J. Chem. Sci. (2022)134:2
https://doi.org/10.1007/s12039-021-01995-2

PERSPECTIVE ARTICLE

Artificial intelligence: machine learning for chemical sciences

AKSHAYA KARTHIKEYAN and U DEVA PRIYAKUMAR*
Center for Computational Natural Sciences and Bioinformatics, International Institute of Information Technology, Hyderabad 500 032, India
E-mail: deva@iiit.ac.in

MS received 9 July 2021; revised 8 September 2021; accepted 14 September 2021, corrected publication 2022

Abstract. Research in molecular sciences witnessed the rise and fall of Artificial Intelligence (AI)/ Machine Learning (ML) methods, especially artificial neural networks, few decades ago. However, we see a major resurgence in the use of modern ML methods in scientific research during the last few years. These methods have had phenomenal success in the areas of computer vision, speech recognition, natural language processing (NLP), etc. This has inspired chemists and biologists to apply these algorithms to problems in natural sciences. Availability of high performance Graphics Processing Unit (GPU) accelerators, large datasets, new algorithms, and libraries has enabled this surge. ML algorithms have successfully been applied to various domains in molecular sciences by providing much faster and sometimes more accurate solutions compared to traditional methods like Quantum Mechanical (QM) calculations, Density Functional Theory (DFT) or Molecular Mechanics (MM) based methods, etc. Some of the areas where the potential of ML methods are shown to be effective are in drug design, prediction of high–level quantum mechanical energies, molecular design, molecular dynamics materials, and retrosynthesis of organic compounds, etc. This article intends to conceptually introduce various modern ML methods and their relevance and applications in computational natural sciences.

Keywords. Deep learning; machine learning; computational chemistry; drug design; molecular design; computational materials; neural networks.

1. Introduction

The application of ML methods to problems in natural sciences started few decades ago. The first publication in this area was by Hiller *et al.* in 1973, which used a three-layer perceptron network for the classification of substituted 1,3-dioxanes as pharmacologically active and inactive.[1] From the 1990s, the use of artificial neural networks (ANNs) was prevalent in computer aided drug design, especially in quantitative structure-activity relationship (QSAR) studies.[2] However, application of ML methods to other areas of scientific research remained a niche domain without much attention until recently.[3] Experiment, theory and computation are recognized as the three cornerstones on which scientific advances are made. The advent of new deep learning (DL) algorithms, along with new datasets, libraries, and better computing infrastructure, has fueled data–driven methods as the fourth paradigm. Figure 1 shows the number of publications with "machine learning" in the abstract according to American Chemical Society (ACS) Journals through the years. It shows ML has grown at a remarkable rate in the past four years as one of the most popular research directions.

An extreme view on AI/ML is that it "has made huge progress in perception". The immense hype around it has attracted the attention of people from all walks of science and technology. Below is a recent example of how modern ML methods have made a high impact on one of the holy grails of biological research - protein structure modeling from its primary sequence.

Critical assessment of protein structure prediction (CASP) is a competition that is conducted once in two years since 1994, where research teams from around the world attempt to predict three-dimensional structures of proteins from just the amino acid sequences. Proteins, whose structures are almost solved, or whose structures have been recently solved but are withheld from the public, are taken up in these competitions. The most recent and the 14th edition of this occurred in November 2020.[4] By comparing the computational predictions with the lab results, each CASP14

*For correspondence

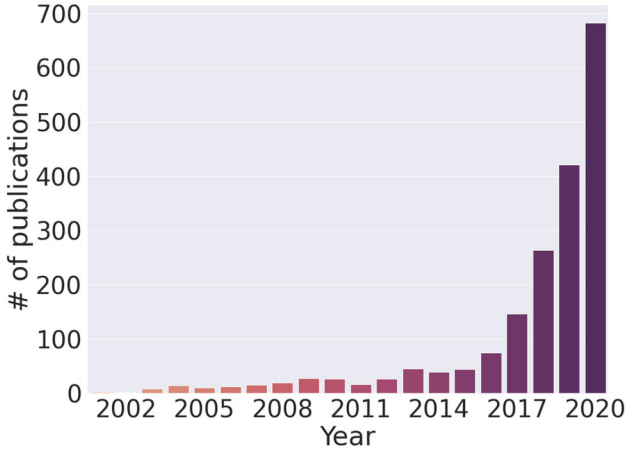

Figure 1. The rise of machine learning over the years evident from the number of publications American Chemical Society journals with "machine learning" anywhere in the article.

competitor received a global distance test (GDT) score. GDT is a structure similarity measure for comparing protein folds. One of the competitors - a company called DeepMind, outperformed others by a huge margin. DeepMind's AlphaFold 2 produced models for about two-thirds of the CASP14 target proteins with GDT scores above 90, indicating that the models are considered roughly equivalent to experimental methods. AlphaFold 2 is so highly accurate that many have hailed it as the solution to the long-standing protein structure prediction problem.[4–6] Such a huge difference between the performances of Deepmind and others was primarily due to the engineering aspects of the ML algorithms used.[7] This is one of the many successes of the modern ML methods and is just one example of how these algorithms along with physics based methods may impact the nature of scientific computing in the years to come.

The rest of the article is structured in the following manner. Initially, a short overview of different types of molecular representations and datasets is presented. Then, selected ML methods are discussed at the conceptual level. This is followed by brief discussions of a few popular areas of molecular sciences where ML has found success. Finally, the challenges faced by ML in molecular sciences are analyzed and there is also a discussion on how this area may evolve in general.

2. The role of ML in AI

The definitions of AI, ML, and DL have changed over the years, and their correlations have also evolved. Conventionally, AI is a general area which can loosely be termed as a class of techniques that enable computers to mimic human intelligence. Recently, AI systems have performed as well as, or even better than humans in several tasks.[8] AI, and its most common subfield of ML, study the methods of enabling machines to skillfully perform intelligent tasks without explicitly being programmed for those tasks. Today, in its various forms, AI is successfully applied across various domains ranging from robotics and image analysis to its application in molecular sciences.

Most researchers today agree that one of the primary requirements for intelligent behavior is learning. This makes ML one of the most rapidly developing subfields of AI. Nowadays, it is being argued that ML has outgrown its parent. DL and Reinforcement Learning (RL) are subcategories of ML that have recently developed in the field. Figure 2 shows the schematic of the conventional relationship between the categories.

2.1 *Machine learning*

Within AI, ML has emerged as the method of choice for developing practical software for machine translation, speech recognition, computer vision, recommendation systems and other applications.[9,10] ML, which includes DL, relies on statistical methods to learn from data. Using these techniques, we can extract complex and often hidden patterns from given data sets and can express them as mathematical

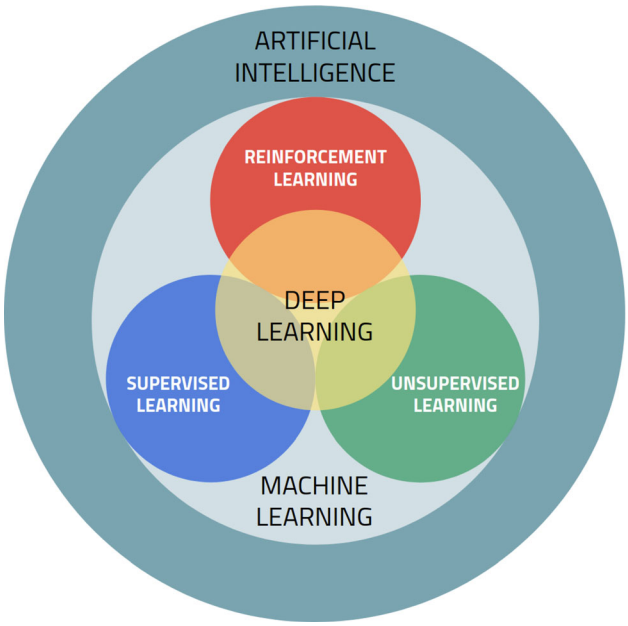

Figure 2. Schematic of the conventional relationship between artificial intelligence (AI), machine learning, deep learning and reinforcement learning.

objects. Many of the AI system developers now agree that, for many tasks, it can be far simpler to train a system by showing it examples of desired input-output behavior than to program it manually.

2.2 *Deep learning*

Traditional ML is limited to the size of its input data. For example, thousands of pixels will be sent to the system for analyzing images of conventional size. This means that reception and grouping of information to select those which are essential to the task will be necessary. DL is capable of handling such problems. It uses multi-layered neural networks, extremely large amounts of data and computing time to make accurate predictions. Unlike ML, it is not necessary to hand-engineer features (discussed later) from the raw data in DL. Function specification (defining what to learn from the given data) and optimization (how to weigh the data appropriately) are taken care of by the algorithm itself that has made DL extremely popular in many fields such as speech recognition,[11] computer vision,[12] NLP,[13] and recently in molecular sciences.

3. Chemical representations and descriptors

3.1 *Chemical representations*

Traditionally, molecules are depicted as structure diagrams with bonds and atoms. However, other representations are required for the computational processing of chemical structures. Chemical representation of a molecule may contain its spatial or topological information in a computer-interpretable format.[14,25] Current representations can be broadly classified into three types: discrete (e.g., text), continuous (e.g., vectors and tensors) and weighted graphs. Atomic coordinates, graph representations, simplified molecular-input line-entry system (SMILES) and international chemical identifier (InChl) are some of the popular representation methods.

A molecular graph representation essentially maps the atoms and bonds in a molecule to sets of nodes and edges respectively. It's formally a 2D matrix that can be used to represent 3D information like atomic coordinates and bond angles. A simple example is representing molecules in the form of an adjacency matrix A, where $a_{ij} = 1$ means there exists a bond between nodes v_i and v_j in the molecular graph, and $a_{ij} = 0$ means otherwise. However, the matrices by which molecules are described are not compact as they scale as the square of the number of atoms. This is not a problem with linear notations like SMILES and InChI.

SMILES is used to translate a chemical's 3D structure into a string of symbols based on a set of rules. It's like a connection table (Ctab) which identifies the nodes and edges of a molecular graph. Another form of line notation, InChI, is a hierarchical layered notation where each new layer describes more complex chemical characteristics. The first few layers include information within the connection table, and the additional layers (if needed) deal with complexities like isomers and isotopic distributions. The InChI provides a unique identifier, while SMILES is commonly used for storage and interchange of chemical structures.

3.2 *Molecular descriptors and fingerprints*

Using algorithms, the physical and chemical information encoded within the symbolic representations of molecules are transformed into useful mathematical representations, known as molecular descriptors or feature vectors.[15,16] Efforts have been made to define the criteria for developing efficient descriptors: they need to be interpretable, invariant to the symmetries of the underlying physics, direct and concise to avoid redundancy and the curse of dimensionality. Molecular descriptors can be experimental values like density, logP, dipolemoment and so on. They are used for various tasks like finding quantitative structure-property relationships (QSPRs) and QSARs, virtual screening (VS), and similarity searching. This is because molecules with similar properties tend to have similar descriptors.[15,17,18]

Molecular descriptors can have a significant impact on the performance of ML models based on how they capture the relevant features for the specific task. In 2013, Hansen *et al.*[19] improved their method of predicting atomization energies of organic molecules largely by modifying the representation used. By using variations of the Coulomb matrix (the representation used for the previous state-of-the-art model), they were the first to highlight the importance of good data representation in QM tasks.

Molecular descriptors are commonly categorized as 0D (0-Dimensional), 1D, 2D, 3D and 4D descriptors (Figure 3).[17] The 0D descriptors contain no information about the molecular structure, like atom and bond counts. 1D descriptors contain information obtained from the molecular formula, like molecular fingerprints. Molecular fingerprints encode the structural features of molecules in a binary bit string format. Circular

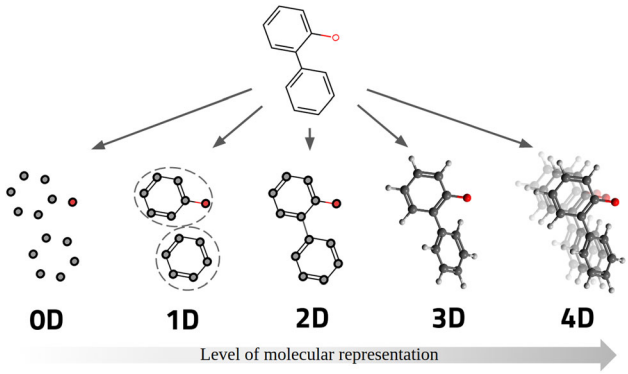

Figure 3. The common classification method of molecular descriptors.

fingerprints, based on the Morgan algorithm,[20] encode which substructures are present in a molecule.[21,22] One of the most common circular molecular fingerprints, extended-connectivity fingerprints (ECFPs),[23] are often used in QSAR models for lead optimization. A new molecular fingerprint called MinHashed atom-pair fingerprint, up to a diameter of four bonds (MAP4), is suitable for small to large molecules and can be adopted as a universal fingerprint.[24]

2D descriptors contain information concerning the size, configuration, and/or electronic distribution of molecules. These include variants of molecular graph representation[25] and CM. CM is a square matrix (atom by atom) that encodes the atomic nuclear charges (Z) and cartesian coordinates (R) of the atoms:

$$CM_{i,i} = 0.5Z_i^{2.4} \tag{1}$$

$$CM_{i,j} = \frac{(Z_i Z_j)}{|R_i - R_j|}, \quad i \neq j \tag{2}$$

where Z_i is the nuclear charge, and R_i is the nuclear radius of $atom_i$. Equation (1) corresponds to the approximate electronic potential energies of a free atom and Eqn. (2) corresponds to the coulomb nuclear repulsion terms. 3D descriptors usually depend on the 3D conformation of the molecule, like van der Waals volume and WHIM descriptors.[26] 4D descriptors are usually obtained through reference grids and molecular dynamics (MD) simulations.

Other examples of molecular featurization include Bag of Bonds (BoB)[27] and BAND[28] descriptor. BoB[27] can be seen as a histogram vector where each unit, called a "bag" counts the number of times a particular bond (such as C-O, C-H, etc.) appears. Like CM, a bag contains internuclear Coulomb repulsion between the atoms involved. In 2019, Laghuvarapu, Pathak, and Priyakumar[28] proposed BAND neural network for predicting atomization energies based on a chemically intuitive representation that captures the essence of molecular mechanics (MM) force fields. The BAND descriptor is computed as the sum of energy contributions from bonds (B), angles (A), nonbonds (N), and dihedrals (D).

4. Molecular datasets

The performance of ML models heavily depends on the increasing availability and quality of data. One of the challenges of using ML is getting the right data in the appropriate format. Getting the right data involves gathering information, which contains signals that correlate with the outcomes of the task. For example, information on NMR spectrum of molecules won't help in solvation energy prediction. High-quality datasets are usually difficult and expensive to create, and supervised learning (discussed later) also requires a significant amount of time to label the data.

The first ML algorithms for molecular modeling in 2010–2012 relied on small datasets having quantum mechanical (QM) properties for 10^2–10^3 molecular systems.[29–31] The chemical compound space (CCS) is estimated to consist an order of 10^{60}-10^{100} molecular systems.[32,33] In the last decade, increasingly larger chemical spaces were built and explored. Large scale QM and MD methods, along with advances in high throughput experiments, are generating data at an incredible rate. Today, DL models are capable of predicting chemical properties with reasonable accuracy by analyzing under just 5% of large molecular datasets. Such data efficiency and quality are crucial for in-silico chemical discovery.

Most studies applying ML for predicting QM properties, like atomization energy, use either QM7(b) dataset or its larger version QM9.[34,35] Both are subsets of the combinatorially generated molecular library GDB, which include over 10^9 stable organic compounds and up to 17 heavy atoms,[36] which essentially covers all small drug-like molecules. Other datasets are used in various ML problems such as predicting drug-target affinity (like Kiba[37] and Davis[38]), solvation energy (like FreeSolv[39] and MNSol[40]), spectrum prediction (like NMRShiftDB[41]), molecule generation (like MOSES[42]) and for many other tasks in molecular sciences. Datasets such as ZINC and ChEMBL include over 10^8 drug-like molecules for studying problems like ligand discovery. PubChem, a database of over 10^8 chemical substances and their activities,[43,44] is used in the fields of, among others, VS, drug repurposing, drug side effect prediction, chemical toxicity prediction and metabolite identification.

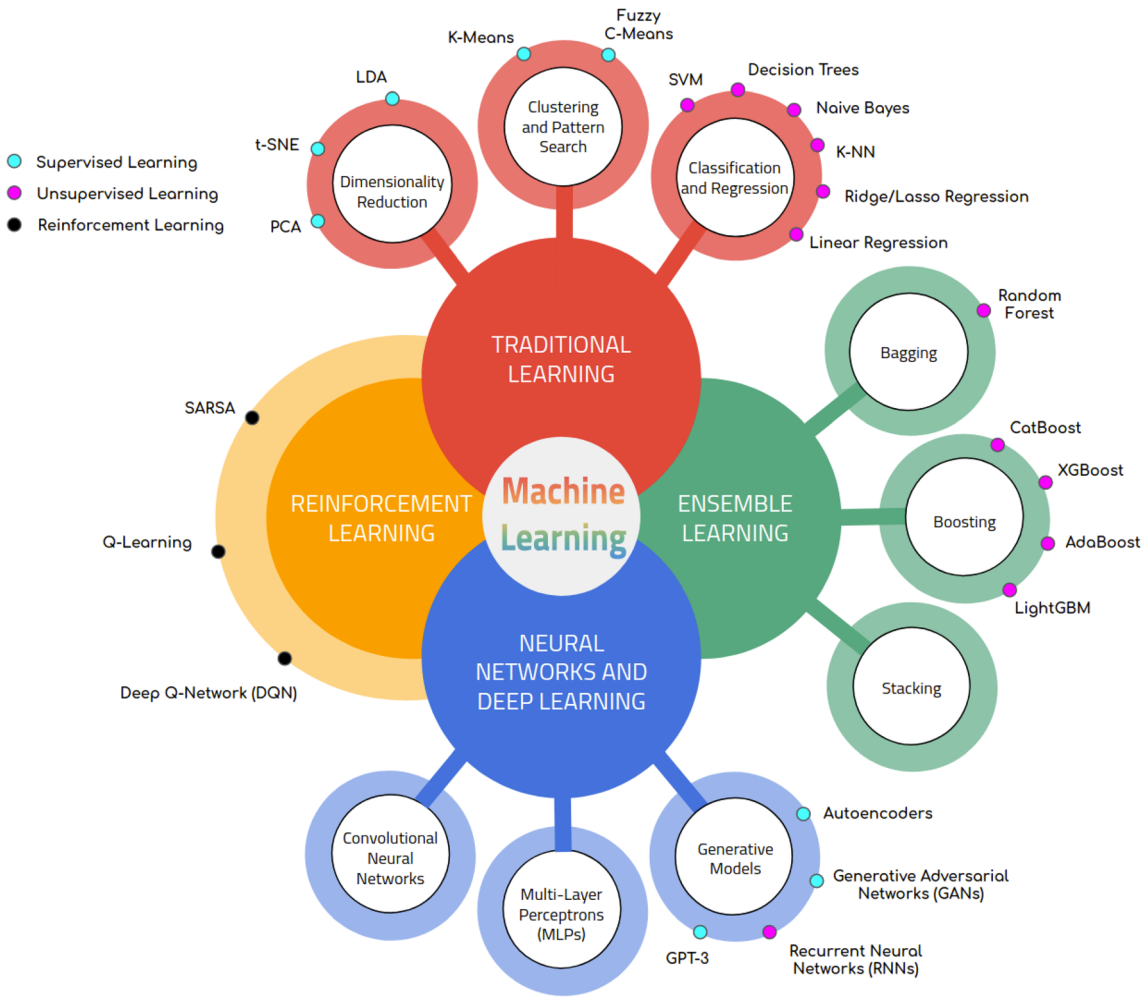

Figure 4. Examples of various machine learning approaches and algorithms.

5. ML approaches in molecular sciences

ML algorithms have successfully been applied to various domains in molecular sciences to obtain faster and more accurate solutions when compared to traditional methods (like QM calculations, DFT or MM-based methods, etc.). The relationship between a molecular structure and its properties is largely deterministic.[45] ML models take advantage of this through their flexibility (e.g. universal approximation theorem for ANNs) and learn the underlying QSPRs of a problem, even from simple chemical representations.[46]

ML approaches can be classified based on various standards. One method of classification is based on whether the ML system needs human supervision. Based on this, ML approaches are broadly categorized into three types: supervised, unsupervised, and reinforcement learning (Figure 4). This section presents a brief account of selected popular ML methods that have been used to tackle molecular science problems.

5.1 *Supervised learning*

The most widely used ML methods are supervised.[47] Molecular property predictions usually fall into this category. Supervised learning is the process of learning a function that maps an input to an output based on input–output pairs labelled by humans. The algorithms aim to minimize the errors pointed out during the learning process. It can extract complex nonlinear patterns and is superior to manually programmed traditional models. The most basic algorithm is linear regression, which is expressed as

$$\hat{y} = h_\theta(x) = \theta^T.x \tag{3}$$

where x is the feature vector, h_θ is the hypothesis function (mathematical formula to model a problem), and θ^T is the model's parameter vector with a bias term. The following sections briefly present examples of supervised algorithms applied to various molecular science tasks.

5.1a *Traditional ML methods*: Traditional ML methods can loosely be said to encompass fundamental algorithms that are often the foundation for more cutting–edge ML. Traditional algorithms are of several types: kernel based methods (like SVMs), decision tree methods (like Random Forests and XGBoost), Bayesian methods, etc. These algorithms can be used to solve classification and regression problems. For example, molecular property prediction is a regression problem where algorithms such as Kernel Ridge Regression (KRR),[27,48,49] Random Forests,[50,51] and Elastic Net[52] have been employed.

Although they have been successfully applied in various fields, traditional models rely on hand-engineered molecular descriptors from the symbolic representation of molecules, which requires domain expertise. Some ML approaches utilize experimental measurements such as physico-chemical properties as descriptors, but the cost of obtaining such optimized descriptors is the bottleneck. Deep neural networks (DNNs) are capable of automatic feature extraction and greatly outperforms traditional methods when it come to dealing with large datasets of complex problems.

However, traditional ML methods are still preferred over DNNs if the dataset size is small, as DNNs tend to overfit. The performance of these methods with respect to dataset size is shown in Figure 5. Often, traditional models are conceptually simpler. Most DNNs work like a "black-box", which is a big limitation in fundamental science where uncertainty measures and interpretability are desired.

5.1b *Artificial Neural Networks (ANNs)*: ANNs (also known as perceptrons), which are similar to the biological neural networks,[53,54] are one of the most widely applied models in computational studies. ANN can be thought of as transforming the input x into a new feature space, in which it becomes correlated with the output y. When ANNs transform features sequentially through several layers, it is referred to as DNNs. They are excellent tools for identifying patterns and correlations which are far too complex or numerous for a human to extract and manually program.

Each layer consists of one or more artificial neurons (Figure 6). These neurons calculate the weighted sum of the outputs from their preceding neurons and add a bias. Before passing their output to the succeeding neurons, an activation function is used to decide if the value should be "activated" or not. Since the value can range from $-\infty$ to $+\infty$, the type of activation function required is chosen depending on the task. For example, Rectified Linear Unit (ReLU) is an activation function that gives an output x if x is positive and 0 otherwise, and it can be employed in large neural networks for sparsity.

When a neuron contributes to predicting the correct results, the connections associated with it are strengthened, i.e., updated weight values are higher. During feed-forward training, the output of each neuron till the last layer is calculated. After the process, the differences between the predicted and the target outputs are compared to find each neuron's contribution to the errors. A numerical optimization technique called gradient descent is used to update the weight values by backpropagating the errors to the input layer. The learning algorithm is typically represented as:

$$w_{i,j}^{n+1} = w_{i,j}^{n} + \eta(y_j - \hat{y}_j)x_i \tag{4}$$

where x_i is the i_{th} input, y_j is the target value of the j_{th} output, $\hat{y}_j$ is the predicted value, $w_{i,j}$ is the weight between i_{th} input and j_{th} output, n is the n_{th} step, and η is the learning rate. The learning rate is chosen such that the model training can converge in a reasonable time.

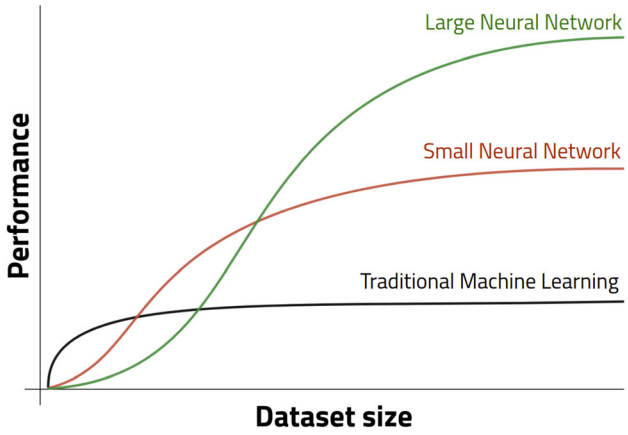

Figure 5. The performance of traditional ML methods and neural networks with respect to dataset size.

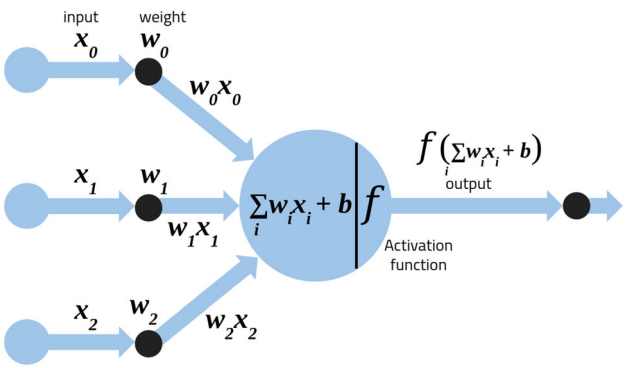

Figure 6. The structure of an Artificial Neuron.

DNNs learn high-level features from data incrementally, with each additional hidden layer capturing higher level features than the previous layer. This eliminates the need for domain expertise and manual feature extraction. Thus, DNNs can automatically learn to extract useful molecular descriptors best suited for the given data. However, since features have to be learned from scratch for every new dataset, these methods can lead to overfitting with limited data.

The most basic type of ANN is a feedforward neural network, in which information travels in only one direction from input to output. There are a variety of others like recurrent neural networks (RNNs), CNNs, etc.

5.1c *Recurrent neural networks (RNNs)*: While training vanilla ANNs, each iteration doesn't remember what it processed in the previous iteration. This is a disadvantage when it comes to identifying patterns and correlations in sequential data, for example, amino acid sequence of proteins. RNNs are ANN architectures capable of remembering data and modelling short-term dependencies due to its recurrent memory cells and are popularly used in sequence modeling and generation. The RNN cell retains the knowledge of what the model saw in the previous time-step when processing the current time-step's information, which may affect the interpretation of the current one. Figure 7 shows a basic pipeline of an RNN sequentially generating molecules *via* SMILES. The output of each RNN cell is fed as input to the next RNN cell. The cells also pass their shared weights that capture the past information in the sequence. Concatenating all the outputs create the completed SMILES for a newly generated molecule.

When training basic RNNs to predict long-term dependencies, the gradient shrinks or explodes as it backpropagates through time - the vanishing and exploding gradient problems.[55,56] This prevents RNNs from learning these features from long sequences. A type of RNN unit, the long short term memory (LSTM) unit or its variant called the gated recurrent unit (GRU), contains "gates" which lessen these gradient problems. These gates decide how much to remember from its past, what to include in its current state, and what to pass on as output to the next gate. The gradients can now be preserved for longer sequences. LSTMs and GRUs are popularly used for inverse molecular design as molecular representations such as SMILES have long-term dependencies like closing parenthesis and rings. For generating molecules using SMILES, the output layer usually gives probabilities for every possible SMILES string token

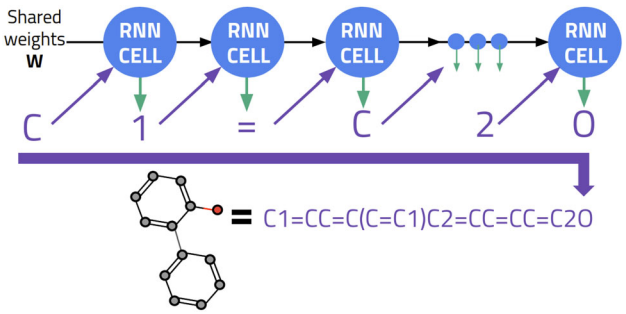

Figure 7. Recurrent Neural Network for sequentially generating molecules *via* SMILES.

and not the character itself because of these strict long-term dependencies. Typically in generative mode, the method is to sample this distribution, while in training mode, the token with the highest probability is chosen.

5.2 *Unsupervised*

Unlike supervised learning, unsupervised learning is the process of learning without labelled data. Instead of picking out specific types of data that are predefined as desired, it simply looks for data that can be grouped based on their similarities. This is why it is also called clustering or grouping. The system is trained using large data and it learns by itself. The following section presents a few examples of unsupervised learning for different tasks.

5.2a *Autoencoders (AEs)*: Studies have aimed to derive molecular descriptors in an unsupervised and data-driven way. In 2016, Gomez-Bombarelli *et al.*[57] created the first ML–based generative model for molecules called CharacterVAE. The model also delivered a data-driven method for molecular descriptors. They developed a variational autoencoder (VAE) to convert the discrete SMILES representation of a molecule to and from a continuous multidimensional representation.

An AE is an ANN architecture for unsupervised feature extraction. It consists of an encoder, a decoder, and a distance function. The encoder compresses the input into a lower-dimensional fixed vector (latent representation), then the decoder reconstructs the vector back into the input. A distance function determines the difference between the original input and the reconstructed output. The objective of the training is to minimize the information loss of the reconstruction. If the input is the chemical representation of a molecule, the bottleneck vector between the networks forces the essential information of the molecule to get compressed, so that the decoder makes as few errors as

possible in the reconstruction. If the compressed vector captures all the necessary information of the given molecule to accurately reconstruct the original chemical representation, it may also capture more general chemical information about the molecule. This idea could be used to acquire molecular descriptors for property prediction ML models.

Vanilla AEs are however not employed for de novo drug design as it is not capable of learning a generalized representation of the molecules. The valid molecules lie on a continuous manifold of functionality, but due to the large number of NN parameters and the relatively small number of training data, it is possible that the AE learns some explicit (non-continuous) mapping of the training set. Thus, the latent space learnt may contain large "dead areas", and the decoder will not be able to decode valid SMILES in the continuous space. VAEs generalise AEs and are capable of forming continuous latent spaces. The model is restricted to learning a latent variable from its input distribution, usually the mean and variance (Figure 8). The restriction encourages all areas of the latent space to correspond to the decoding of valid molecules. When VAEs are trained to reproduce molecules and properties together, the latent space reorganizes in a way that molecules with similar properties are nearby each other.[58,59]

5.2b *Generative adversarial networks (GANs)*: GANs[60] are a rapidly evolving research area. They are a clever way of training a generative model that consists of two sub-models: the generator model G_θ and the discriminator model D_ϕ. These two models are

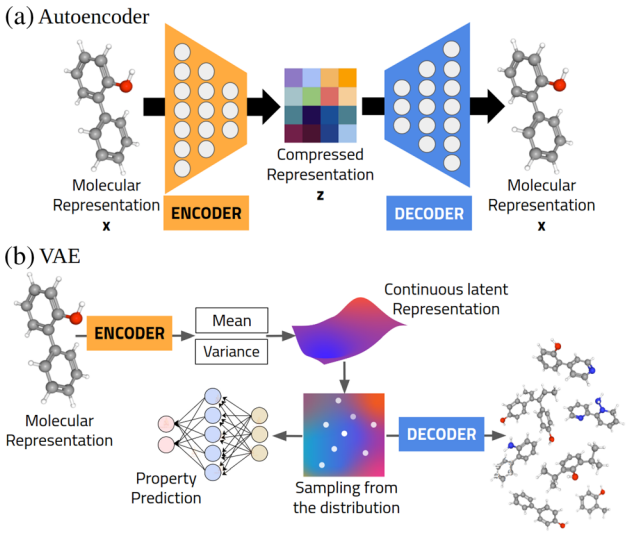

Figure 8. (a) An AE encodes the molecules into a feature space and decodes them back (b) A VAE encodes the molecules into the latent space, which is a continuous numerical representation.

ANNs typically trained together with stochastic gradient descent (SGD). The key idea is that the discriminator's job is to differentiate whether the sample it is looking at was generated by the generator or came from the training dataset. In the de novo molecular design, the sample generated is a molecule, and the training data is a library of valid molecules (Figure 9). The G_θ learns the training data distribution to fool D_ϕ. The distribution is compressed into a latent space, from which the generator draws inputs for creating new molecules.

G_θ and D_ϕ have different objectives, and they can be seen as two players in a minimax game:

$$\min_\theta \max_\phi V(D_\phi, G_\theta) = \mathbb{E}_{x\in p_d(x)}[\log D_\phi(x)] + \mathbb{E}_{z\in p_z(z)}[\log(1 - D_\phi(G_\theta(z)))] \quad (5)$$

where $p_d(x)$ is the data distribution. GANs are implicit generative models, i.e., there's inference of model parameters without the specification of a likelihood. The two models are trained until D is fooled about half the time, meaning G is generating valid molecules from the distribution of the training data. Figure 9 shows the general GAN architecture used for molecular design.

5.2c *Reinforcement learning (RL)* RL is an autonomous, self-teaching algorithm that learns through trial and error dynamically. Like a pet trained using treats and punishments, these algorithms are rewarded when they make the right decisions and penalized when they make the wrong ones. It performs actions with the aim of maximizing rewards. RL has been used in domains like robotics, self-driving cars, and board games.

In RL, the information given to the system is intermediate between supervised and unsupervised learning.[61] The samples for RL don't contain the desired input-output pairs. Instead, they give indications on whether an action is correct or incorrect.

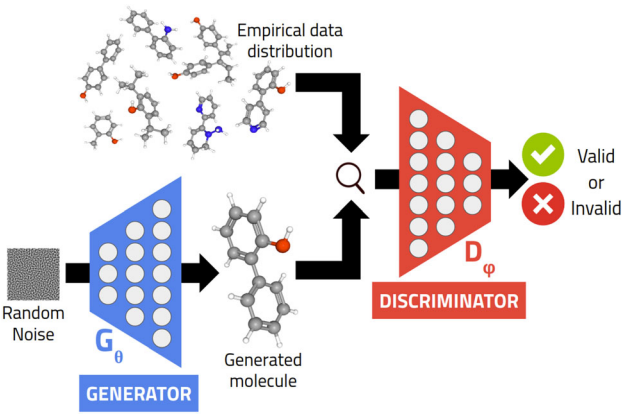

Figure 9. GAN architecture for molecular design.

Given a state s ∈ S, an RL agent has to choose which action a ∈ A has to take, where *S* and *A* are the set of possible states and actions, respectively. For this, the agent learns a policy $\pi(a|s)$ for an unknown dynamic environment, which defines its behavior. Essentially, the policy maps the perceived states to the actions taken therein, with the objective of maximizing its expected reward over time. The reward indicates how good it was to take an action at a certain state.

RL problems are generally framed as Markov decision processes (MDPs). This means there is a fully observable environment with deterministic dynamics where the current state would contain all information necessary to choose an action. Awareness of the past states doesn't add more knowledge. However, this is only an approximation for many real problems. In partially observable Markov decision process (a generalization of MDP), the agent can interact with an incomplete representation of the environment. This has been useful in instances like SMILES generation, as the drug likeliness makes sense to completed SMILES string.

There is a renewed interest in RL,[62] especially when it is combined with DNNs. This is known as deep RL. This can create something fantastic like Deepmind's AlphaGo, an algorithm that beat the world champions of the Go board game. The game has a theoretical complexity of more than 10^{140} possible solutions.[63] An analogy can be seen with the complexity of CCS exploration, showing the potential of the algorithm. RL has been successfully applied in de novo drug design. One of the popular RL approaches involves the agent building new molecules in step–wise fashion.[64,65] Simm *et al.*[64] designed molecules by sequentially drawing atoms from a given bag and placing them onto a 3D *canvas*. Intuitively, the agent is rewarded for placing atoms so that the energy of the resulting molecules is low. Figure 10 shows a general pipeline of a deep RL approach for generating molecules with desired properties *via* SMILES. Here, the agent generates molecules and is rewarded if the molecular properties predicted through the QSAR are desirable. Deep RL can also be employed for optimization of molecules with desired properties.[66,67]

6. Goals and advances

Application of ML methods to problems in chemistry, biology, materials, etc., has taken a giant leap during the last few years.[68] This section presents selected popular fields that have witnessed immense progress through ML.

6.1 *Molecular property prediction*

Since the emergence of atomistic theory, chemists have strived to predict the properties of molecular systems without actually synthesizing them. Molecular property prediction has applications in many fields like quantum mechanics, physical chemistry, biophysics, and physiology.[10,69,70] The molecular properties range from solubility (angstroms) to protein-ligand binding (nanometers) to in vivo toxicity (meters). Recently, it has attracted much attention since it accelerates the discovery of substances with desired characteristics, such as drug design with a specific target.[71–75]

Molecular properties like the total energy of a system are most accurately calculated by QM or Density Functional Theory (DFT) methods, but the process is computationally expensive for an exhaustive exploration of the CCS.[76] The Schrödinger equation (SE) helps us find the electron density for simple systems of small size, but solving it for complex many-body systems is almost impossible. DFT, the computational modelling methods derived or approximated from the SE, are impractical for large systems because the complexity is $O(N_4)$, where N is the number of atoms. For modelling such systems, methods like those involving MM force fields are adopted. Essentially, force fields provide the potential energy of a molecule as a function of nuclear positions.[77] However, these methods improve speed by compromising accuracy. ML methods are replacing traditional calculations at an increasing rate since they can predict properties that are of DFT accuracy and are comparable to MM in

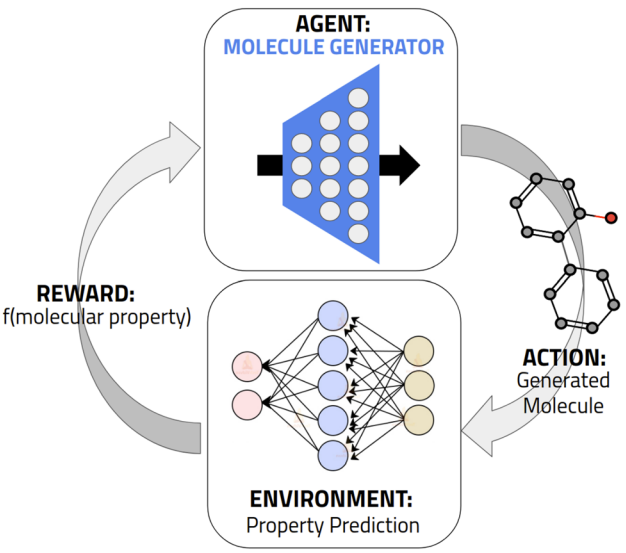

Figure 10. A Reinforcement Learning method where the desired molecular properties are used as a reward for generating desired structures.

terms of speed. These ML methods aim to learn a function that maps a molecule to the property of choice. Just last year, there have been a notable number of scientific papers on ML applications in the prediction of molecular properties.[78–82] There are 3 main steps in learning QSPRs: generating a training set with measured properties, preparing suitable molecular descriptors or inputs, and building an ML architecture to predict the measured properties from the inputs.

Early studies applying ML to QSPR tasks employed linear regression models, which were quickly surpassed by Bayesian neural networks and other approaches.[83,84] In 2012, von Lilienfeld proposed an ML method based on non-linear statistical regression to predict the atomization energies of organic molecules. The supervised learning method used a subset of 7000 stable organic compounds from GDB. Their cartesian coordinates and nuclear charges were encoded into a CM as inputs, without any explicit feature engineering. With a training set of only 1000 compounds, the model achieved a mean absolute error (MAE) of 14.9 kcal/mol. This extraordinary result showed that an ML method could predict QM properties with reasonable accuracy without having to solve the SE explicitly. Over the years, various traditional ML methods have been employed.[85] These methods generally rely on rule-based feature engineering. ANNs are popular among recent state-of-the-art publications.[86]

DL models are capable of automatic feature learning and are widely employed for prediction.[57,87,88] Laghuvarapu *et al.*[28] developed BAND neural network, a DL framework for atomization energy prediction and geometry optimization of small organic molecules. The model was remarkably accurate and robust over the conformational, configurational, and reaction space. It also performed reasonably well on larger molecules than the ones in their training set. Most studies are on organic molecules. Inorganic molecules, especially clusters, need to be studied more. Modee *et al.*[89] introduced the Deep Learning Enabled Topological (DART) model, which uses Topological Atomic Descriptor (TAD) as a feature vector for energy prediction of metal clusters.

Although DL has been successful in property prediction, it is still in its infancy.[69,90,91] In 2017, Goh *et al.* proposed ChemNet for prediction by using 2D RGB images of molecular diagrams as inputs.[88] Grid-like transformations like these usually cause loss of molecular information lying in non-euclidean space, where the molecule's internal spatial and distance information are not complete.[92] Geometric DL encompasses the emerging techniques that aim to generalize DNNs to non-Euclidean domains, such as graphs and manifolds.[92] Graph neural networks (GNNs) achieved superior performance in various domains and have shown great potential for molecular property prediction, as they can directly handle non-euclidean data.[78,82,91,93–96] Variants of G[71]NNs like Message Passing Neural Networks (MPNNs)[93], Schnet[94] and Multiscale Graph Convolutional Networks (MGCNs)[71] use graph representation of molecules for prediction. They have several neural layers to project each node of the graph into latent space with a low dimensional embedding. The node embeddings (interaction messages) are propagated and updated using the embeddings of their neighborhood iteratively. This is called message passing. The node embeddings are then pooled for property prediction. Pathak and others[95,96] developed a GNN-based solution that accurately predicts solvation free energies and is interpretable. The first phase of the model utilized MPNN to compute inter-atomic interaction within both solute and solvent molecules expressed as molecular graphs.

Though GNNs are successful, they are generally data-hungry. Labeled molecules usually span a small portion of the CCS since they can only be generated by expensive and time-consuming techniques. Other unlabelled valid molecules may also have structural benefits. Methods like unsupervised, semi-supervised, and self-supervised learning provide effective solutions to incorporate these unlabelled molecules.[79–81]

Property prediction ML models have achieved high scalability and high prediction quality across both chemical and conformational space. Due to this, they are also employed in various MD simulation tasks like analyzing MD trajectories, and to enhance sampling.[97,99,100]

As explained above, ML has shown extraordinary potential in accurate predictions of quantum mechanical properties such as the electronic energies. These efforts have been accomplished by using supervised learning based on a large amount of pre–computed data. Availability of such data has allowed for circumventing the explicit need to solving the Schrödinger equation. While analytical solution is elusive for multi–electron systems, accurate numerical solutions using configuration interaction and coupled–cluster methods are computationally prohibitive. In practice, a trade–off between computational efficiency (expense) and accuracy is made in making a choice of an appropriate wavefunction approximation.

ANNs are universal approximate functions and few studies have explored their application for obtaining

an *ab initio* solution for many–electron Schrödinger equations. Carleo and Troyer proposed the neural networks to represent the wavefunction that are trained in an unsupervised manner using the variational principle.[101] They showed high accuracy in describing the ground and excited states of interacting spin models in up to two dimensions demonstrating the possibility of applying ANNs for solving quantum many–body systems. Han *et al.* used deep NNs as trial wavefunctions and used variational Monte Carlo method for obtaining the optimal wavefunction (DeepWF).[102] Pfau *et al.* introduced Fermionic neural network (FermiNet) that obeys Fermi-Dirac statistics. They showed quantitative accuracy in calculating the dissociation curves of nitrogen molecule and H_{10}.[103] More recently, in a seminal paper, Hermann *et al.* reported a deep NN representaiton of electronic wavefucntion named PauliNet. They demonstrated that this method outperforms traditional variational methods on systems up to 30 electrons.[104] Using these approaches, the curse of limited basis sets, a major source of inaccuracies in computational quantum mechanical methods is overcome. Applying ANNs for solving many body quantum systems have just begun and research in this direction opens up exciting opportunities in modeling chemical systems efficiently and accurately.

6.2 *Molecular dynamics simulations*

With the advance in algorithms and power of computing resources, MD simulations have become an integral tool for analyzing molecular systems.[10,105] It has helped us analyze thermodynamic and dynamic properties of molecules, create 4D molecular descriptors, probe complex processes such as protein folding and facilitated many other purposes.[106,107] MD is a computer simulation approach for analyzing the time evolution of an interacting molecular system.[108,109] The motion of the system (atomic trajectories) is generated by solving the classical Newtonian dynamic equations for a specific interatomic potential defined by the initial and boundary conditions.[110,111]

The predictive power of the simulations depends on the underlying potential energy surface (PES).[112,113] Hence, they require a precise PES $U(x)$, which is a function of atomic coordinates x. Molecular modeling techniques are mostly based on either QM methods (e.g., DFT), or on force fields (e.g., Stillinger-Weber potentials). Both techniques stand at the opposite sides of the cost-accuracy trade-off. The approximations to $U(x)$ lack transferability. Studies have shown that ML methods are capable of creating interatomic potentials that surpass conventional methods both in terms of accuracy and versatility. As mentioned earlier, they are much faster than QM methods and have comparable accuracy.

In 2007, Behler & Parrinello[73] proposed an ANN solution to extract PES. They achieved transferability through parameter sharing and the summation principle, meaning the network could adjust to molecules of any size. Since then, other ML PES models have emerged, like Deep Potential net and ANI networks. Most ML PES models are based on nonlinear kernel learning or ANNs, each having its own advantages.[99] For elemental solids, Gaussian approximation potentials (GAP)[114,115] are nowadays used in MD simulations. It provides insights into various domains, for example, amorphous states of matter.[116] Pattnaik *et al.*[117] used the data obtained using DFT on small systems and simulated large systems by taking liquid argon as a test case. ML models have been shown to have the potential to mimic MD trajectories produced through simulations.[118–120] Tsai *et al.*[120] used LSTMs to learn the evolution of MD trajectories that were mapped into a sequence of characters in some languages.

In addition to force fields, ML has designed molecular models at resolutions coarser than atomistic models, as atomistic models are computationally expensive to simulate. For example, CGnets can be used to coarse grain away all the solvent molecules in a protein and map the atoms of each residue to the corresponding Ca atom.

ML has made a variety of contributions to the analysis and simulation of MD trajectories.[98,99] For instance, it has enabled the estimation of free energy surfaces. Along with enhanced sampling methods, it has also attempted to learn the free energy surface on the fly. Studies have also employed ML in building Markov state models and dynamic graphical models of molecular kinetics. For example, VAMPnets was developed as a substitution to the complex and error-prone technique of constructing Markov state models. Other contributions of ML in this domain include ML-driven definition of optimal reaction coordinates, enhancement of sampling through learning bias potentials and selection of starting configurations through active learning.

In the field of molecular design, ML can quickly explore vast spaces of CCS for generating molecules of desired properties, avoiding MD simulations altogether. The next section presents this idea.

6.3 *Inverse molecular design*

Molecular design algorithms aim to virtually create and analyze molecules with relevant optimized properties like synthetic accessibility, ADMET (absorption, distribution, metabolism, elimination, and toxicity) profile etc.[121,122] Finding new chemical compounds for drug discovery can be portrayed using the metaphor "finding a needle in a haystack". (Schneider *et al.*, 2019) In this case, the haystack is the universe of synthetically feasible molecules in the CCS, wherein a single molecule with various desired properties is searched for. A clever navigation is required to explore vast chemical spaces efficiently.

Forward strategies for molecular design lead from CCS to the properties using experiments, simulations, gradient-based algorithms, Monte Carlo or genetic algorithms, or combinations thereof. This means that the input is the molecular structure, and the output is the properties of molecules. These direct methods have been successful in their application domains; however, they are unable to quickly cover relevant large chemical spaces.[123]

Inverse molecular design has emerged as an attractive approach to take on these challenges.[58,124] As its name suggests, it inverts the direct approach by taking the desired properties as input and identifying an optimized molecular structure as output. The approach need not necessarily identify one unique structure but a distribution of probable structures. Valid molecules with similar functionalities lie nearby on a continuous curve or manifold. Inverse design uses optimization, sampling, and search methods to navigate the functionality manifold of CCS.[125]

One of the earliest attempts in inverse design was high-throughput virtual screening (HTVS). HTVS is performed to ascertain an initial set of candidate molecules, called "hits". In HTVS, molecules from large small-molecule drug libraries are evaluated for properties such as the binding affinity, against a target receptor. More recent techniques involving optimization can be roughly divided into two types: evolutionary techniques and ML algorithms.[58] Recently, Mehta *et al.*[126] proposed an ML framework "MEMES" based on Bayesian optimization for efficient sampling of chemical space. The architecture identifies 90% of the top-1000 molecules from a dataset of about 100 million molecules, while calculating the docking score only for about 6% of the dataset.

Recent ML-driven methods have accelerated the search for new molecules with desired properties. Generative models such as VAEs,[57,127] RNNs,[128,129] GANs[130] and Generative Pre-Training (GPT)[131] can model complex SPRs and use them to create molecular designs. Pathak *et al.*[59] proposed a deep learning based inorganic material generator (DING) framework that employs conditional variational autoencoders (CVAE) as a generator and DNNs as a predictor of enthalpy of formation, volume per atom and energy per atom. Bagal *et al.*[131] trained a GPT model, named MolGPT, to predict a sequence of SMILES tokens for molecular generation. The model can be trained conditionally to optimize multiple properties of the generated molecules, including scaffold conditioning.

However, these models require large training data for learning valid molecular distributions. In RL, an agent builds new molecules in a step-wise fashion.[64–66] Training an RL agent only requires samples from a reward function. So, the need for a training data is reduced.

The generative process must be restricted or biased towards desirable qualities as mentioned earlier in"AEs"section. In VAEs, the latent space allows direct gradient-based optimization of desired properties, as it's continuous. Nevertheless, the functionality manifold has local minimas. Bayesian optimization or constrained optimization, with Gaussian processes, is applied to explore a smoothed version of the manifold.[58]

In the case of GANs and RNNs dealing with non-continuous data, a gradient estimator is required to backpropagate the generator. RL has been employed as an approach to bias the generation process by rewarding the generator's behaviors. Some examples are methods involving Q-learning and policy gradients (SeqGANs and BGANs). Several studies have adopted RL for the generation of drug-like molecules. Popova *et al.* proposed Reinforcement Learning for Structural Evolution (ReLeaSE), a de novo molecular design method.[132] Molecular applications have adopted models that are a combination of generative algorithms to utilize the advantages from each. For example, druGAN[133] adopts an adversarial autoencoder network, RANC[134] adopts both RL and adversarial network.

Few promising research directions in this domain include structured architectures such as multilevel VAE and inverse RL. Developments in inverse RL may allow for the discovery of reward functions associated with different molecular design tasks.[58]

6.4 *Materials discovery and design*

New materials can contribute to the immense progress in tools and technology.[135,136] Materials discovery and design aim to find candidate materials with desired properties that are synthesizable.[137] This would allow experimental researchers to perform targeted explorations.

Materials screening *via* traditional experiments or computational simulations involve element replacement and structure transformation.[135] The chemical compositional and structural search space tends to be constrained in these methods.[135,138]

ML is employed for finding solutions to various problems in materials science as it has led to a decrease in materials development time and cost.[135,136,139–143] There are now many examples, such as thermoelectrics and photovoltaic materials,[144] metal organic frameworks (MOFs),[145] metallic glass,[146] polymers,[147] and DNA nanostructures,[148] in which ML has been applied to move away from the traditional methods. ML has performed well in areas such as materials property prediction,[149–151] novel materials discovery,[59,152–155] process optimization,[156,157] finding density functionals,[158] and other materials-related studies.[135,159,160]

Finding new chemical components and their crystal structures that likely match the composition and properties of desired materials, is an essential step in novel materials discovery.[136] ML is used to learn and screen for potential combinations of chemical components and structures from a large dataset containing real and synthesized materials. Then, the most-probable crystal structures need to be identified and tested for stability. The number of candidate compounds is still huge because of the extremely large combination space of compositions and structures.[137] Therefore, these candidate new compounds still need to be tested by first-principles calculation (e.g. DFT). Hautier *et al.*[161] demonstrated how the search for novel materials can be accelerated using a combination of ML techniques and high-throughput *ab initio* computations.

Methods involving VAEs have recently been applied to solid-state materials[154] and porous materials.[162] GANs are finding their position in materials design too. A recent application is ZeoGAN[155] – employed in the generation of an energy grid of guest molecules and zeolite structures. RL has been effective for exploring chemical space for different applications, such as MOFs for gas adsorption, and synthesis planning. Dieb *et al.*[163] used RL to design depth-graded multilayer structures, known as supermirrors, for X-ray optics applications. Active learning approaches are also gaining attention in the field. It allows the exploration of new regions of space that were not in the initial dataset.[142,164] This is done by adding new data points to the training set on the fly based on model uncertainty.

6.5 *Other domains*

ML has played roles in several other problems, such as protein–protein interactions, viable retrosynthetic pathways, stability of solids, etc. ML-based scoring functions have been shown to perform significantly better than software like AutoDock Vina for predicting both binding poses and affinities.[165] Finding functionally relevant binding sites on the 3D structure of a protein is crucial for drug design. Aggarwal *et al.*[166] proposed a method that is a combination of geometry–based software and DL, called DeepPocket, that utilises 3D CNNs for making this process accurate.

Results from ML methods in molecular sciences have been applied for many practical purposes. For example, many results of generative models have been used in pharmaceutics.[167] They aid in drug design by generating molecular systems and optimizing relevant medicinal properties such as solubility in water, ADMET profile and synthesizability. Healthcare systems also employ ML to analyse various health-related issues and accelerate decision-making processes efficiently.[168,169] To illustrate, the COVID-19 pandemic has witnessed numerous ML methods such as those by Alle *et al.*[170] and Karthikeyan *et al.*,[171] who have provided risk stratification and mortality prediction models for patients with COVID-19.

Another area of rapid development is imaging and -omics technologies, which will further blur the barrier between cheminformatics and bioinformatics.[172,173] Thus, molecular biology, transcriptomics, proteomics etc. are getting more relevant for ML researchers in molecular sciences.[166,174]

7. Challenges and outlook

Apart from successfully performing desired tasks, ML methods also provide novel insights and transformational ideas. For instance, analysing the weights of trained ML prediction models can potentially lead to automatic discovery of scientific laws and principles, which can cause a revolutionary development in science.[143] Another impressive example is from ML for molecular discovery, where the corresponding

statistical view and analysis of the discovered chemical space leads to fresh insights, discoveries of molecules with unexpected properties, hints for new chemical reaction mechanisms, and more. However, current successful applications of ML in molecular sciences have only scratched the surface of possibilities.[100]

One of the challenges is encoding the essential characteristics of a molecule into its numerical representation. This is one of the most effective ways to infuse physics in ML and generalise better. Attempts have been made to define criteria for the development of molecular descriptors, but adhering to all the criteria is difficult. From the perspective of atomic interactions, current molecular representations describe local chemical interactions well, but completely miss long-range interactions like polarization and van der Waals dispersion. Moreover, capturing highly complex QM interactions like distracted attraction and exchange repulsion, especially in the large molecules (Kollman 1985), has been difficult. An important direction for future progress in studying large complex molecular systems would be incorporating intermolecular interaction theory, such as Hamiltonians for electronic interactions based on SFT, molecular orbital techniques, or the many- body dispersion method, into ML. Further research into the criteria and creation methods of molecular descriptors will be necessary.[46]

Another challenge is the limited amount of labeled molecular data available compared to other domains. This poses the inherent danger of ML models overfitting to benchmarks. Thus, progress needs to be made in reducing the cost of data generation. Due to the combinatorial scaling in CCS, it's also crucial to infuse physics and invariance information in ML and achieve robustness and accuracy using smaller datasets. A few of the promising methods in this context include employing smart sampling methods, identifying valuable data points for training, and employing recent techniques such as transfer learning, meta-learning, or active learning.[175,176] Recently, a bayesian framework performed as well as humans on one-shot learning problems with limited data.[143]

Applying ML in molecular sciences is a young domain. Hence, much of the infrastructure is still in its early stages or waiting to be developed. Drug discovery operates as a feedback loop, where the large number of molecules designed by generative models must be synthesized and validated experimentally to provide feedback for further decision making.[122] These experiments are slow and expensive. Although prediction models can be coupled with generative models to streamline this process, the synthetic tractability of these molecules remain a challenge.[177] Efforts taken in future towards closing the loop need to consider incorporating AI/ML, intelligent systems, embedded systems and robotics into one framework.[58] This can lead to automated laboratories.[178]

This rapidly growing field in computational science, supported by increasing computing power, data sharing and open-source tools, has the potential to solve many theoretical and practical challenges. Beyond these numerous unsolved challenges lies the "chemical discovery revolution!".[116]

Acknowledgements

The authors acknowledge IHub-Data, IIIT Hyderabad for funding. We thank Ms. Indhu Ramachandran for carefully proofreading the manuscript.

References

1. Hiller S A, Golender V E, Rosenblit A B, Rastrigin L A and Glaz A B 1973 Cybernetic methods of drug design. I. Statement of the problem–the perceptron approach *Comput. Biomed. Res.* **6** 411
2. Baskin I I, Winkler D and Tetko I V 2016 A renaissance of neural networks in drug discovery *Expert Opin. Drug Discov.* **11** 785
3. Ramakrishnan R and von Lilienfeld O A 2017 Machine learning, quantum chemistry, and chemical space *Rev. Comput. Chem.* **30** 225
4. AlQuraishi M 2019 AlphaFold at CASP13 *Bioinformatics* **35** 4862
5. Wei G W 2019 Protein structure prediction beyond AlphaFold *Nat. Mach. Intell.* **1** 336
6. Fersht A R 2021 AlphaFold-a personal perspective on the impact of machine learning *J. Mol. Biol.* 167088
7. Senior A W, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson A W, Bridgland A and Penedones H 2019 Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13) *Proteins Struct. Funct. Bioinf.* **87** 1141
8. Mnih V, Kavukcuoglu K, Silver D et al. 2015 Human-level control through deep reinforcement learning *Nature* **518** 529
9. Jordan M I and Mitchell T M 2015. Machine learning: Trends, perspectives, and prospects *Science* **349** 255
10. Hong Y, Hou B, Jiang H and Zhang J 2020 Machine learning and artificial neural network accelerated computational discoveries in materials science *Wiley Interdiscipl. Rev. Comput. Mol. Sci.* **10** e1450
11. Zhu W, Lan C, Xing J, Zeng W, Li Y, Shen L and Xie X 2016 Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. *Proc. AAAI Conf. Artif. Intell.* **30** 1
12. Lecun Y and Bengio Y 1995 Convolutional networks for images, speech, and time-series. In M A Arbib (Ed.) *The handbook of brain theory and neural networks* (MIT Press)

13. Collobert R and Weston J 2008 A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning* (ACM) 160
14. David L, Thakkar A, Mercado R et al. 2020 Molecular representations in AI-driven drug discovery: A review and practical guide *J. Cheminform.* **12** 56
15. Chandrasekaran B, Abed S N, Al-Attraqchi O, Kuche K and Tekade R K 2018 Computer-aided prediction of pharmacokinetic (ADMET) properties. In *Dosage Form Design Parameters 2018 Jan 1* (Academic Press) p. 731
16. Randić M 1991 Generalized molecular descriptors *J. Math. Chem.* **7** 155
17. Todeschini R and Consonni V 2008 *Handbook of Molecular Descriptors* (Wiley)
18. Khan M T and Sylte I 2007 Predictive QSAR modeling for the successful predictions of the ADMET properties of candidate drug molecules *Curr. Drug Discov. Technol.* **4** 141
19. Hansen K, Montavon G, Biegler F, Fazli S, Rupp M, Scheffler M, Von Lilienfeld O, Tkatchenko A and Mller K-R 2013 Assessment and validation of machine learning methods for predicting molecular atomization energies *J. Chem. Theory Comput.* **9** 3404
20. Morgan H L 1965 The generation of a unique machine description for chemical structures—A technique developed at chemical abstracts service *J. Chem. Doc.* **5** 107
21. Gle, R C, Bender A, Arnby C H, Carlsson L, Boyer S and Smith J 2006 Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME *IDrugs Investig. Drugs J.* **9** 199
22. Morgan H L 1965 The generation of a unique machine description for chemical structure *J. Chem. Document.* **5** 107
23. Rogers D and Hahn M 2010 Extended-connectivity fingerprints *J. Chem. Inf. Model.* **50** 742
24. Capecchi A, Probst D and Reymond J L 2020 One molecular fingerprint to rule them all: Drugs, biomolecules, and the metabolome *J. Cheminform.* **12** 43
25. David L, Thakkar A, Mercado R and Engkvist O. Molecular representations in AI-driven drug discovery: A review and practical guide *J. Cheminform.* **12** 1.
26. Todeschini R and Gramatica P 2002 New 3D molecular descriptors: The WHIM theory and QSAR applications *3D QSAR Drug Des.* **355**.
27. Hansen K, Biegler F, Ramakrishnan R, Pronobis W, Von Lilienfeld O A, Muller K R and Tkatchenko A 2015 Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space *J. Phys. Chem. Lett.* **6** 2326
28. Laghuvarapu S, Pathak Y and Priyakumar U D 2020 Band nn: A deep learning framework for energy prediction and geometry optimization of organic small molecules *J. Comput. Chem.* **41** 790
29. Rupp M, Tkatchenko A, Müller K R and Von Lilienfeld O A 2012 Fast and accurate modeling of molecular atomization energies with machine learning *Phys. Rev. Lett.* **108** 058301
30. Ballester P J and Mitchell J B 2010 A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking *Bioinformatics* **26** 1169
31. Pozun Z D, Hansen K, Sheppard D, Rupp M, Müller K R and Henkelman G 2012 Optimizing transition states via kernel-based machine learning *J. Chem. Phys.* **136** 174101
32. Kirkpatrick P and Ellis C 2004 Chemical space *Nature* **432** 823
33. Lipinski C and Hopkins A 2004 Navigating chemical space for biology and medicine *Nature* **432** 85
34. Montavon G, Rupp M, Gobre V, Vazquez-Mayagoitia A, Hansen K, Tkatchenko A, Müller K-R and Anatole von Lilienfeld O 2013 Machine learning of molecular electronic properties in chemical compound space *New J. Phys.* **15** 095003
35. Ramakrishnan R, Dral P O, Rupp M and von Lilienfeld O A 2014 Quantum chemistry structures and properties of 134 kilo molecules *Sci. Data* **1** 140022
36. Ruddigkeit L, van Deursen R, Blum L C and Reymond J-L 2012 Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17 *J. Chem. Inform. Model.* **52** 2864
37. Tang J et al. 2014 Making sense of large-scale kinase inhibitor bioactivity data sets: A comparative and integrative analysis *J. Chem. Inf. Model.* **54** 735
38. Davis M I et al. 2011 Comprehensive analysis of kinase inhibitor selectivity *Nat. Biotechnol.* **29** 1046
39. Mobley D L, Guthrie J P 2014 FreeSolv: A database of experimental and calculated hydration free energies, with input files *J. Comput. Aided Mol. Des.* **28** 711
40. Marenich A V, Kelly C P, Thompson J D, Hawkins G D, Chambers C C, Giesen D J, Winget P, Cramer C J, Truhlar D G 2020 *Minnesota Solvation Database (MNSOL) Version 2012*. Retrieved from the Data Repository for the University of Minnesota
41. Steinbeck C, Kuhn S 2004 NMRShiftDB—compound identification and structure elucidation support through a free community-built web database *Phytochemistry* **65** 2711
42. Polykovskiy D, Zhebrak A, Sanchez-Lengeling B, Golovanov S, Tatanov O, Belyaev S, Kurbanov R, Artamonov A, Aladinskiy V, Veselov M and Kadurin A 2020 Molecular sets (MOSES): a benchmarking platform for molecular generation models *Front. Pharmacol.* **18** 1931
43. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker B A, Thiessen P A, Yu B and Zaslavsky L 2021 PubChem in 2021: New data content and improved web interfaces *Nucleic Acids Res.* **49** D1388
44. Wang Y, Xiao J, Suzek T O, Zhang J, Wang J and Bryant S H 2009 PubChem: A public information system for analyzing bioactivities of small molecules *Nucleic Acids Res.* **37** W623
45. Hansch C, Maloney P P, Fujita T and Muir R M. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients *Nature* **194**178
46. Haghighatlari M and Hachmann J 2019 Advances of machine learning in molecular modeling and simulation *Curr. Opin. Chem. Eng.* **23** 51

47. Hastie T, Tibshirani R and Friedman J 2011 *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (New York: Springer)
48. Faber F A, Hutchison L, Huang B, Gilmer J, Schoenholz S S, Dahl G E, Vinyals O, Kearnes S, Riley P F and Von Lilienfeld O A 2017 Prediction errors of molecular machine learning models lower than hybrid DFT error *J. Chem. Theory Comput.* **13** 5255
49. Huang B and Von Lilienfeld O A 2016 Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity *J. Chem. Phys.* **145** 161102
50. McDonagh J L, Silva A F, Vincent M A and Popelier P L 2017 Machine learning of dynamic electron correlation energies from topological atoms *J. Chem. Theory Comput.* **14** 216
51. Meyer J G, Liu S, Miller I J, Coon J J and Gitter A 2019 Learning drug functions from chemical structures with convolutional neural networks and random forests *J. Chem. Inf. Model.* **59** 4438
52. Zou H and Hastie T 2005 Regularization and variable selection via the elastic net *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **67** 301
53. Freeman B, Lowel S and Singer W 1987 Deoxyglucose mapping in the cat visual-cortex following carotid-artery injection and cortical flat-mounting *J. Neurosci. Methods* **20** 115
54. Krogh A 2008 What are artificial neural networks? *Nat. Biotechnol.* **26** 195
55. Hochreiter S *et al.* 2001 Gradient flow in recurrent nets: The difficulty of learning long-term dependencies
56. Agar J C, Naul B, Pandya S and van Der Walt S 2019 Revealing ferroelectric switching character using deep recurrent neural networks *Nat. Commun.* **10** 1
57. Gómez-Bombarelli R, Wei J N, Duvenaud D, Hernàndez-Lobato J M, Sànchez-Lengeling B, Sheberla D, Aguilera-Iparraguirre J, Hirzel T D, Adams R P and Aspuru-Guzik A 2018 Automatic chemical design using a data-driven continuous representation of molecules *ACS Central Sci.* **4** 268
58. Sanchez-Lengeling B and Aspuru-Guzik A 2018 Inverse molecular design using machine learning: Generative models for matter engineering *Science* **361** 360
59. Pathak Y, Juneja K S, Varma G, Ehara M and Priyakumar U D 2020 Deep learning enabled inorganic material generator *Phys. Chem. Chem. Phys.* **22** 26935
60. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y 2020 Generative adversarial networks *Commun. ACM* **63** 139
61. Sutton R S and Barto A G 2018 *Reinforcement Learning: An Introduction* (MIT Press)
62. Krakovsky M 2016 Reinforcement Renaissance *Commun. ACM* **59** 12
63. van den Herik H J, Uiterwijk J W H M and van Rijswijck J 2002 Games solved: Now and in the future *Artif. Intell.* **134** 277
64. Simm G, Pinsler R and Hernàndez-Lobato J M 2020 Reinforcement learning for molecular design guided by quantum mechanics. In *International Conference on Machine Learning 2020 Nov 21* (PMLR) 8959
65. Olivecrona M, Blaschke T, Engkvist O and Chen H 2017 Molecular de-novo design through deep reinforcement learning *J. Cheminform.* **9 1**
66. Zhou Z, Kearnes S, Li L, Zare R N and Riley P 2019 Optimization of molecules via deep reinforcement learning *Sci. Rep.* **9** 1
67. Ahuja K, Green W H and Li Y P 2021 Learning to optimize molecular geometries using reinforcement learning *J. Chem. Theory Comput.* **17** 818
68. Murugan N A, Poongavanam V and Priyakumar U D 2019 Recent advancements in computing reliable binding free energies in drug discovery projects In *Structural Bioinformatics: Applications in Preclinical Drug Discovery Process 2019* (SpringerChem) 221
69. Wu Z, Ramsundar B, Feinberg E N, Gomes J, Geniesse C, Pappu A S, Leswing K and Pande V 2018 MoleculeNet: A benchmark for molecular machine learning *Chem. Sci* **9** 513
70. Ramakrishnan R, Dral P O, Rupp M and von Lilienfeld O A 2015 Big data meets quantum chemistry approximations: The △ machine learning approach *J. Chem. Theory Comput.* **11** 2087
71. Lu C, Liu Q, Wang C, Huang Z, Lin P and He L 2019 Molecular property prediction: A multilevel quantum interactions modeling perspective In *Proceedings of the AAAI Conference on Artificial Intelligence 2019 Jul 17* (Vol. 33, No. 01) 1052
72. Schneider G and Wrede P 1998 Artificial neural networks for computer-based molecular design *Prog. Biophys. Mol. Biol.* **70** 175
73. Behler J and Parrinello M 2007 Generalized neural-network representation of high-dimensional potentialenergy surfaces *Phys. Rev. Lett.* **98** 146401
74. Varnek A and Baskin I 2012 Machine learning methods for property prediction in chemoinformatics: Auo vadis? *J. Chem. Inf. Model.* **52** 1413
75. Bartók A P, Kondor R and Csànyi G 2013 On representing chemical environments *Phys. Rev. B* **87** 184115
76. Ratcliff L E, Mohr S, Huhs G, Deutsch T, Masella M and Genovese L 2017 Challenges in Large Scale Quantum Mechanical Calculations *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **7** e1290
77. Pissurlenkar R R, Shaikh M S, Iyer R P and Coutinho E C. Molecular mechanics force fields and their applications in drug design *Anti Infect. Agents Med. Chem. (Form. Curr. Med. Chem. Anti Infect. Agents)* **8** 128
78. Lamb G and Paige B 2020 Bayesian Graph Neural Networks for Molecular Property Prediction arXiv preprint arXiv:2012.02089
79. Sun F Y, Hoffmann J, Verma V and Tang J 2019 Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization arXiv preprint arXiv:1908.01000.
80. Chithrananda S, Grand G and Ramsundar B 2020 ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction arXiv preprint arXiv:2010.09885
81. Hao Z, Lu C, Huang Z, Wang H, Hu Z, Liu Q, Chen E and Lee C 2020 ASGN: An active semi-supervised graph neural network for molecular property

prediction In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining 2020* Aug 23 731
82. Wang Z, Liu M, Luo Y, Xu Z, Xie Y, Wang L, Cai L and Ji S 2020 MoleculeKit: Machine learning methods for molecular property prediction and drug discovery. arXiv preprint arXiv:2012.01981
83. Ajay, Walters W P and Murcko M A 1998 Can we learn to distinguish between drug-like and nondrug-like molecules? *J. Med. Chem.* **41** 3314
84. Burden F R, Ford M G, Whitley D C and Winkler D A 2000 Use of automatic relevance determination in qsar studies using bayesian neural networks *J. Chem. Inf. Comput. Sci.* **40** 1423
85. Bose S, Dhawan D, Nandi S, Sarkar R R and Ghosh D 2018 Machine learning prediction of interaction energies in rigid water clusters *Phys. Chem. Chem. Phys* **20** 22987
86. Smith J S, Nebgen B T, Zubatyuk R, Lubbers N, Devereux C, Barros K, Tretiak S, Isayev O and Roitberg A E 2019 Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning *Nat. Commun.* **10** 1
87. Goh G B, Siegel C, Vishnu A, Hodas N O and Baker N 2017 Chemception: A deep neural network with minimal chemistry knowledge matches the performance of expert-developed QSAR/QSPR models arXiv preprint arXiv:1706.06689
88. Goh G B, Siegel C M, Vishnu A and Hodas N O 2017 Chemnet: A transferable and generalizable deep neural network for small-molecule property prediction (No. PNNL-SA-129942) Pacific Northwest National Lab.(PNNL), Richland, WA (United States)
89. Modee R, Agarwal S, Verma A, Joshi K and Priyakumar U D 2021 DART: Deep Learning Enabled Topological Interaction Model for Energy Prediction of Metal Clusters and its Application in Identifying Unique Low Energy Isomers. ChemRxiv. 2021. https://doi.org/10.26434/chemrxiv.14672682.v1
90. Mayr A, Klambauer G, Unterthiner T, Steijaert M, Wegner J K, Ceulemans H, Clevert D A, Hochreiter S 2018 Large-scale comparison of machine learning methods for drug target prediction on ChEMBL *Chem. Sci.* **9** 5441
91. Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, Guzman-Perez A, Hopper T, Kelley B, Mathea M and Palmer A 2019 Analyzing learned molecular representations for property prediction *J. Chem. Inf. Model.* **59** 3370
92. Bronstein M M, Bruna J, LeCun Y, Szlam A and Vandergheynst P 2017 Geometric deep learning: going beyond euclidean data *IEEE Signal Process. Mag.* **34** 18
93. Gilmer J, Schoenholz S S, Riley P F, Vinyals O and Dahl G E 2017 Neural message passing for quantum chemistry. In *International Conference on Machine Learning 2017 Jul 17* (PMLR) 1263
94. Schütt K T, Kindermans P J, Sauceda H E, Chmiela S, Tkatchenko A and Müller K R 2017 Schnet: A continuous-filter convolutional neural network for modeling quantum interactions arXiv preprint arXiv:1706.08566
95. Pathak Y, Laghuvarapu S, Mehta S and Priyakumar U D. Chemically interpretable graph interaction network for prediction of pharmacokinetic properties of drug-like molecules In *Proceedings of the AAAI Conference on Artificial Intelligence 2020 Apr 3* (Vol. 34, No. 01) 873
96. Pathak Y, Mehta S and Priyakumar U D 2021 Learning atomic interactions through solvation free energy prediction using graph neural networks *J. Chem. Inf. Model.* **61** 689
97. Wang Y, Ribeiro J M and Tiwary P 2020 Machine learning approaches for analyzing and enhancing molecular dynamics simulations *Curr. Opin. Struct. Biol.* **61**139
98. Chattopadhyay A, Zheng M, Waller MP, Priyakumar U D 2018 A probabilistic framework for constructing temporal relations in replica exchange molecular trajectories *J. Chem. Theory Comput.* **14** 3365
99. Noé F, Tkatchenko A, Müller K R, Clementi C 2020 Machine learning for molecular simulation *Annu. Rev. Phys. Chem.* **71** 361
100. von Lilienfeld O A, Müller K R and Tkatchenko A 2020 Exploring chemical compound space with quantum-based machine learning *Nat. Rev. Chem.* **4** 347
101. Carleo G, Troyer M 2017 Solving the quantum many-body problem with artificial neural networks *Science* **355** 602
102. Han J, Zhang L and Weinan E 2019 Solving many-electron Schrodinger equation using deep neural networks *J. Comput. Phys.* **399** 108929
103. Pfau D, Spencer J S, Matthews A G and Foulkes W M 2020 Ab initio solution of the many-electron Schrödinger equation with deep neural networks *Phys. Rev. Res.* **2** 033429
104. Hermann J, Schätzle Z and Noé F 2020 Deep-neural-network solution of the electronic Schrödinger equation *Nat. Chem.* **12** 891
105. Hospital A, Goñi J R, Orozc, M and Gelpí J L 2015 Molecular dynamics simulations: Advances and applications *Adv. Appl. Bioinform. Chem. AABC* **8** 37
106. Barducci A, Bonomi M, Prakash M K and Parrinello M 2013 Free-energy landscape of protein oligomerization from atomistic simulations *Proc. Nat. Acad. Sci.* **110** E4708
107. Palazzesi F, Prakash M K, Bonomi M and Barducci A 2015 Accuracy of current all-atom force-fields in modeling protein disordered states *J. Chem. Theory Comput.* **11** 2
108. McCammon J A, Gelin B R and Karplus M 1977 Dynamics of folded proteins *Nature* **267** 585
109. Warshel A and Levitt M 1976 Theoretical studies of enzymic reactions–dielectric, electrostatic and steric stabilization of carbonium-ion in reaction of lysozyme *J. Mol. Biol.* **103** 227
110. Roy K, Kar S and Das R N 2015 *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment* (Academic Press) 151
111. Brown S, Tauler R and Walczak B (Eds.) 2020 *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis* (Elsevier) 26

112. Chiriki S, Jindal S and Bulusu S S 2017 Neural network potentials for dynamics and thermodynamics of gold nanoparticles *J. Chem. Phys.* **103** 227
113. Chiriki S and Bulusu S S 2016 Modeling of DFT quality neural network potential for sodium clusters: Application to melting of sodium clusters (Na20 to Na40) *Chem. Phys. Lett.* **652** 130
114. Bartók A P, Payne M C, Kondor R, Csányi G 2010 Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons *Phys. Rev. Lett.* **104**, 136403
115. Bartóók A P, Csányi G 2015 Gaussian approximation potentials: A brief tutorial introduction *Int. J. Quantum Chem.* **115** 1051
116. Tkatchenko A 2020 Machine learning for chemical discovery *Nat. Commun.* **11** 1
117. Pattnaik P, Raghunathan S, Kalluri T, Bhimalapuram P, Jawahar C V and Priyakumar U D 2020 Machine learning for accurate force calculations in molecular dynamics simulations *J. Phys. Chem. A.* **124** 6954
118. Eslamibidgoli M J, Mokhtari M and Eikerling M H 2019 Recurrent neural network-based model for accelerated trajectory analysis in aimd simulations. Preprint at arxiv:1909.10124
119. Pathak J, Hunt B, Girvan M, Lu Z and Ott E 2018 Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach *Phys. Rev. Lett.* **120** 024102
120. Tsai S T, Kuo E J and Tiwary P 2020 Learning molecular dynamics with simple language model built upon long short-term memory neural network *Nat. Commun.* **11** 1
121. Elton D C, Boukouvalas Z, Fuge M D and Chung P W 2019 Deep learning for molecular design-a review of the state of the art *Mol. Syst. Des. Eng.* **4** 828
122. Brown N 2015 *In Silico Medicinal Chemistry: Computational Methods to Support Drug Design* Royal Society of Chemistry
123. Kuhn C and Beratan D N 1996 Inverse strategies for molecular design *J. Phys. Chem.* **100** 10595
124. Shiraogawa T and Ehara M 2020 Theoretical design of photofunctional molecular aggregates for optical properties: an inverse design approach *J. Phys. Chem. CC.* **124** 13329
125. Pollice R, dos Passos Gomes G, Aldeghi M, Hickman R J, Krenn M, Lavigne C, Lindner-D'Addario M, Nigam A, Ser C T, Yao Z and Aspuru-Guzik A 2021 Data-driven strategies for accelerated materials design *Acc. Chem. Res.* **2** 1120
126. Mehta S, Laghuvarapu S, Pathak Y, Sethi A, Alvala M and Priyakumar U D 2021 *Enhanced Sampling of Chemical Space for High Throughput Screening Applications using Machine Learning* ChemRxiv. Cambridge: Cambridge Open Engage
127. Jin W, Barzilay R and Jaakkola T 2018 Junction tree variational autoencoder for molecular graph generation In *International Conference on Machine Learning 2018 Jul 3* (PMLR) 2323
128. Kim K, Kang S, Yoo J, Kwon Y, Nam Y, Lee D, Kim I, Choi Y S, Jung Y, Kim S and Son W J 2018 Deep-learning-based inverse design model for intelligent discovery of organic molecules *npj Computational Materials* **4** 1
129. Segler M H, Preuss M and Waller M P 2018 Planning chemical syntheses with deep neural networks and symbolic AI *Nature* **555** 604
130. De Cao N and Kipf T 2018 MolGAN: An implicit generative model for small molecular graphs. arXiv preprint arXiv:1805.11973
131. Bagal V, Aggarwal R, Vinod P K and Priyakumar U D 2021 *LigGPT: Molecular Generation Using a Transformer-Decoder Model ChemRxiv* (Cambridge: Cambridge Open Engage)
132. Popova M, Isayev O and Tropsha A 2018 Deep reinforcement learning for de novo drug design *Sci Adv.* **4** eaap7885
133. Kadurin A, Nikolenko S, Khrabrov K, Aliper A, Zhavoronkov A 2017 druGAN: An advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico *Mol. Pharm.* **14** 3098
134. Putin E, Asadulaev A, Ivanenkov Y, Aladinskiy V, Sanchez-Lengeling B, Aspuru-Guzik A and Zhavoronkov A 2018 Reinforced adversarial neural computer for de novo molecular design *J. Chem. Inf. Model.* **58** 1194
135. Liu Y, Zhao T, Ju W and Shi S 2017 Materials discovery and design using machine learning *J. Materiomics.* **3** 159
136. Saal J E, Oliynyk A O and Meredig B 2020 Machine learning in materials discovery: Confirmed predictions and their underlying approaches *Annu. Rev. Mater. Res.* **50** 49
137. Liu Y, Guo B, Zou X, Li Y and Shi S 2020 Machine learning assisted materials design and discovery for rechargeable batteries *Energy Storage Mater.*
138. Meredig B, Agrawal A, Kirklin S, Saal J E, Doak J W, Thompson A, Zhang K, Choudhary A and Wolverton C 2014 Combinatorial screening for new materials in unconstrained composition space with machine learning *Phys. Rev. B* **89**
139. Moosavi S M, Jablonka K M and Smit B 2020 The role of machine learning in the understanding and design of materials *J. Am. Chem. Soc.* **142** 20273
140. Juan Y, Dai Y, Yang Y and Zhang J 2020 Accelerating materials discovery using machine learning *J. Mater. Sci. Technol.*
141. Schmidt J, Marques M R, Botti S and Marques M A 2019 Recent advances and applications of machine learning in solid-state materials science *NPJ Comput. Mater.* **5** 1
142. Vasudevan R, Pilania, Balachandran P V 2021 *Machine Learning for Materials Design and Discovery* 070401.
143. Butler K T, Davies D W, Cartwright H, Isayev O and Walsh A 2018 Machine learning for molecular and materials science *Nature* **559** 547
144. Butler K T, Frost J M, Skelton J M, Svane K L and Walsh A 2016 Computational materials design of crystalline solids *Chem. Soc. Rev.* 45, 6138
145. Yaghi O M, Kalmutzki M J and Diercks C S 2019 *Introduction to Reticular Chemistry: MetalOrganic Frameworks and Covalent Organic Frameworks* (Wiley)

146. Ward L, O'Keeffe S C, Stevick J, Jelbert G R, Aykol M and Wolverton C 2018 A machine learning approach for engineering bulk metallic glass alloys *Acta Mater.* **159** 102
147. Allcock H R 1992 Rational design and synthesis of new polymeric material *Science* **255** 11061112
148. Jones M R, Seeman N C, Mirkin C A 2015 Programmable materials and the nature of the DNA bond *Science* **347** 1260901
149. Dureckova H, Krykunov M, Aghaji M Z, Woo T K. Robust machine learning models for predicting high CO_2 working capacity and CO_2/H_2 selectivity of gas adsorption in metal organic frameworks for pre-combustion carbon capture *J. Phys. Chem. C* **123** 4133
150. Rajan A C, Mishra A, Satsangi S, Vaish R, Mizuseki H, Lee K R and Singh A K 2018 Machine-learning-assisted accurate band gap predictions of functionalized MXene *Chem. Mater.* **30** 4031
151. Kapse S, Janwari S, Waghmare U V and Thapa R 2021 Energy parameter and electronic descriptor for carbon based catalyst predicted using QM/ML *Appl. Catal. B Environ.* **286** 119866.
152. Kim K, Kang S, Yoo J, Kwon Y, Nam Y et al 2018 Deep-learning-based inverse design model for intelligent discovery of organic molecules *NPJ Comput. Mater.* **4** 67
153. Gómez-Bombarelli R, Aguilera-Iparraguirre J, Hirzel T D, Duvenaud D, Maclaurin D *et al.* 2016 Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach *Nat. Mater.* **15** 1120
154. Noh J, Kim J, Stein H S, Sanchez-Lengeling B, Gregoire J M, Aspuru-Guzik A and Jung Y 2019 Inverse design of solid-state materials via a continuous representation *Matter* **1** 13701384
155. Kim B, Lee S and Kim J 2020 Inverse design of porous materials using artificial neural networks *Sci. Adv.* **6** eaax9324
156. Han Y F, Zeng W D, Shu Y, Zhou Y G, Yu H Q 2011 Prediction of the mechanical properties of forged Ti-10V-2Fe-3Al titanium alloy using FNN *Comput. Mater. Sci.* **50** 1009
157. Zhu Q, Abbod M F, Talamantes-Silva J, Sellars C M, Linkens D A and Beynon J H 2003 Hybrid modelling of aluminium-magnesium alloys during thermomechanical processing in terms of physically-based, neuro-fuzzy and finite element models *Acta Mater.* **51** 5051
158. Snyder J C, Rupp M, Hansen K, Muller K R and Burke K Finding density functionals with machine learning *Phys. Rev. Lett.* **108**253002
159. Cai J, Chu X, Xu K, Li H and Wei J 2020 Machine learning-driven new material discovery *Nanoscale Adv.* **2** 3115
160. Singh S, Pareek M, Changotra A, Banerjee S, Bhaskararao B, Balamurugan P and Sunoj R B 2020 A unified machine-learning protocol for asymmetric catalysis as a proof of concept demonstration using asymmetric hydrogenation *Proc. Nat. Acad. Sci.* **117** 1339
161. Hautier G, Fischer C C, Jain A, Mueller T and Ceder G 2010 Finding nature's missing ternary oxide compounds using machine learning and density functional theory *Chem. Mater.* **22** 3762
162. Yao Z, Sanchez-Lengeling B, Bobbitt N S, Bucior B J, Kumar S G H, Collins S P, Burns T, Woo T K, Farha O, Snurr R Q and Aspuru-Guzik A 2021 Inverse design of nanoporous crystalline reticular materials with deep generative models. *Nat. Mach. Intell.* **3** 76
163. Dieb S, Song Z, Yin W-J, and Ishii M 2020 Optimization of depth-graded multilayer structure for x-ray optics using machine learning *J. Appl. Phys.* **128** 074901
164. Tian Y, Yuan R, Xue D, Zhou Y, Ding X, Sun J and Lookman T 2020 Role of uncertainty estimation in accelerating materials development via active learning *J. Appl. Phys.* **128** 014103
165. Ragoza M, Hochuli J, Idrobo E, Sunseri J and Koes D R 2017 Protein-ligand scoring with convolutional neural networks *J. Chem. Inf. Model.* **57** 942
166. Aggarwal R, Gupta A, Chelur V, Jawahar C V and Priyakumar U D 2021 DeepPocket: Ligand binding site detection and segmentation using 3D convolutional neural networks. *J. Chem. Inf. Model.* 2021
167. Schneide G 2018 Generative models for artificially-intelligent molecular design *Mol. Inf.* **37** 1880131
168. Khare Y, Bagal V, Mathew M, Devi A, Priyakumar U D and Jawahar C 2021 MMBERT: Multimodal BERT Pretraining for Improved Medical VQA. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), 2021* 1033
169. Sruthi C K, Biswal M R, Saraswat B, Joshi H, Prakash M K. Predicting and interpreting COVID-19 transmission rates from the ensemble of government policies. medRxiv. 2020 Jan 1.
170. Alle S, Siddiqui S, Kanakan A, Garg A, Karthikeyan A, Mishra N, Waghdhare S, Tyagi A, Tarai B, Hazarika P P and Das P 2020 *COVID-19 Risk Stratification and Mortality Prediction in Hospitalized Indian Patients* medRxiv. https://doi.org/10.1101/2020.12.19.20248524
171. Karthikeyan A, Garg A, Vinod P K and Priyakumar U D 2021 Machine learning based clinical decision support system for early COVID-19 mortality prediction *Front. Public Health* **9**
172. Moolamalla S T R, Chauhan R, Priyakumar U D and Vinod P K 2020 Host metabolic reprogramming in response to SARS-Cov-2 infection *bioRxiv*. https://doi.org/10.1101/2020.08.02.232645
173. Nagamani S and Sastry G N 2021 *Mycobacterium tuberculosis* cell wall permeability model generation using chemoinformatics and machine learning approaches *ACS Omega* **6** 17472
174. Yashas B L Samaga, Shampa Raghunathan and Deva Priyakumar U 2021 SCONES: Self-consistent neural network for protein stability prediction upon mutation *J. Phys. Chem. B* **125** 10657
175. Cohn D A, Ghahramani Z and Jordan M I 1995 Active learning with statistical models In *Advances in Neural Information Processing Systems 7*. Tesauro G,

Touretzky DS, Leen TK (Eds) (The MIT Press, Cambridge, MA, USA) p. 705
176. Reker D and Schneider G 2015 Active-learning strategies in computer-assisted drug discovery *Drug Discov. Today* **20** 458
177. Sellwood M A, Ahmed M, Segler M H and Brown N 2018 Artificial intelligence in drug discovery *Future Med. Chem.* **10** 2025
178. Schneider G 2018 Automating drug discovery *Nat. Rev. Drug Discov.* **17** 97