement learning. In smart grid management, for instance, these three AI capabilities can be leveraged to extract essential features of the power system to create a digital twin that mimics system dynamics, derives optimized control policy for improved system efficiency, and predicts both energy generation and consumption for balanced supply and demand.

The aforementioned three types of AI algorithms can be integrated into a unified framework centered on an industry-grade digital twin to enable advanced applications in the sustainable energy economy and ultimately overcome the intractable challenges described above. Figure 6 depicts the proposed framework, which includes the three modules listed below:

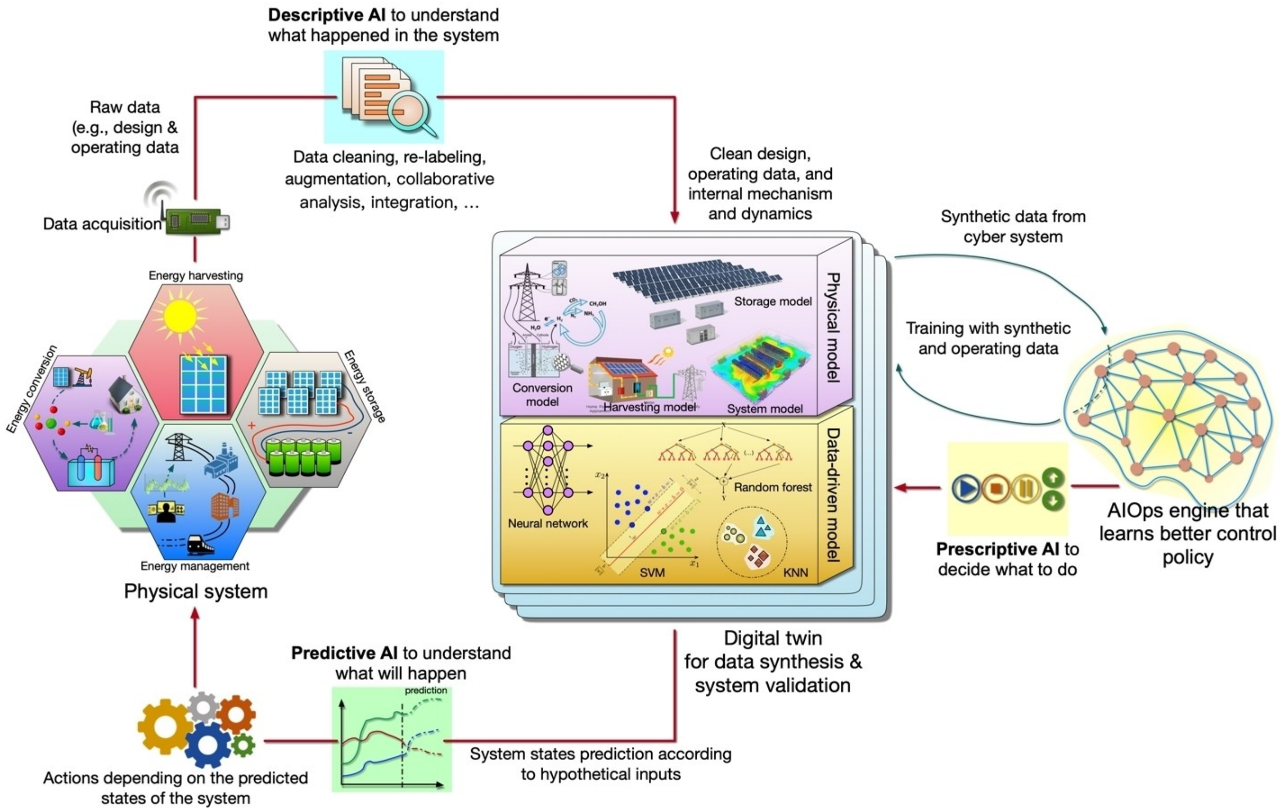
Descriptive AI to understand what happened in the system
Raw data (e.g., design & operating data)
Data acquisition
Data cleaning, re-labeling, augmentation, collaborative analysis, integration, …
Clean design, operating data, and internal mechanism and dynamics
Synthetic data from cyber system
Training with synthetic and operating data
Energy harvesting
Energy conversion
Energy storage
Energy management
Physical system
Physical model
Storage model
Conversion model
Harvesting model
System model
Data-driven model
Neural network
Random forest
SVM
KNN
Prescriptive AI to decide what to do
AIOps engine that learns better control policy
Digital twin for data synthesis & system validation
Predictive AI to understand what will happen
prediction
Actions depending on the predicted states of the system
System states prediction according to hypothetical inputs

Figure 6. A unified framework for systems-level management. This framework comprises descriptive, predictive, and prescriptive AI in a digital twin to solve complex problems in smart grid management.

1) Physical system – represents the actual environment that the AI algorithms are targeting (e.g., the smart grid). The physical system is usually a collection of states that change in response to certain variables (e.g., temperature, humidity, irradiance) and physical laws.
2) Digital twin – denotes the virtual representation of the physical system. The digital twin can be categorized into a physical rule-based twin and data-driven twin. Specifically, the physical rule-based twin is constructed using basic design data and physical laws, whereas the data-driven twin is trained using historical data obtained from the physical system.
3) AIOps engine – represents the AI agent that optimizes, diagnoses, or controls the physical system. This engine can be trained using data from the physical system or digital twin directly. To eliminate the risks of AI implementation on the physical system, preliminary AIOps engine training is implemented on the digital twin to reduce its uncertainties in decision-making.

The proposed framework collects various raw data (e.g., design, configuration, operational data) from the physical system. Descriptive AI is used to improve data quality and extract key features which are then transformed into a form acceptable by the digital twin for digital twin model construction. Subsequently, AIOps interacts with the digital twin to generate massive training data with high diversity for prescriptive AI training so that system behaviours can be learned and then optimized. Aside from training, the digital twin can also be used to validate the recommended actions of AIOps in advance, which greatly helps to alleviate the risk-averse mindset prevalent in deploying AI-based approaches. To further improve the decision-making process, predictive AI technology can be used to forecast future system states (e.g., lifetime, power generation, workload) based on specified inputs, and appropriate actions can be derived for the physical system based on the prediction results to improve system efficiency.

Such a framework can potentially tackle the data scarcity and risk-averse mindset challenges that impede AI technology deployment in real-world scenarios through a combination of data augmentation and risk-aware deployment. The digital twin provides sufficient data with high accuracy and diversity to enhance the performance of AI algorithms. For example, the synthesized data of a digital twin has been found to improve the accuracy of prognostics and health management by up to 30% compared with the method that only uses the physical data from the wind turbine, exemplifying the effectiveness of the digital twin on data augmentation.[121] Furthermore, the AI algorithm can be trained with the digital twin instead of the physical system, to ensure the safety of the training. Due to the bias of the data distribution and uncertainties of the AI algorithm, the deployment safety of AI cannot be fully guaranteed. To circumvent this, the high-fidelity digital twin can be used to validate the recommended policies from the AI algorithm in advance to verify the deployment safety. For example, the digital twin has been proven to be useful for early validation of performance degradation and control policies in safety-critical systems such as advanced driving assistant systems,[122] aircraft,[123] manufacturing,[124] and energy systems.[125] The combination of predictive AI techniques with digital twins thus empowers operators to predict anomalies and emergencies in systems taking a proactive approach towards system maintenance and diagnosis, while saving operating expenses.[126]

12. Data Center Systems in Smart Grids

In addition, data centers can be used as a potential "interruptible" load to stabilize the power grid with renewable sources. Traditionally, data centers consist of the information technology (IT) subsystem (e.g., servers, racks, switches) and the facility subsystem (e.g., cooling units). Data centers have recently received increased attention from the energy sector due to their high electricity consumption and carbon emission in operation. According to a report, data centers in the United States consumed 70 billion kWh of electricity in 2014, accounting for 1.8% of total electricity consumption[127] in the country, and this figure is expected to rise rapidly. As a result, the optimization of data centers is critical and urgent.

Many attempts have been made to improve data center energy efficiency. For example, some works have used siloed ML approaches to optimize job allocation[128] (for the IT subsystem) and cooling control (for the facility subsystem).[129] Such existing works are unsatisfactory because almost all of them are not adopted in physical systems due to the challenges of "data scarcity" and "risk-averse mindset" in deploying AI-based approaches. Only giant enterprises like Google which have their own data center and accepts a certain level of risk, have successfully deployed learning-based technologies onto the physical system.[130] Researchers have further looked into joint IT-facility optimization to improve data center energy efficiency.[131] In spite of this, they all assume static or dynamic models for the target system which are often insufficient to capture the internal dynamics. Consequently, these techniques have not been successfully applied to the physical system.

The fluctuation of energy output in fact provides opportunities for system operation and maintenance. Grid fluctuation, for example, can be used to absorb different types of loads, such as rigid or interruptible loads. Data centers can be leveraged as an "interruptible" load for demand response,[131] providing both economic incentives to data center operators as well as grid stability handles for power grid operators. Previously, a large battery array was introduced to mitigate the fluctuations of renewable energy sources to improve the stability of the energy system. Such an approach would, however, significantly increase the construction costs of the system. As the charge and discharge rate of the battery increases, so does its price. Operators must thus devote a large amount of manpower and material resources to the operation of batteries. The replacement of batteries also creates waste and pollutes the environment.

The high dynamics in data center electricity usage present the possibility of it being utilized as an interruptible load that can alternate itself as a regulator of the power supply of a large grid. A recent survey has revealed that the data center's load can be reduced by 5% within 5 minutes and 10% within 15 minutes.[132] Data centers thus present an extremely promising flexible load that can replace battery storage to improve power grid stability. Given the storage characterization, one will be able to calculate the value of data center demand response in terms of its "equivalent" storage capacity. The capacity equivalence in using battery systems or interruptible loads for demand response is called supply-demand parity, which, is an innovative approach. The envisioned paradigm (Figure 7) involves two stakeholders: data center operators and power grid operators. Data center operators may use learning-based approaches to measure and manage their power consumption to match variable yields generated by power grid operators. This new operational model has the potential to

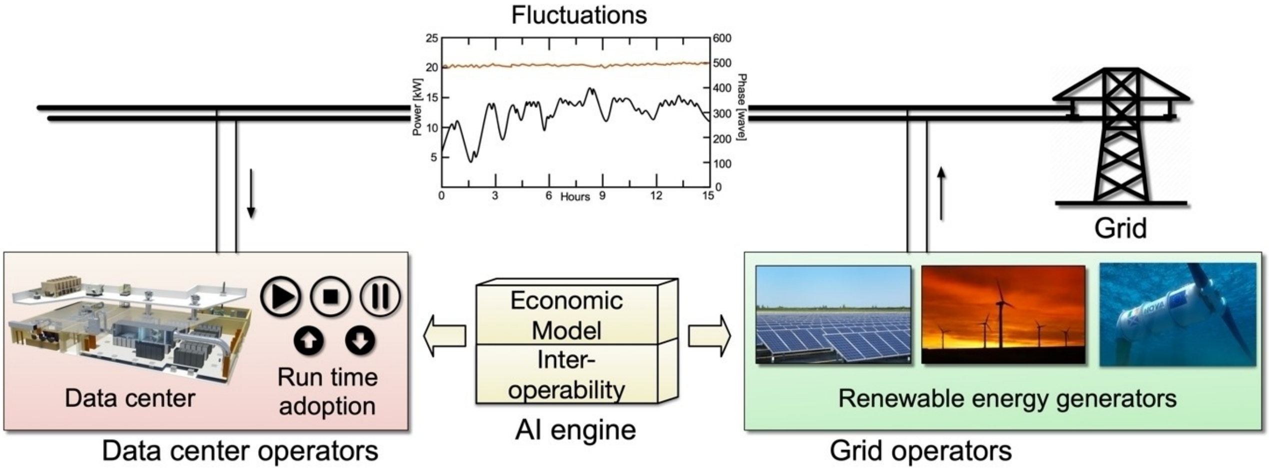

Figure 7. A conceptual paradigm for smart grid management. In this paradigm, a green data center can be leveraged as an "interruptible" load for grid stabilization.

transform data center operations from a cost center to a revenue generator – a possibility that will profoundly transform the data center industry.

13. Outlook

Although ML has shown tremendous potential in enhancing battery material research and smart grid management systems, it is undeniable that ML is still overall a relatively new method, at least in the scope of how long science and research has existed. Therefore, there are still 3 major contradictions brought up regarding applying machine learning reasonably and efficiently to the materials community. Namely, they are the contradiction between learning results and domain knowledge, the contradiction between model complexity and ease of use, and the contradiction between high dimension and small sample data.[133,134]

First and foremost, the contradiction between learning results and domain knowledge. Due to the primary purpose of ML being the search for things that we may have originally missed, it is not surprising that many of the ML models that exist today do not consider knowledge that is already known. While this worked well in the past, with the increase in the complexity of the subjects considered and thus the complexity of the data that is being analyzed, the computing power required to run the standard ML models is increasing exponentially over the years. Additionally, due to imperfect data, it can also lead to inaccurate results. We believe that one such way to address this is using Explainable AI, which we mentioned earlier in the review. Explainable AI aids in the understanding of how the ML algorithms make their decisions and as such, experts would be able to compare their own understanding against the decisions made by the algorithm to check if it is something that we have misunderstood all this time, or if it was merely an incorrect output by the algorithm due to imperfect data or due to imperfect algorithm design.

Secondly, the contradiction between model complexity and ease of use. Due to there being many different experimental setups in the world with different features, different outcomes and thus different relationships, generally speaking, from dataset to dataset, different models have to be used and the hyperparameters for the model have to be optimized and finetuned for every single dataset. However, due to some datasets having linearly related features and others having non-linearly related features and more complexly, having a mixture of linearly related features and non-linearly related features, more and more complex ML algorithms have to be created to handle the increasing complexity of the problems. However, this has resulted in some ML algorithms becoming "black boxes" where the decision-making process is no longer able to be understood by humans, not even by the designer themselves. Once again, this is something that Explainable AI was primarily designed to solve. Explainable AI is meant to fade these "black boxes" by helping the user to understand how the models are making their decisions. For example, the random forests method assigns a feature importance score to every feature present in the dataset and allows the user to easily see the ranking and the significance that the ML model places on each feature. This would allow for increased ease of use due to an increased understanding, even while the complexity of the model increases.

Lastly, the contradiction between high dimension and small sample data. This indeed is a significant issue when it comes to ML, because it is generally believed that the greater the number of features we begin with, the more bases we can cover. However, having too many features in itself can be detrimental as it can result in the increased complexity of the ML model and hence, result in an overfitted model. Therefore, the standard approaches to solving this issue are either dimensionality reduction (reducing the total number of features in use by choosing only the best features), sample augmentation (obtaining more data points via papers or data posted online, or via data that is generated through the use of generative models), active learning (only the most informative samples are used for the model training), and ensemble learning (combines multiple learning models together similar to the random tree method, where the individual mistakes of each learning model are covered by the rest of the learning models, due to the low likelihood of all the models sharing the same mistakes). In this manuscript, we addressed this issue using sample augmentation, where we did not just rely on the data that we have generated on our own but instead relied on data that has been published online as well. However, we are solely looking at experimental data and as such do not use generative models to obtain additional data. Instead, we are supporting the push for more open and effective data sharing that follows the 'FAIR' data principles. This requires data to be 'Findable' and easily 'Accessible' to all interested parties. The data should also be represented in established standards, making them 'Interoperable' and 'Reusable' by anyone who uses the data, even when outside of the initial intent of the data.

14. Conclusion

ML algorithms offer a promising pathway to accelerate the pace of battery materials research and provide much needed solutions for pressing climate concerns. Doing so requires frameworks to be developed that are highly transferable, enabling it to solve a multitude of problems. Preferably, these frameworks should also be designed to be less dependent on domain expertise and thus accessible to a larger pool of researchers. The application of ML in battery research is still rapidly developing, with many opportunities abound. We anticipate that, with the aid of XAI and transfer learning, these application agnostic concepts can be extended to a multitude of other important technologies for fully autonomous materials development in the future.

With regards to smart grid management, ML has the potential to revolutionize the grid, provided the existing challenges of data scarcity and risk-averse mindset are solved. The critical regulatory role that data centers, as major electricity consumers, can play in smart grid management, should not be neglected. We hope that the proposed unified framework that incorporates ML approaches and digital twins can motivate and promote the use of AI algorithms in real-world scenarios. Specifically, the insights regarding systems-level optimization of smart grids using ML can be applied to other systems to create interesting prospects as well.

Author contributions

All authors contributed to researching data and writing the manuscript.

Acknowledgements

This work is supported by the Singapore National Research Foundation (NRF-NRFF2017-04 and Energy Research Test-Bed and Industry Partnership Funding Initiative, part of the Energy Grid (EG) 2.0 programme), and Agency for Science, Technology and Research (Central Research Fund Award).

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Keywords: artificial intelligence · batteries · inverse design · machine learning · robotics

[1] S. Chu, Y. Cui, N. Liu, *Nat. Mater.* **2017**, *16*, 16–22.
[2] D. J. Davidson, *Nat. Energy* **2019**, *4*, 254–256.
[3] J. B. Goodenough, K.-S. Park, *J. Am. Chem. Soc.* **2013**, *135*, 1167–1176.
[4] M. S. Whittingham, *Chem. Rev.* **2014**, *114*, 11414–11443.
[5] A. Y. S. Eng, C. B. Soni, Y. Lum, E. Khoo, Z. Yao, S. K. Vineeth, V. Kumar, J. Lu, C. S. Johnson, C. Wolverton, Z. W. Seh, *Sci. Adv.* **2022**, *8*, eabm2422.
[6] S. Suh, J. D. Bergesen, T. Gibon, E. G. Hertwich, M. Taptich, *United Nations Environment Programme, International Resource Panel Report*, **2017**.
[7] M. Lu, C. Chang, W. Lee, L. Wang, *IEEE Trans. Ind. Appl.* **2009**, *45*, 2109–2115.
[8] A. Mahmood, N. Javaid, M. A. Khan, S. Razzaq, *Int. J. Energy Res.* **2015**, *39*, 1437–1450.
[9] S. Rajanna, R. P. Saini, *Renewable Sustainable Energy Rev.* **2016**, *58*, 376–396.
[10] T. Lombardo, M. Duquesnoy, H. El-Bouysidy, F. Årén, A. Gallo-Bueno, P. B. Jørgensen, A. Bhowmik, A. Demortière, E. Ayerbe, F. Alcaide, M. Reynaud, J. Carrasco, A. Grimaud, C. Zhang, T. Vegge, P. Johansson, A. A. Franco, *Chem. Rev.* **2022**, *122*, 10899–10969.
[11] C. Lv, X. Zhou, L. Zhong, C. Yan, M. Srinivasan, Z. W. Seh, C. Liu, H. Pan, S. Li, Y. Wen, Q. Yan, *Adv. Mater.* **2022**, *34*, 2101474.
[12] Z. Deng, V. Kumar, F. T. Bölle, F. Caro, A. A. Franco, I.E. Castelli, P. Canepa, Z. W. Seh, *Energy Environ. Sci.* **2022**, *15*, 579–594.
[13] S. N. Steinmann, Z. W. Seh, *Nat. Rev. Mater.* **2021**, *6*, 289–291.
[14] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. J. Martinez-de-Pison, F. Antonanzas-Torres, *Sol. Energy* **2016**, *136*, 78–111.
[15] P. Chemouil, P. Hui, W. Kellerer, N. Limam, R. Stadler, Y. Wen, *IEEE J. Sel. Areas Commun.* **2020**, *38*, 2229–2233.
[16] F. Cong, C. Mingjian, H. Bri-Mathias, Z. Jie, *Appl. Energy* **2017**, *190*, 1245–1257.
[17] V. Cyril, N. Gilles, K. Soteris, N. Marie-Laure, P. Christophe, M. Fabrice, F. Alexis, *Renewable Energy* **2017**, *105*, 569–582.
[18] W. Huaizhi, L. Zhenxing, Z. Xian, Z. Bin, P. Jianchun, *Energy Convers. Manage.* **2019**, *198*, 111799.
[19] 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.
[20] A. Van der Ven, Z. Deng, S. Banerjee, S. P. Ong, *Chem. Rev.* **2020**, *120*, 6977–7019.
[21] 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.
[22] M. I. Jordan, T. M. Mitchell, *Science* **2015**, *349*, 255–260.
[23] Y. LeCun, Y. Bengio, G. Hinton, *Nature* **2015**, *521*, 436–444.
[24] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, *Nature* **2018**, *559*, 547–555.
[25] P. Luna, J. Wei, Y. Bengio, A. Aspuru-Guzik, E. Sargent, *Nature* **2017**, *552*, 23–27.
[26] Q. Zhao, L. Zhang, B. He, A. Ye, M. Avdeev, L. Chen, S. Shi, *Energy Storage Mater.* **2021**, *40*, 386–393.
[27] B. Sanchez-Lengeling, A. Aspuru-Guzik, *Science* **2018**, *361*, 360–365.
[28] A. Zunger, *Nat. Chem. Rev.* **2018**, *2*, 0121.
[29] X. Feng, Q. Zhang, Z. W. Seh, *Adv. Mater. Technol.* **2022**, *7*, 2200616.
[30] A. Chen, X. Zhang, Z. Zhou, *InfoMat* **2020**, *2*, 553–576.
[31] R. P. Joshi, J. Eickholt, L. Li, M. Fornari, V. Barone, J. E. Peralta, *ACS Appl. Mater. Interfaces* **2019**, *11*, 18494–18503.
[32] O. Allam, B. W. Cho, K. C. Kim, S. S. Jang, *RSC Adv.* **2018**, *8*, 39414–39420.
[33] M. Kim, S. Kang, H. G. Park, K. Park, K. Min, *Social Science Research Network* **2022**, *10.2139/ssrn.4117000*.
[34] E. D. Cubuk, A. D. Sendek, E. J. Reed, *J. Chem. Phys.* **2019**, *150*, 214701.
[35] A. D. Sendek, E. D. Cubuk, E. R. Antoniuk, G. Cheon, Y. Cui, E. J. Reed, *Chem. Mater.* **2019**, *31*, 342–352.
[36] 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.
[37] Y. Takagishi, T. Yamanaka, T. Yamaue, *Batteries* **2019**, *5*, 54.
[38] Z. Ahmad, T. Xie, C. Maheshwari, J. C. Grossman, V. Viswanathan, *ACS Cent. Sci.* **2018**, *4*, 996–1006.
[39] Y. Zhang, X. Xu, *Ind. Eng. Chem. Res.* **2021**, *60*, 343–354.
[40] J. Noh, J. Kim, H. S. Stein, B. Sanchez-Lengeling, J. M. Gregoire, A. Aspuru-Guzik, Y. Jung, *Matter* **2019**, *1*, 1370–1384.
[41] 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*, 76–86.
[42] S. Shi, J. Gao, Y. Liu, Y. Zhao, Q. Wu, W. Ju, C. Ouyang, R. Xiao, *Chin. Phys. B* **2016**, *25*, 018212.
[43] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, *Inf. Fusion* **2020**, *58*, 82–115.
[44] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, G.-Z. Yang, *Sci. Robot.* **2019**, *4*, eaay7120.
[45] A. Wang, Z. Zou, D. Wang, Y. Liu, Y. Li, J. Wu, M. Avdeev, S. Shi, *Energy Storage Mater.* **2021**, *35*, 595–601.
[46] Q. Zhao, M. Avdeev, L. Chen, S. Shi, *Sci. Bull.* **2021**, *66*, 1401–1408.
[47] L. Breiman, *Mach. Learn.* **2001**, *45*, 5–32.
[48] S. Stijven, W. Minnebo, K. Vladislavleva, in *Proceedings of the 13th annual conference companion on Genetic and evolutionary computation*, Association for Computing Machinery, Dublin, Ireland, **2011**, pp. 623–630.
[49] B. H. Menze, B. M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, F. A. Hamprecht, *BMC Bioinf.* **2009**, *10*, 213.
[50] M. T. Ribeiro, S. Singh, C. Guestrin, *ArXiv* **2016**, *abs/1602.04938*.
[51] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, *PLoS One* **2015**, *10*, e0130140.
[52] S. M. Lundberg, S.-I. Lee, *ArXiv* **2016**, *abs/1611.07478*.
[53] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, K.-R. Müller, *Pattern Recognit. Lett.* **2017**, *65*, 211–222.
[54] M.-F. Ng, J. Zhao, Q. Yan, G. J. Conduit, Z. W. Seh, *Nat. Mach. Intell.* **2020**, *2*, 161–170.

[55] 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*, 383–391.

[56] 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*, 397–402.

[57] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, A. Jain, *Nature* **2019**, *571*, 95–98.

[58] 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.

[59] J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway, A. Aspuru-Guzik, *J. Phys. Chem. Lett.* **2011**, *2*, 2241–2251.

[60] C. Fan, M. Chen, X. Wang, J. Wang, B. Huang, *Front. Energy Res.* **2021**, *9*.

[61] C. Fan, F. Xiao, H. Madsen, D. Wang, *Energy Build.* **2015**, *109*, 75–89.

[62] F. Xiao, C. Fan, *Energy Build.* **2014**, *75*, 109–118.

[63] C. Fan, F. Xiao, C. Yan, *Autom. in Constr.* **2015**, *50*, 81–90.

[64] M. M. Jenghara, H. Ebrahimpour-Komleh, V. Rezaie, S. Nejatian, H. Parvin, S. K. Yusof, *Knowl. Inf. Syst.* **2018**, *56*, 123–139.

[65] P. Kang, *Neurocomputing* **2013**, *118*, 65–78.

[66] X. Li, C. P. Bowers, T. Schnier, *IEEE Trans. Ind. Electron.* **2010**, *57*, 3639–3644.

[67] J. Liu, J. Liu, H. Chen, Y. Yuan, Z. Li, R. Huang, *Energy Build.* **2018**, *175*, 148–162.

[68] Y. Liu, J.-M. Wu, M. Avdeev, S.-Q. Shi, *Adv. Theory Simul.* **2020**, *3*, 1900215.

[69] X. Yu, S. Ergan, G. Dedemen, *Appl. Energy* **2019**, *253*, 113497.

[70] M. Ashouri, B. C. M. Fung, F. Haghighat, H. Yoshino, *Energy* **2020**, *194*, 116813.

[71] M. Le Cam, A. Daoud, R. Zmeureanu, *Energy* **2016**, *101*, 541–557.

[72] 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*, 73–76.

[73] C. Finn, T. Yu, T. Zhang, P. Abbeel, S. Levine, *ArXiv* **2017**, *abs/1709.04905*.

[74] Y. Duan, M. Andrychowicz, B. Stadie, J. Ho, J. Schneider, I. Sutskever, P. Abbeel, W. Zaremba, *Arxiv* **2017**, *abs/1703.07326*.

[75] T. Yu, C. Finn, A. Xie, S. Dasari, T. Zhang, P. Abbeel, S. Levine, *ArXiv* **2018**, *abs/1802.01557*.

[76] L. Bassman, P. Rajak, R. K. Kalia, A. Nakano, F. Sha, J. Sun, D. J. Singh, M. Aykol, P. Huck, K. Persson, P. Vashishta, *Npj Comput. Mater.* **2018**, *4*, 74.

[77] T. Lookman, P. V. Balachandran, D. Xue, R. Yuan, *Npj Comput. Mater.* **2019**, *5*, 21.

[78] D. Jha, K. Choudhary, F. Tavazza, W.-k. Liao, A. Choudhary, C. Campbell, A. Agrawal, *Nat. Commun.* **2019**, *10*, 5316.

[79] H. Yamada, C. Liu, S. Wu, Y. Koyama, S. Ju, J. Shiomi, J. Morikawa, R. Yoshida, *ACS Cent. Sci.* **2019**, *5*, 1717–1730.

[80] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, *Sci. Data* **2016**, *3*, 160018.

[81] P. Friederich, F. Häse, J. Proppe, A. Aspuru-Guzik, *Nat. Mater.* **2021**, *20*, 750–761.

[82] T. Bereau, R. A. DiStasio, A. Tkatchenko, O. A. von Lilienfeld, *J. Chem. Phys.* **2018**, *148*, 241706.

[83] A. Hajibabaei, K. S. Kim, *J. Phys. Chem. Lett.* **2021**, *12*, 8115–8120.

[84] Z. Liu, W. Fu, E. A. Payzant, X. Yu, Z. Wu, N. J. Dudney, J. Kiggans, K. Hong, A. J. Rondinone, C. Liang, *J. Am. Chem. Soc.* **2013**, *135*, 975–978.

[85] Y. Seino, T. Ota, K. Takada, A. Hayashi, M. Tatsumisago, *Energy Environ. Sci.* **2014**, *7*, 627–631.

[86] F. Häse, L. M. Roch, A. Aspuru-Guzik, *Chem.* **2019**, *1*, 282–291.

[87] D. P. Tabor, L. M. Roch, S. K. Saikin, C. Kreisbeck, D. Sheberla, J. H. Montoya, S. Dwaraknath, M. Aykol, C. Ortiz, H. Tribukait, C. Amador-Bedolla, C. J. Brabec, B. Maruyama, K. A. Persson, A. Aspuru-Guzik, *Nat. Rev. Mater.* **2018**, *3*, 5–20.

[88] S. Maruyama, O. Kubokawa, K. Nanbu, K. Fujimoto, Y. Matsumoto, *ACS Comb. Sci.* **2016**, *18*, 343–348.

[89] M. Roberts, J. Owen, *ACS Comb. Sci.* **2011**, *13*, 126–134.

[90] S. Vogt, Y. S. Chu, A. Tkachuk, P. Ilinski, D. A. Walko, F. Tsui, *Appl. Surf. Sci.* **2004**, *223*, 214–219.

[91] J. Sottmann, R. Homs-Regojo, D. S. Wragg, H. Fjellvag, S. Margadonna, H. Emerich, *J. Appl. Crystallogr.* **2016**, *49*, 1972–1981.

[92] A. Kwade, W. Haselrieder, R. Leithoff, A. Modlinger, F. Dietrich, K. Droeder, *Nat. Energy* **2018**, *3*, 290–300.

[93] W. Sun, T. Dacek Stephen, P. Ong Shyue, G. Hautier, A. Jain, D. Richards William, C. Gamst Anthony, A. Persson Kristin, G. Ceder, *Sci. Adv. 2*, e1600225.

[94] J. M. Granda, L. Donina, V. Dragone, D.-L. Long, L. Cronin, *Nature* **2018**, *559*, 377–381.

[95] M. Roch Loïc, F. Häse, C. Kreisbeck, T. Tamayo-Mendoza, P. E. Yunker Lars, E. Hein Jason, A. Aspuru-Guzik, *Sci. Robot.* **2018**, *3*, eaat5559.

[96] P. Ravestein, G. van der Schrier, R. Haarsma, R. Scheele, M. van den Broek, *Renewable Sustainable Energy Rev.* **2018**, *97*, 497–508.

[97] R. Ehsan, S. Saeed, R. R. Leon, M. Marc, *Int. J. Electr. Power Energy Syst.* **2016**, *77*, 337–344.

[98] H. Omar, B. Kankar, *Renewable Energy* **2012**, *45*, 7–15.

[99] A. Khosravi, R. N. N. Koury, L. Machado, J. J. G. Pabon, *J. Cleaner Prod.* **2018**, *176*, 63–75.

[100] A. Gensler, J. Henze, B. Sick, N. Raabe, *Deep Learning for solar power forecasting – An approach using AutoEncoder and LSTM Neural Networks*, from 2016 IEEE Int. Conf. Syst. Man Cybern. SMC 2016, 002858–002865.

[101] L. Yin, T. Yu, X. Zhang, B. Yang, *Energy* **2018**, *149*, 11–23.

[102] T. Yu, H. Z. Wang, B. Zhou, K. W. Chan, J. Tang, *IEEE Trans. Power Syst.* **2015**, *30*, 1669–1679.

[103] N. Wu, H. Wang, *J. Cleaner Prod.* **2018**, *204*, 1169–1177.

[104] X. Wang, A. Palazoglu, N. H. El-Farra, *Appl. Energy* **2015**, *143*, 324–335.

[105] D. L. Marino, K. Amarasinghe, M. Manic, *Building energy load forecasting using Deep Neural Networks*, from Proc.: IECON 2016–42nd Annu. Conf. IEEE Ind. Electron. Soc., 7046–7051.

[106] R. Seunghyoung, N. Jaekoo, K. Hongseok, *Deep neural network based demand side short term load forecasting*, from 2016 Int. Conf. Smart Grid Clean Energy Technol. ICSGCE 2016, 308–313.

[107] E. Mocanu, P. H. Nguyen, W. L. Kling, M. Gibescu, *Energy Build.* **2016**, *116*, 646–655.

[108] P. D. Lund, J. Lindgren, J. Mikkola, J. Salpakari, *Renewable Sustainable Energy Rev.* **2015**, *45*, 785–807.

[109] B. G. Kim, Y. Zhang, M. v. d. Schaar, J. W. Lee, *IEEE Trans. Smart Grid* **2016**, *7*, 2187–2198.

[110] I. Dusparic, A. Taylor, A. Marinescu, V. Cahill, S. Clarke, *Maximizing renewable energy use with decentralized residential demand response*, from 2015 IEEE Int. Smart Cities Conf. ISC2 2015, 1–6.

[111] I. Dusparic, C. Harris, A. Marinescu, V. Cahill, S. Clarke, *Multi-agent residential demand response based on load forecasting*, from 2013 IEEE Conf. Technol. Sustain. SusTech 2013, 90–96.

[112] L. Hardesty, *Siting wind farms more quickly, cheaply*, **2015**, MIT News Office.

[113] L. Mearian, *IBM's machine-learning crystal ball can foresee renewable energy availability*, **2015**, IDG Communications, Inc.

[114] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, Y. Zhang, *IEEE Trans. Smart Grid* **2019**, *10*, 841–851.

[115] T. Wuest, D. Weimer, C. Irgens, K.-D. Thoben, *Prod. Manuf. Res.* **2016**, *4*, 23–45.

[116] A. M. Omer, *Renewable Sustainable Energy Rev.* **2007**, *11*, 1481–1497.

[117] P. de Jong, A. S. Sánchez, K. Esquerre, R. A. Kalid, E. A. Torres, *Renewable Sustainable Energy Rev.* **2013**, *23*, 526–535.

[118] Z. Hameed, S. H. Ahn, Y. M. Cho, *Renewable Energy* **2010**, *35*, 879–894.

[119] S. Zolfani, J. Šaparauskas, *Eng. Econ.* **2013**, *24*, 408–414.

[120] K. C. Toh, K. J. Tseng, S. K. Panda, S. E. Lee, *Green Data Centre Technology Primer: A Summary*, **2014**, National Climate Change Secretariat and National Research Foundation.

[121] F. Tao, M. Zhang, Y. Liu, A. Y. C. Nee, *CIRP Ann.* **2018**, *67*, 169–172.

[122] S. Yun, J. Park, W. Kim, *Data-centric middleware based digital twin platform for dependable cyber-physical systems*, from 2017 Ninth Int. Conf. Ubiquitous Future Netw. ICUFN 2017, 922–926.

[123] E. Glaessgen, D. Stargel, *The digital twin paradigm for future NASA, U. S. air force vehicles*, **2012**.

[124] R. Rosen, G. von Wichert, G. Lo, K. D. Bettenhausen, *IFAC-PapersOnLine* **2015**, *48*, 567–572.

[125] H. Bounechba, A. Bouzid, H. Snani, A. Lashab, *Int. J. Electr. Power Energy Syst.* **2016**, *83*, 67–78.
[126] C. Zhuang, J. Liu, H. Xiong, *Int. J. Adv. Manuf.* **2018**, *96*.
[127] E. O'Shaughnessy, J. Heeter, J. Sauer, *Status and Trends in the U. S. Voluntary Green Power Market (2017 Data)*, **2018**, National Renewable Energy Laboratory.
[128] Z. Liu, I. Liu, S. Low, A. Wierman, *Sigmetrics '14* **2014**, 111–123.
[129] N. Lazic, T. Lu, C. Boutilier, M. Ryu, E. Wong, B. Roy, G. Imwalle, in *2018 32nd Int. Conf. Neur. Inf. Processing Syst. NeurIPS 2018*, Curran Associates Inc., Montréal, Canada, **2018**, pp. 3818–3827.
[130] J. Wan, X. Gui, R. Zhang, L. Fu, *IEEE Syst. J.* **2018**, *12*, 2461–2472.
[131] W. Xia, Y. Wen, K.-C. Toh, Y. Wong, *IEEE Commun. Mag.* **2015**, *53*, 192–198.
[132] N. Liu, Z. Li, Z. Xu, J. Xu, S. Lin, Q. Qiu, J. Tang, Y. Wang, *ArXiv* **2017**, *abs/1703.04221*.
[133] Y. Liu, B. Guo, X. Zou, Y. Li, S. Shi, *Energy Storage Mater.* **2020**, *31*, 434–450.
[134] Y. Liu, T. Zhao, W. Ju, S. Shi, *J. Materiomics* **2017**, *3*, 159–177..

Manuscript received: July 4, 2022
Revised manuscript received: August 11, 2022
Accepted manuscript online: August 29, 2022
Version of record online: September 15, 2022