HHS Public Access

Author manuscript

Chem Rev. Author manuscript; available in PMC 2024 April 07.

Published in final edited form as:

Chem Rev. 2023 July 12; 123(13): 8736–8780. doi:10.1021/acs.chemrev.3c00189.

Machine Learning Methods for Small Data Challenges in Molecular Science

Bozheng Dou,
Research Center of Nonlinear Science, School of Mathematical and Physical Sciences, Wuhan Textile University, Wuhan 430200, P, R. China

Zailiang Zhu,
Research Center of Nonlinear Science, School of Mathematical and Physical Sciences, Wuhan Textile University, Wuhan 430200, P, R. China

Ekaterina Merkurjev,
Department of Mathematics, Michigan State University, East Lansing, Michigan 48824, United States

Lu Ke,
Research Center of Nonlinear Science, School of Mathematical and Physical Sciences, Wuhan Textile University, Wuhan 430200, P, R. China

Long Chen,
Research Center of Nonlinear Science, School of Mathematical and Physical Sciences, Wuhan Textile University, Wuhan 430200, P, R. China

Jian Jiang,
Research Center of Nonlinear Science, School of Mathematical and Physical Sciences, Wuhan Textile University, Wuhan 430200, P, R. China

Department of Mathematics, Michigan State University, East Lansing, Michigan 48824, United States

Yueying Zhu,
Research Center of Nonlinear Science, School of Mathematical and Physical Sciences, Wuhan Textile University, Wuhan 430200, P, R. China

Jie Liu,
Research Center of Nonlinear Science, School of Mathematical and Physical Sciences, Wuhan Textile University, Wuhan 430200, P, R. China

Bengong Zhang,

Corresponding Authors **Jian Jiang** – Research Center of Nonlinear Science, School of Mathematical and Physical Sciences, Wuhan Textile University, Wuhan 430200, P, R. China; Department of Mathematics, Michigan State University, East Lansing, Michigan 48824, United States; jjiang@wtu.edu.cn, **Guo-Wei Wei** – Department of Mathematics, Michigan State University, East Lansing, Michigan 48824, United States; Department of Electrical and Computer Engineering and Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, Michigan 48824, United States; weig@msu.edu.

The authors declare no competing financial interest.

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.chemrev.3c00189

Research Center of Nonlinear Science, School of Mathematical and Physical Sciences, Wuhan Textile University, Wuhan 430200, P, R. China

Guo-Wei Wei

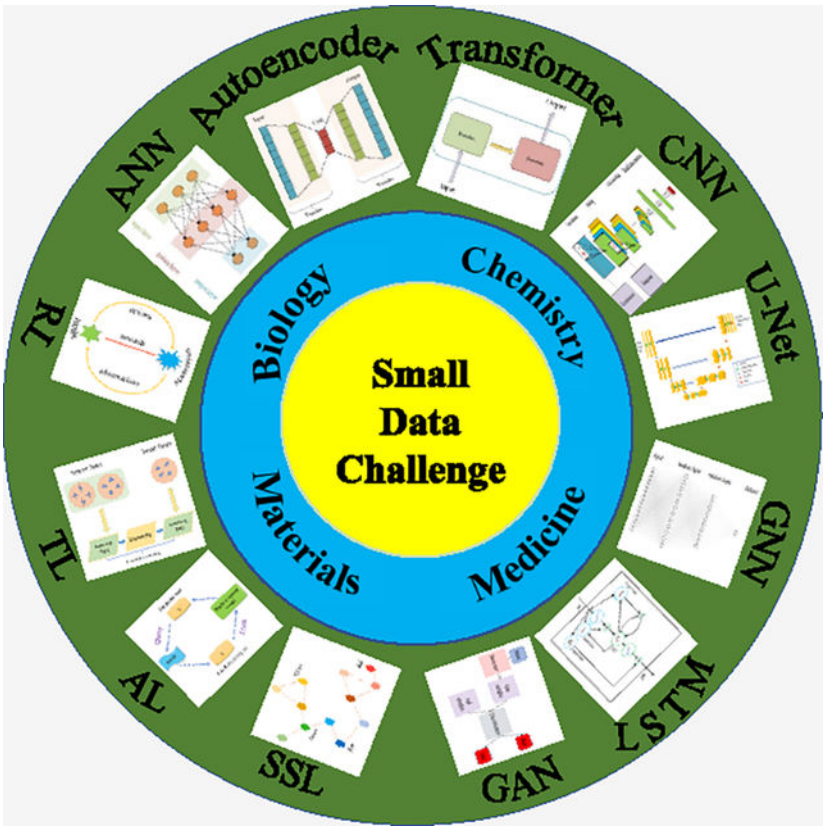
Department of Mathematics, Michigan State University, East Lansing, Michigan 48824, United States

Department of Electrical and Computer Engineering and Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, Michigan 48824, United States

Abstract

Small data are often used in scientific and engineering research due to the presence of various constraints, such as time, cost, ethics, privacy, security, and technical limitations in data acquisition. However, big data have been the focus for the past decade, small data and their challenges have received little attention, even though they are technically more severe in machine learning (ML) and deep learning (DL) studies. Overall, the small data challenge is often compounded by issues, such as data diversity, imputation, noise, imbalance, and high-dimensionality. Fortunately, the current big data era is characterized by technological breakthroughs in ML, DL, and artificial intelligence (AI), which enable data-driven scientific discovery, and many advanced ML and DL technologies developed for big data have inadvertently provided solutions for small data problems. As a result, significant progress has been made in ML and DL for small data challenges in the past decade. In this review, we summarize and analyze several emerging potential solutions to small data challenges in molecular science, including chemical and biological sciences. We review both basic machine learning algorithms, such as linear regression, logistic regression (LR), *k*-nearest neighbor (KNN), support vector machine (SVM), kernel learning (KL), random forest (RF), and gradient boosting trees (GBT), and more advanced techniques, including artificial neural network (ANN), convolutional neural network (CNN), U-Net, graph neural network (GNN), Generative Adversarial Network (GAN), long short-term memory (LSTM), autoencoder, transformer, transfer learning, active learning, graph-based semi-supervised learning, combining deep learning with traditional machine learning, and physical model-based data augmentation. We also briefly discuss the latest advances in these methods. Finally, we conclude the survey with a discussion of promising trends in small data challenges in molecular science.

Graphical Abstract



1. INTRODUCTION

In recent years, machine learning (ML), including deep learning (DL), has made remarkable advancements in a wide range of research fields, including science, engineering, technology, medicine, and industry,[1–4] marking a significant milestone in data-driven discovery. Sophisticated algorithms, such as graph convolutional networks (GCNs),[5] convolutional neural networks (CNNs),[6] recurrent neural networks (RNNs),[7] and Generative Adversarial Networks (GANs),[8] are aided by powerful computing resources, such as graphics processing units (GPUs), to achieve success in ML and DL. The main reason behind these achievements is the ability to accurately estimate the behavior in unknown domains by quantitatively learning patterns from a sufficient number of training samples. However, in scientific fields, it is often challenging to obtain large labeled training samples due to various restrictions or limitations such as privacy, security, ethics, high cost, and time constraints. Fields such as computer vision,[9] language translation,[10] speech recognition,[11] and game playing[12] may have large-scale data sets with billions or even trillions of data points, but this is typically not the case in scientific research. For example, in drug discovery,[13,14] the discovery of properties of new molecules to identify useful ones as new drugs is constrained by toxicity, potency, side effect, partition coefficient (log *P*), solubility (log *S*), and various other pharmacokinetics and pharmacodynamics metrics. As a result, there are few records of successful clinical candidates for a given target. When the number of training samples is very small, the ability of ML-based or DL-based models to learn from observed data sharply

decreases, resulting in poor predictive performance. Therefore, it is very important for the scientific community to learn and generalize effectively the data from very few training samples.

Efficiently learning from very few training samples holds great theoretical and practical significance in the fields of ML and DL. First, it can help avoid the prohibitively high cost of acquiring data and performing costly annotations in certain data-intensive applications. Second, it can enable the construction of a low-cost and speedy model for an emerging task that only has a few temporarily available samples, which can illuminate potential laws earlier in the exploration process. Driven by these promising advantages and the practical need for affordable learning, learning from very few training samples has become a popular research topic. However, despite related ML approaches such as small or one-sample learning, zero-shot learning,[15] one-shot,[16] or few-shot learning,[17,18] the research progress on this problem has been slower in the past decade compared to that of large sample learning, due to its intrinsic difficulty. For instance, if a learning algorithm is executed on a task with very few training samples using just vanilla learning techniques without any advanced learning strategies or specific model design, serious overfitting may occur, significantly reducing the predictive power of the model.[19]

Overall, there are several viable strategies to improve the predictive power of ML or DL models when dealing with small scientific data sets. Commonly used strategies include transfer learning,[20,21] combining DL and ML,[22,23] GANs,[24,25] variational autoencoder (VAE),[26,27] self-supervised learning (SSL),[28,29] long short-term memory (LSTM),[30,31] data augmentation based on physical models,[32,33] active learning (AL),[34,35] and semi-supervised learning.[36,37] However, no paper has provided an organized taxonomy linking these techniques. Therefore, in this review, we conduct a survey on ML or DL prediction using small scientific data sets and aim to create a taxonomy that connects these techniques.

The remaining sections of this paper are organized as follows. ML preliminaries are presented in section 2. Section 3 provides a brief overview of several of the main methods used for dealing with data scarcity. Section 3.11 details the theory of transfer learning and its applications in the context of small data sets. Sections 4.1 and 3.7 discuss the methods of combining DL with traditional learning and those based on GANs, respectively. In section 3.8, we outline VAE-based methods for dealing with small training set sizes. Section 3.9 surveys the approach of SSL to small data sets. Sections 3.6 and 3.12 cover LSTM techniques and AL methods, respectively. In sections 3.13 and 4.2, we delineate the Merriman–Bence–Osher method and physical model-based data enlargement, respectively. Section 4 discusses several perspectives for dealing with small data challenges in molecular science, including combining DL with traditional ML, physical model-based data augmentation, natural language processing (NLP), and generative networks. Finally, section 5 offers an outlook on future developments.

2. MACHINE LEARNING PRELIMINARIES

2.1. Supervised, Unsupervised, Semi-Supervised, and Self-Supervised Learning Strategies

In supervised learning,[38,39] a data set containing input and output pairs is used to train a function that maps feature vectors (input) to labels (output). The data set is split into a training set and a test set, with the former used to adjust the model parameters for accurately predicting the outputs for the input examples in the training set. After the model is trained, its generalization ability is evaluated by testing its performance in the test set.

Supervised learning encompasses various types, including classification, regression, naïve Bayes (NB) models, [40] random forest (RF) models,[41] support vector machines (SVM),[42] and neural networks (NNs),[8] among others. These algorithms have found widespread application in biological and chemical fields. For instance, Lazarovits et al.[43] constructed NN models to investigate the mechanism of liver and spleen uptake by nanoparticles, finding that it was due to protein adsorption on their surfaces. Sandfort et al.[44] concatenated 24 fingerprint representations into a 71 375 dimensional vector, which was then used for various supervised learning tasks related to chemical reactivity. Additionally, there is a growing interest in applying supervised learning techniques to predict drug side effects.[45] Munoz et al.[46] used different models, including logistic regression, RF, decision trees, and others, to predict the side effects of biological molecules. Similarly, Zhou et al.[47] applied boosted RF classifiers to predict the side effects of protein targets, therapeutics, transport proteins, enzymes, pathways, and chemical structures of drugs.

In unsupervised learning, the available data set does not have labeled training examples, and the objective is to uncover patterns or relationships in the data.[48,49] One of the most common types of unsupervised learning are clustering, which involves grouping unlabeled data points based on their similarities and differences. This process aims to group data points into clusters in such a way that those in the same group have the highest similarity to each other while points in different groups have little or no similarities. Another type of unsupervised learning is data compression or dimensionality reduction,[50] which aims to represent high-dimensional data in a lower-dimensional space while preserving as much information as possible. This technique can significantly reduce computing or storage costs while making the ML model run much faster. Unsupervised learning has become a crucial tool for handling the increasing amount of data generated by atomic and molecular simulations in biochemistry. Glielmo et al.[51] provided a discussion of the latest algorithms for feature representation of molecular systems used in downscaling and clustering models. Basdogan et al.[52] employed a nonlinear dimensionality reduction algorithm to create a two-dimensional visual representation of the similarity between solute environments in microsolvation clusters of different sizes.

Semi-supervised learning[53] involves a data set that includes both labeled and unlabeled examples. The objective is to learn a function that can predict the labels for the unlabeled samples using the labeled ones. Semi-supervised learning can be used for various tasks, such as classification, regression, clustering, and association, and offers the benefit of reducing expenses on manual annotation and data preparation time. There are several approaches to

semi-supervised learning, including self-training,[54] cotraining,[55] and multiview learning,[56] with the selection of a specific approach depending on the data set's properties and the task requirements. MicroRNAs are noncoding RNAs closely associated with many human diseases in the biomedical field. Ji et al.[57] treated MicroRNAs disease association prediction as a semi-supervised learning problem and proposed a novel method to predict potential MicroRNA-disease associations, a new method for predicting potential MicroRNAs-disease associations. In healthcare, Yin et al.[58] introduced deep forest and semi-supervised self-training to address disease classification and gene selection for different types of diseases. Experimental results demonstrated that the proposed model could achieve good results in both disease classification and causative gene identification.

SSL, also known as predictive or pretext learning, is an ML process where a model trains itself to learn one part of the input from another part of the input.[59] The goal is to learn useful representations from unlabeled data that can help with downstream learning tasks such as classification or object detection. In SSL, the unsupervised problem is transformed into a supervised problem by autogenerating the labels. To make use of the huge quantity of unlabeled data, it is crucial to set the right learning objectives to get supervision from the data itself. Hence, SSL is particularly useful for tasks where it is difficult to obtain labeled data or where the amount of labeled data is limited. Some common SSL tasks include predicting missing elements in data,[60] reconstructing data from corruptions or perturbations,[61] and so on. SSL has recently achieved tremendous success in the fields of biology and chemistry. Wang et al.[62] proposed a cloze-style SSL model, MolCloze, in 2021 to obtain a generic information representation for molecular property prediction tasks. In 2022, Zhang et al.[63] introduced the concept of SSL to develop HelixADMET, a robust and end point scalable absorption, distribution, metabolism, excretion, and toxicity (ADMET) system. HelixADMET generated a pretrained model to efficiently screen out unwanted drug candidates in the early stages of drug discovery.

2.2. Regression, Classification, Clustering, and Dimensionality Reduction Tasks

Regression ML is used to understand the relationship between dependent and independent variables and commonly predicts a continuous value based on the input variables. The main goal of regression problems is to estimate a mapping function based on the input and output variables. There are various types of regression algorithms that are widely used in the biochemical domain, including linear regression,[64] decision tree regression,[65] principal components regression,[66] RF regression,[67] support vector regression,[68] and polynomial regression.[69] For example, in computer-aided drug design, multiple linear regression models are often used for pattern recognition, structural similarity, and binding energy prediction to screen promising drug candidates for COVID-19 therapy and quantitative structure–activity relationships when assessing the structural stability and densification of drugs in complex with the major protease of SARS-CoV-2.[70] Yan et al.[71] used multiple linear regression model and SVM methods to predict the inhibitory activity of 117 Aurora-A kinase inhibitors, respectively. Additionally, Ye et al.[72] applied the established molecular docking-based SVM regression model to the design of new NF-B-inducing kinase inhibitors.

Classification predictive modeling involves approximation of a mapping function from input variables to discrete output variables, which are typically labeled or categorized. The mapping function is used to predict the class or category for a given observation. In many cases, classification algorithms predict a continuous value as the probability of an example belonging to each output class. These probabilities can be interpreted as the likelihood or confidence of an example belonging to a particular class. Common classification algorithms include linear classifiers, SVMs, decision tree classification, *k*-nearest neighbor (KNN), and RF classification, among others.[40,73,74] There are numerous applications of these methods in biology and chemistry. For instance, Arian et al.[75] utilized the KNN algorithm to distinguish between active and inactive protein kinase inhibitors and evaluated the performance of the model using SVM and NB classification methods.

Clustering or cluster analysis[76,77] is the task of grouping a set of objects into homogeneous groups or clusters while ensuring that objects in different groups are dissimilar. Clustering can be considered an unsupervised task, as it aims to describe the hidden structure of the objects. Each object is described by a set of features. The key step in dividing objects into clusters is to define the similarity or distance between the different objects. There are many clustering algorithms, including hierarchical clustering,[78] centroid-based clustering,[79] distribution-based clustering,[80] density-based clustering,[81] and grid-based clustering.[82] In order to improve the classification of primary breast cancer and identify disease subgroups relevant to patient management, Ferro et al.[83,84] used four different clustering methods. Their findings showed that applying unsupervised learning to primary breast cancer data was a promising approach to enhance the classification of primary breast cancer and define subclasses of treated patients.

Dimensionality reduction[85] is the process of reducing the number of random variables in a data set while retaining as much relevant information as possible. The goal of dimensionality reduction is to transform high-dimensional data into data of lower dimensions, making them easier to analyze, visualize, and understand. Dimensionality reduction is commonly used as a preprocessing step before supervised learning and to remove noise in the data. There are several common dimensionality reduction methods, including principal component analysis,[86] factor analysis,[87] t-distributed stochastic neighbor embedding (t-SNE),[88] the uniform manifold approximation and projection,[89] and residue-similarity scores,[90] among others. Several applications of biochemistry have utilized ML techniques. For example, to investigate the structure and binding interactions of HIV-1 protease and P2 ligands, Karnati et al.[91] performed principal component analysis to identify differences in conformational changes induced by inhibitor binding. In 2021, Bort et al.[92] used t-SNE to explore the structure of bioactive organic molecule data sets.

3. METHODS FOR SMALL MOLECULAR DATA CHALLENGES

3.1. Basic Machine Learning Algorithms

Since Arthur Samuel proposed the concept of ML in 1956,[93] a large number of algorithms have been developed. Some of the most traditional algorithms include KNN, which was proposed by COVER in 1968,[94] SVM, which was introduced in 1995 by Cortes,[95] and RF which was also proposed in 1995 by Ho.[96] These classic and fundamental algorithms have

been widely applied in various fields such as data mining,[97–99] statistical learning,[100,101] and computer vision.[102,103] ML is becoming increasingly popular in the fields of biological medicine and chemistry, particularly in healthcare and COVID-19 research.[104–106] This is due to its potential to aid in disease diagnosis,[107] data pattern detection,[108,109] patient management,[110] and other areas. In this context, basic ML algorithms will be introduced in the field of small molecules. Special algorithms like CNNs and artificial neural networks (ANN) will be discussed in detail later.

In the field of drug–target interaction, traditional chemical experiments can be both expensive and time-consuming. Although many methods based on different principles have been developed to measure the similarity of drugs or targets, their results are often unsatisfactory. In 2018, Chen et al. proposed a new method for identifying drug–target interaction using the Gradient Boosted Decision Trees (GBDT) ML algorithm.[40] This method combined drug and protein identifiers, descriptors, and negative information to predict drug–target interactions. The data set used in the experiment consisted of 4950 drugs and 2313 human protein interactions, with 609 drug characteristics and 1819 protein characteristics. The GBDT algorithm was compared to six other methods in the experiment, and the results are shown in Table 1. The experimental results indicated that the GBDT algorithm outperformed other advanced methods, particularly when the data set was small.

Drug-induced toxicity is a significant side effect that requires consideration during drug development. However, current experimental methods used to evaluate drug-induced toxicity are often time-consuming and expensive, which make them unsuitable for large-scale assessments during the early stages of drug discovery. In 2014, Zhou et al. proposed a computational prediction model of drug-induced toxicity based on SVM.[111] The study included 572 samples from a small toxicity data set, and to compare the performance of the proposed model, the researchers applied NB[112] and recursive partitioning[113] methods to the same data set. Among all the prediction models, drug-induced toxicity based on SVM achieved the best performance, with prediction accuracies of 85.33% and 83.05% for the two independent test sets, respectively.[111] In comparison, the Bayesian model yielded prediction accuracies of 76.09% and 74.58% in the two independent test sets, while the recursive partitioning model resulted in prediction accuracies of 79.89% and 77.97% in the same two independent test sets. Based on these experimental results, the drug-induced toxicity based on the SVM model outperformed the other two models.

There is growing interest in the application of ML and DL across the life sciences, including drug discovery. In 2022, Siemers et al. identified the minimal data requirements for learning with activity-based composite classification, which serves as an example application.[114] ML binary classification models were constructed using increasingly larger training sets, starting from a minimal set that included only one active compound (and two inactive compounds) and extending up to a training set containing 600 active compounds (and 1200 inactive compounds), which was used for model construction. In chemical informatics, message-passing neural networks are increasingly used for the deep representational learning of molecular diagrams and the prediction of molecular properties. As a control, a simple KNN classifier was used. To explore the sustained high performance of the KNN classifier, the researchers systematically extracted simulated sequences from 20 randomly selected activity

classes using the composite core relation algorithm. The calculation protocol is shown in Figure 1. The results of the control calculations showed an increasing size of the training set consisting of the same number of active and inactive instances.[114] This work has the potential to impact the prospective use of prediction models, particularly for emerging targets where there are typically limited compound data.

MicroRNAs play a crucial role in many pathological processes by inhibiting translation through interaction with specific target mRNAs. These miRNAs can act as either oncogenes or tumor suppressors. In 2022, Yu et al. used a variety of ML algorithms to build models predicting up–down and up–down pairs on the training parts of data set1 and data set2, respectively.[115] The flowchart of this study is illustrated in Figure 2. A model was constructed using data set1 (2096 positive pairs and 2096 manually constructed negative pairs) to predict upregulated pairs of small molecules and MicroRNAs. Similarly, data set2, with 1591 positive and 1591 negative pairs, was used to build a model for predicting down-regulated pairs. The RF algorithm showed the best performance. On the test data set, the maximum area under the curve (AUC) value of the up-regulated model was 0.911, and that of the down-regulated model was 0.896. Additionally, the accuracy values of the down-regulated and upregulated models on independent verification pairs were 0.91 and 0.90, respectively.[115] This study is expected to have implications in identifying potential therapeutic targets for the development of antitumor drugs.

As ML continues to evolve, its various methods have future directions in various fields. Currently, probabilistic graphical models, neural networks, and other methods based on probability are research hotspots, in addition to the NB algorithm.[116] NB models have stable classification efficiency and perform well on small-scale data, handle multiclassification tasks, and are suitable for incremental training. However, it can lead to poor predictions due to the assumed prior model or classification decision errors. Similarly, SVM has developed rapidly in finite dimensions, but further research is needed in infinite dimensions,[117] and its robustness[118] also needs improvement. RF is capable of efficiently operating on large data sets[119] but sometimes results in a large number of decision trees, which enlarges the space and time required for training.[120] Fortunately, RF will play a significant role in this era of increasing data volume. Later, we will discuss some special ML algorithms that are suitable for small sample applications.

3.2. Artificial Neural Networks

With the continuous development of AI, significant progress has been made in the field of AI. However, in some complex research areas, AI cannot completely replace the human brain in solving complex problems. With increasing exploration by researchers, ANN has proven to be effective in replacing the human brain to solve difficult problems. In 1943, McCulloch et al.[121] proposed the first ANN computational model, called the M-P model, which promoted the development of ANN research. ANN, which is also known as a collection of connected units of artificial neurons, is a framework for many different algorithms from ML. A basic ANN structure usually contains three parts: an input layer, a hidden layer, and an output layer. Due to their advantages of self-learning and powerful computing power, ANNs have been widely used in various fields, such

as face recognition,[122–124] medical diagnosis,[125–127] and speech recognition.[128,129] The generalization ability of neural networks is mainly dependent on the size of the training set and the network architecture. Generally, the performance of the neural network improves as the number of samples in the data set increases. In recent years, ANNs have played a vital role in small data set research in the fields of biology and chemistry.[130] Researchers have applied neural networks to obtain the optimal experimental results. In the following section, we will summarize and analyze the research on ANNs in small data sets.

In the field of drug design, researchers have used ANNs to develop predictors for log *P*. For instance, Chen et al.[131] developed a predictor for log *P* using a fully connected ANN model. This study is expected to contribute to the identification of potential therapeutic targets for antitumor drug development. This, in turn, can help reduce time, effort, costs, and attrition rates in drug discovery by enabling the rejection or prioritization of compounds without the need for synthesis and testing. While, in order to explore anticancer properties of thioguanine, Hoseini et al.[132] applied an ANN approach to generate quantitative structure–property relationships models for log *P* prediction. Additionally, Dadfar et al.[133] developed genetic algorithm–multiple linear regressions (GA-MLR) and genetic algorithm–artificial neural network (GA-ANN) models to predict the log *P* of sulfonamides. Sulfonamides are compounds with a wide range of biological activities and serve as the basis for several groups of drugs.

In clinical practice, predicting blood-to-plasma concentration ratios is crucial for determining drug administration regimens. However, only a few studies have investigated methods for predicting concentration ratios. In 2021, Mamada et al. developed an concentration ratios prediction model incorporating typical human pharmacokinetics parameters.[134] They compiled experimental concentration ratio values for 289 compounds, providing reliable predictions by extending the range of application. The authors used human pharmacokinetics parameters, including the volume of distribution, clearance, mean residence time, and plasma protein binding rate calculated from plasma drug concentration and 2702 molecular descriptors to construct a quantitative structure–pharmacokinetics relationship model for concentration ratios. Among the algorithms analyzed, the ANN algorithm had the best performance. After optimizing with six molecular descriptors and log Vd, the correlation coefficient of the model is 0.64, and the root-meansquare error (RMSE) is 0.205, which is better than other concentration ratio prediction methods reported in the past.

In 2022, Mayer et al. investigated the nucleation of dislocations in homogeneous lattices, which is related to small-scale plasticity or ultrafast loading.[135] They prepared training data using molecular dynamics (MD) simulations and performed polynomial extrapolation beyond the nucleation limit to improve the accuracy of the trained ANN and make theoretical predictions more accurate. The authors considered atomic configurations observed during dislocation nucleation and subsequent development and presented an approximation method that required smaller and simpler MD data for training. Their method gave a strain rate dependence for the nucleation threshold that was close to that of a rigorous theory of dislocation nucleation. A schematic diagram is shown in Figure 3.

The performance of small molecule receptors in organic solar cells is determined by their chemical structure. To avoid trial-and-error based design, multiscale simulation is necessary, which can ultimately save time and resources. In 2022, Mahmood et al.[136] collected data on 164 small molecule nonfullerene acceptors from the literature. Computational analysis is a quick and efficient way to narrow down potential candidates for synthesis. This work shows that properly regulating sp^2-hybrid nitrogen substitution is an effective way to tune the properties of the electron acceptors. This study also demonstrates the potential of multiscale theoretical modeling, which makes it possible to envision structural changes from atomic to molecular levels.

In the future, ANNs will be more frequently applied to neurobiology,[137] enabling researchers to derive testable insights and predictions from neurobiological experiments. ANNs have already been integrated with other advanced methods, such as fuzzy logic[138] and wavelet analysis,[139] to enhance their ability for data interpretation and modeling, as well as to avoid subjectivity in the operation of the training algorithm.[140] ANNs are expected to show their talents in more fields in the future, mainly due to their powerful data processing capabilities. They can outperform almost all other ML algorithms in some cases, such as cancer detection, which is a demanding task,[141] where better performance can lead to more people being treated. However, one of the disadvantages of ANNs is data gluttony, as they generally require more data than traditional ML algorithms. Additionally, ANNs have other downsides such as the classic black box problem,[142] as well as being time-consuming and labor-intensive in their training.

3.3. Convolutional Neural Networks

CNN is a valuable tool in the analysis of biological data[143] and is a type of DL algorithm inspired by the natural visual perception mechanism in biology.[144] LeCun et al. proposed LeNet-5 in 1998 for standard handwritten character recognition.[145] The network structure is relatively complete, which is one of the fundamental components of modern CNNs, making LeNet-5 the beginning of the class of CNNs. Over time, Krizhevsky et al. came up with AlexNet,[146] which performed exceptionally well in image classification. Later, VGG-Net[147] and GoogLeNet[148] were created in the same year and achieved remarkable performance in the ImageNet classification task. It is worth mentioning that ResNet[9] made a significant innovation in the network structure and pioneering work in computer vision and DL. In addition, DenseNet[149] was proposed, a CNN with dense connections, which further improved the network's performance. Next, we introduce the CNN structure. CNNs are a collection of neurons organized in interconnected layers, with convolutional, pooling, and fully connected layers.[143] The convolution layer is used to extract local features, the pooling layer is responsible for significantly reducing parameter size, and the fully connected layer is used to output the desired results, similar to the traditional neural network part.[150] In fact, CNNs have been applied in various fields and have yielded remarkable results, such as image processing,[151–154] action classification,[155–157] NLP,[158–161] physics,[162–164] and more. CNNs are popular in the fields of biology and chemistry for studying quantitative conformational relationships (QSAR). For example, Hu et al.[165] proposed an end-to-end encoder-decoder model and CNN architecture for QSAR prediction, and Karpov et al.[166] constructed a transformer–CNN framework for generating higher quality, interpretable

QSAR models. Additionally, Hamza et al.[167] used a CNN model for bioactivity prediction. We next focus on how CNNs excel in predicting biochemical molecules on small data sets.

Drug-induced liver injury poses a significant challenge in drug development and postmarketing safety monitoring, as it can cause clinical trial failures and drug withdrawals. Traditional safety testing methods are inadequate to address this pharmacological problem due to their limited predictive capabilities. In 2020, Nguyen-Vo et al.[168] proposed a novel NLP-inspired computational framework using CNN and molecular fingerprint embedding features to address this issue. The construction of their model is illustrated in Figure 4. Their development set included 1597 samples, consisting of 946 DILI compounds and 651 non-DILI compounds, while their independent test set included 322 samples, including 128 DILI compounds and 194 non-DILI compounds. The study achieved an average accuracy of 0.89, a Matthews correlation coefficient (MCC) of 0.80, and an AUC of 0.96. The results indicate that the proposed model significantly outperformed the latest and best model with a 6.67% improvement in AUC from 0.90 to 0.96. Additionally, the findings suggest that molecular fingerprint embedding features are an effective method for molecular representation in biological research, complementing traditional molecular fingerprinting applications.

Quantitative structure–activity relationships (QSARs) play important roles in the environmental field. In 2021, Zhong et al.[169] used molecular images combined with CNN to develop QSARs to predict the rate constants of hydroxyl radical generation from compounds. The data set contained 1159 organic compounds, which were initially classified into 357 classes based on all functional groups. However, 250 of the 357 classes contained fewer than three compounds and could not be divided into the training, validation, and test data sets. Therefore, based on functional group similarity, they merged classes with less than 3–4 compounds with larger groups to form 98 classes.[170] The study developed molecular image-CNN models using transfer learning and data augmentation techniques. These techniques greatly improve the robustness of the model and prediction performance. Experimental results show that the proposed model has a better prediction performance than the model based on molecular fingerprints.

MD simulations are effective in analyzing the transport characteristics of liquids on solid surfaces with different nanometer-scale roughness, but they require high computational costs.[171] In 2022, Li et al. proposed a DL encoder-decoder CNN to predict the adsorption density distribution of atoms and organic liquids at various molecular-scale surface roughnesses.[172] The data set consisted of monatomic liquids (sample size: 384) and polyatomic liquids (sample size: 384), with 344 samples in the training set and 40 samples in the test set for both single atoms and multiple atoms. The CNN structure and parameter settings are shown in Figure 5. The proposed method achieved high accuracy in predicting adsorption densities at different microinterfaces with a small data set. The experimental results show that MD and DL methods have good coupling, which can help in designing surface geometry to obtain ideal molecular liquid interface transport characteristics and complement the nanoscale model system for interactive visualization.

In summary, CNN is a great approach for solving small data set problems in various fields, such as pharmacology,[168,173,174] chemical efficacy testing,[175] and protein structure

prediction.[143,176] One of the most significant advantages of CNN[177–179] is its ability to identify important features without human supervision. Additionally, CNN is highly accurate at image recognition and classification.[180] Another major advantage of CNN is its weight sharing property, which reduces the amount of computation required compared with regular neural networks. However, CNN also has some drawbacks.[181] First, it fails to encode the position and orientation of objects.[182] Second, CNN's effectiveness depends on having a large amount of training data. Third, CNN tends to be slower due to operations such as maxpool.[183] Finally, because CNN is made up of multiple layers, the training process can take a long time if the computer lacks a powerful GPU. Despite these drawbacks, CNN has shown promising results when applied to small data sets of biochemical molecules. Nevertheless, researchers need to consider issues such as efficiency, experimental costs, and result generation. After years of research and application, CNN has become one of the representative DL algorithms, reflecting its powerful functions in many aspects. In the future, CNN will have more applications in the field of small molecules and will play a greater role. Thus, further research and exploration are necessary to improve CNN.

3.4. U-Net

Semantic segmentation has been successfully used in various fields, including geological detection, automatic driving,[184,185] and agriculture.[186] It is a fundamental task in computer vision, and its first model, a fully convolutional network, was proposed by[187] in 2015. Since then, several other models have been introduced, such as the U-Net model,[188] SegNet,[189] dilated convolutions,[190] and Deeplab.[191] In particular, we focus on U-Net and its application to small data sets in the biochemical molecular field. U-Net belongs to the Encoder–Decoder structure,[192] and its framework is shown in Figure 6c.[193] The original intention of U-Net was to solve problems in biomedical images.[194] Because of its excellent performance, U-Net has been widely used in the fields of biology and chemistry, such as drug and material design,[195] protein structure prediction,[196,197] as well as other topics such as satellite image segmentation[198] and industrial defect detection.[199]

In the molecular field, the U-Net model has shown promising results for small data sets. In 2021, Nazem et al.[200] developed a 3D U-Net model based on voxels for predicting binding sites in protein structures. The algorithm was trained and validated on a subset of scPDB, which is the largest and highest quality binding site database selected from the PDB. To test the model's performance, the authors used three data sets: Chen11, B210, and DT198. Chen11 contains 251 structures with the maximum number of relevant pockets; B210 is a set of 210 protein structures in the bound state from LIGSITE-csc, and DT198 contains 198 drug target complex structures. The model was also assessed on the B48/U48 database to show its performance on the apo structures of proteins. The evaluation metric used was the F1-score, and the F1-scores for the three data sets were: B210 (0.41), DT198 (0.40), and Chen11 (0.37). All of these scores were higher than those achieved by the LIGSITE-csc and DeepSite methods.

In 2021, Kotowski et al.[201] proposed a single-sequence-based protein prediction method, called ProteinUnet, which leveraged the U-Net convolutional network architecture. The article aimed to predict protein function and structure from sequence and used protein data

sets from CullPDB and named them TR9993 and TS1199. Specifically, TR9993 consists of 9993 different chains from 9622 proteins as the training set, and the test set TS1199 consists of 1199 chains from 1187 different proteins. The authors concluded that their model had better classification accuracy compared to the SPIDER3-single model, and more detailed results are shown in Table 2. This table provides the mean accuracies of Q3 and Q8 predictions at the sequence level in TS1197 and CASP13, along with the standard deviations and *p*-values of the two-sided Wilcoxon signed-rank test between the models.

In 2021, Prasad et al.[202] developed an automatic liver parenchyma segmentation network based on the U-Net architecture. The authors used a data set consisting of highly variable venous phase enhanced computed tomography (CT) volumes, with 10 males and 10 females as the source, 75% of whom had liver tumors. However, due to the small size of the data set, the model was overfitting, and the authors had to take some measures, such as reducing the convolution and dropout layers. They also added Gaussian noise to prevent overfitting and solved the problem of inconsistent intensity by pixel normalization. To build a model with better performance, it was important to choose an appropriate loss function. The authors evaluated four loss functions: Dice loss, binary cross-entropy loss, Tversky loss, and focal Tversky loss, and we found that the Dice loss function performed the best, achieving a score of 94.5%. Their work may play a crucial role in assisting oncologists and surgeons with accurate analysis of various pathological conditions, ultimately saving time.

As an improved version of the fully convolutional network model, U-Net has several characteristics that make it suitable for large medical image segmentation.[203] These include multiscale capability,[204] simple structure,[205] and the use of skip links.[206] However, U-Net also has some drawbacks, such as slow running efficiency[207] and the limitation of being able to predict on a single scale.[208] In the future, supervised,[38,39] semi-supervised,[53,209] and unsupervised learning[48,49] could be potential areas of research for U-Net, as medical image data often lacks sufficient labeled examples. Additionally, the combination of U-Net and AL[210,211] could also be a promising direction for addressing the challenge of data labeling.

3.5. Graph Neural Networks

In recent years, graph neural networks (GNN) have become powerful and practical tools for ML tasks in graph domains. The GNN model was first introduced by Gori et al.[212] and Scarselli et al.[213] and Micheli et al.[214] developed and improved upon the algorithm. The success of GNN in many domains such as recommender systems,[215,216] computer vision,[217] and NLP[218,219] is attributed in part to its effectiveness in extracting latent representations from Euclidean data. However, as data are increasingly represented in the form of graphs, including non-Euclidean domains such as e-commerce,[220,221] chemistry,[222,223] and citation networks,[224,225] there is a growing need for GNNs. Additionally, molecular property prediction is a popular application of GNNs, as molecules can be represented as topological graphs, with atoms as nodes and bonds as edges. Currently, the most advanced GNNs can be categorized as GCNs,[226] graph autoencoders,[227] recurrent GNNs, and spatial-temporal GNNs.[228] It is worth noting that there are still open questions about how GNNs handle small data on molecular science or small molecular data that need to be addressed.

In 2021, Yaqing Wang et al.[229] proposed property-aware relation networks which are compatible with existing graph-based molecular encoders to address the limitations of quantitative structure–property relationships and the issue of existing works failing to leverage related graphs among molecules. The overall architecture of property-aware relation networks is shown in Figure 7. The authors conducted experiments on widely used benchmark few-shot molecular property prediction data sets from MoleculeNet:[230] Tox21, SIDER, MUV, and ToxCast, which consist of 8014, 1427, 93127, and 8615 molecules, respectively. The article used the ROC-AUC metric to evaluate the model performance on these benchmark molecular property prediction data sets. Empirical results consistently showed that property-aware relation networks achieved state-of-the-art performance on the few-shot molecular property prediction problem.

In 2020, Pappu and colleagues[231] investigated the use of pretraining and the meta-learning technique MAML (as well as variants FO-MAML and ANIL) to enhance the performance of GNNs via transfer learning from related tasks, allowing for their use even in settings with limited data availability. The authors created a new data set comprising 645 binary classification tasks from the ChEMBL database, filtered for five distinct task types. The study found that the performance of the GNN model was initially lower than fingerprint methods but significantly improved with the use of MAML and FO-MAML, outperforming both fingerprint and pretraining methods as measured by the area under the precision-recall curve of the models. The results suggested that meta-learning can improve the use of GNNs in low-data settings compared to fingerprint methods.

Generally, existing DL methods for molecular property prediction require large training data sets for each property, which limits their performance in cases where there is only a limited amount of experimental data, especially for new molecular properties. To address this issue, Zhichun Guo et al. proposed Meta-MGNN, a novel model for few-shot molecular property prediction in 2021.[232] Meta-MGNN's skeleton framework can be seen in Figure 8. To evaluate the performance of Meta-MGNN, the authors used the Tox21 and Sider data sets, which consist of 7831 and 1427 samples, respectively. The overall performance of all methods was evaluated using the AUC metric, and the results showed that Meta-MGNN outperformed all baseline models on both the Tox21 and Sider data sets. Specifically, for 1-shot learning, the average improvements were +1.04% and +1.80% on the Tox21 and Sider data sets, respectively, and +0.84% and +1.87% for 5-shot learning.

GNNs have become widely used not only in the fields of biology and chemistry, such as protein–protein interaction networks,[233] protein structure prediction,[234] and chemical property estimation[235] but also in various other ML applications such as reinforcement learning,[236,237] semi-supervised,[238,239] and unsupervised[240,241] learning. However, due to the complexity of the graph structures, GNN models are not always effective in all graph conditions. To address this issue, several future research directions have been proposed. For instance, the robustness of GNN models should be enhanced because they are vulnerable to adversarial attacks.[242] Moreover, because GNN models are often treated as black-boxes, there is a need for improved interpretability on graphs. Thus, research in generating example-level explanations for GNN models has been proposed.[243,244] Additionally, graph pretraining[245] and the challenges associated with complex graph structures are also

important research directions. Nonetheless, GNNs also have certain limitations such as their performance being limited by their depth and width,[246] their inability to work with insufficient data,[247] and issues related to high computational costs.

3.6. Long Short-Term Memory

LSTM is a type of RNN that addresses the vanishing gradient problem during training and is capable of learning long-term dependencies. It was introduced by Hochreiter et al.[248] in 1997 and has been refined and applied in various fields. Unlike other DL models, the LSTM is specifically designed to handle long-term information without incurring significant cost. LSTM employs back-propagation as its main parameter training algorithm, which involves four steps: forward pass to calculate the output values, error computation using a loss function, backward error propagation among the neurons, and weight parameter updates. LSTM has been successful in a variety of biological and chemical fields, including chemical–protein relation extraction,[249] chemical substance classification,[250] and drug molecular design.[251] Additionally, it has been used in fields like imaging,[252,253] speech recognition,[254,255] NLP,[256] and more. However, when the data set is small or has few labeled samples, using LSTM may not always yield desired results. This is particularly relevant in bioinformatics, biochemistry,[257] and other fields where data sets typically have fewer than 5000 elements.

In the field of protein research, predicting the structure of proteins is essential for understanding their function and designing drugs. Traditional techniques for protein structure prediction are often time-consuming and expensive, and developing new advanced methods remains a major challenge. The secondary structure of proteins is critical for analyzing protein function and designing drugs. Various computational methods have been proposed to improve the performance of protein secondary structure prediction. In 2019, Guo et al.[258] proposed a novel deep neural network approach called deep asymmetric convolutional LSTM neural network (DeepACLSTM) for predicting protein secondary structure from protein sequence features and profile features. DeepACLSTM utilized the eigenvector dimension of the protein feature matrix to effectively combine asymmetric CNNs[259] and bidirectional LSTM (BLSTM) neural networks to predict protein secondary structure. It comprised three main modules. In this paper, DeepACLSTM was compared with several methods such as SSpro8,[260] conditional neural field (CNF), DeepCNF (CNF based on DL),[261] and CBRNN.[262] To evaluate the performance of DeepACLSTM, experiments were conducted on three publicly available data sets: CB513, CASP10, and CASP11. The results demonstrated that DeepACLSTM outperformed the state-of-the-art baseline on all three data sets.

In the field of medicinal science, cancer remains a significant threat to human health. Anticancer peptides (ACPs) present a promising avenue for cancer treatment and offer many advantages. However, traditional experimental methods for identifying novel anticancer peptides can be costly and inefficient. In 2019, Yi et al.[263] proposed a DL-LSTM neural network model, ACP-DL, for effectively identifying new anticancer peptides. The authors combined binary contour features and a *k*-mer sparse matrix of simplified amino acid letter features to construct an efficient feature representation that maximized the use of peptide

sequence information. Additionally, a deep LSTM model was utilized to automatically learn to discriminate between anticancer and ordinary peptides. The workflow of this approach is shown in Figure 9. To evaluate the performance of the method, the authors used the ACP740 (with a sample size of 740) and ACP240 (with a sample size of 240) data sets and compared the results using 5-fold cross-validation, which verified the state-of-the-art performance of ACP-DL.

In 2017, Li et al.[264] proposed a predictive model called ProDec-BLSTM for investigating protein remote homology detection. The model included an input layer, a BLSTM layer, a time-distributed dense layer, and an output layer. The framework diagram for ProDec-BLSTM is shown in Figure. 10. The performance of the model was evaluated using the SCOP data set, which had a sample size of 4019. ProDec-BLSTM was compared with GPkernel,[265] GPextended,[265] GPboost,[265] SVM-Pairwise,[266] Mismatch, eMOTIF,[267] LA-kernel,[268] PSI-BLAST,[269] and LSTM[270] on the same data set. ProDec-BLSTM achieved a mean receiver operating characteristic curve (ROC) of 0.969 on the evaluation metric, which was higher than those of the other methods.

In addition to the previously mentioned applications of LSTM, successful applications have been documented in the fields of biophysics and bioinformatics. Various variant models have also been proposed by combining LSTM techniques to improve accuracy, such as the flow-based LSTM model proposed by Gers et al.[271] in 2000. Recently, Zhu et al.[272] proposed a variant model called ACP-check, which utilized BLSTM networks and multifeature fusion. The model extracted time-dependent information features from peptide sequences by using a BLSTM network and combined them with amino acid sequence features. To validate the performance of the model, six benchmark data sets were selected, including ACPred-Fuse, ACPred-FL, ACP240, ACP740, main, and alternate data sets of AntiCP2.0. ACP-check achieved prediction accuracies of 0.91, 0.91, 0.90, 0.87, 0.78, and 0.93, respectively, with improvements ranging from 1% to 49%. These results demonstrated the excellent predictive performance of ACP-check. Other improved models, such as Bidirectional LSTM,[273] have also been proposed for short-term load forecasting. These successful examples of LSTM-based variant models indicate that the use of LSTM techniques is not only effective in improving experimental results but also has a wide range of applications. However, further research is needed to determine the optimal method for combining LSTM with small data sets.

3.7. Generative Adversarial Networks

Although DL has made significant breakthroughs in various research fields, the quality and quantity of data often affect its results. In 2014, Goodfellow et al. proposed an innovative GAN model.[274] Unlike other DL algorithms, GAN has a discriminant model composed of two main parts: the generator and the discriminator. The generator is responsible for creating synthetic data samples, while the discriminator tries to differentiate between real and synthetic samples. These two parts compete against each other during the training phase, where the generative model learns the distribution of the sample data. Note that the discriminative model is often a dichotomous classifier used to distinguish between real and generated data. The flowchart of the GAN framework is shown in Figure 11.

With the rapid development of GAN, many generalizations have been proposed to improve the original method. Examples of these include AdaGAN,[275] MADGAN,[276] PacGAN,[277] D2GAN,[278] SGAN,[279] and DCGAN.[280] GAN and its extensions are often used to augment data sets and mitigate overfitting issues in downstream ML and DL tasks. In the following sections, we summarize how GANs have been utilized in small data sets of chemical and biological molecules.

In 2019, Han et al.[281] proposed protein log *S* generative adversarial nets (ProGAN), a data augmentation method to address the issue of insufficient data for protein log *S* prediction. The data set comprised 3148 samples from the ESOL database,[282] and the evaluation metric used was the coefficient of determination, R^2. ProGAN was employed solely for data augmentation and combined with the DNN method to enhance the prediction performance. The optimal results were achieved using the sigmoid activation function.[283] The test set R^2 for DNN was 0.40 ± 0.0074, and for DNN+ProGAN, it was 0.42 ± 0.0067. The DNN+ProGAN method with the sigmoid activation function yielded the highest experimental result, with an R^2 of 0.45 ± 0.0018. The results of this work have the potential to enhance the production yield of recombinant proteins in biocatalysis applications.

In 2019, Liu et al.[284] developed a model that combined GAN and deep neural networks (DNN) for multiple classifications with small cancer-staging sample sizes. First, the original data were split into a training set and a test set, and the GAN was trained using the training set to generate synthetic samples that expanded the training set. Then, the DNN classifier was trained using the synthetic samples, and the classifier was tested with the test set using different metrics to verify the effectiveness of the method. The data set used in the experiment had less than 100 samples, which were divided into a training set (60%) and a test set (40%). Classical ML methods such as RF[285] and NB were used as a comparison, and the SMOTE[286] method was used to generate oversampled samples to train the classifiers. In the WGAN-based framework, a large number of synthetic samples generated by WGAN were used to train the classifiers, and then the classifiers were validated with real samples. The experimental results were presented in Table 3. The evaluation metrics used were accuracy, *F*-measure (the harmonic mean of precision and recall),[287] and the geometric mean of recall,[288] which demonstrated that the proposed method substantially improved the results of the classification experiments under the condition of increasing the number of synthetic samples.

In the field of cancer research, the issue of insufficient data often leads to poor performance of ML models. To address this problem, Wei et al. proposed Gene-GAN in 2022,[74] a model for classifying cancer data. As the data set contained less than 500 samples, they used GAN to augment the data and employed the reconstruction loss to stabilize model training, resulting in high-quality generated samples. The excellent performance of Gene-GAN was demonstrated by comparing it with different classifiers in Table 4, which also highlighted the importance of data augmentation using GAN. In the table, Gene-GAN (mixed) indicated that the generated data was used in combination with the original data, while Gene-GAN (nonamplified) meant that the augmented data was not used. The experimental results confirm that the generative model is an effective solution to the problem of insufficient sample size.

In 2020, Hsu et al.[289] proposed the Wasserstein-based data augmentation algorithm, which utilized GANs to augment data during model training. The authors conducted experiments on a breast cancer data set containing 582 samples. The results of Wasserstein-based data augmentation demonstrated higher accuracy, AUC, and concordance index values compared to those of the data augmentation algorithm. Specifically, the accuracy value was 0.6726 ± 0.0278, the AUC value was 0.7538 ± 0.0328, and the concordance index value was 0.6507 ± 0.0248. The results suggest that GANs can be effectively used to train deep models in medical applications, even when limited data is available.

In 2021, Li et al.[290] presented a derivative model, BrainNetGAN, based on GAN for the synthesis of conditional brain networks. The brain network matrix was used as input to generate a fake brain network connection matrix through BrainNetGAN, and then, the potential distribution and topological characteristics of real brain network data were inferred. The experiments evaluated the data augmentation performance of BrainNetGAN and compared its results with the experimental results of Baseline without augmented data. Specifically, BrainNetGAN attained an accuracy of 0.812, which was higher than the baseline of 0.791.

The experiment conducted by Lin et al.[291] in 2021 aimed to develop a sequence-based binary classifier to determine whether short peptides exhibited antiviral activity. The antiviral data set used in the experiment consisted of 2934 samples. To address the issue of imbalanced data, the authors employed a GAN model to augment the number of positive data samples, which were then added to the original data set. As a result, the model achieved an accuracy of 84% in the final prediction, which outperformed the accuracy achieved using the original data set without augmenting the data generated by the GAN method.

Nowadays, GAN has gained popularity in both academia and industry due to its numerous applications, not only in the fields of peptide and protein design,[292] chemical material design,[293,294] and medicine,[295–297] but also in image genersion, among others. These studies demonstrate the broad range of applications of the GAN methods. Despite their significant success, GANs still have shortcomings in various research fields. For example, the interpretability and controllability of GANs have not been fully understood, and further research on these aspects will remain crucial in the future. Additionally, GANs often suffer from poor stability, which can lead to model collapse.[307] Therefore, future research on how to prevent model collapse during GANs training will be important.

3.8. Autoencoders

In recent years, NLP models have become increasingly popular. Among them, (variational) autoencoder (VAE) is considered to be one of the most promising techniques for unsupervised learning. VAE was proposed by Kingma et al.[308] in 2013. It not only plays an important role in generating data but also has a wide range of applications in imaging and other fields. With the continuous development of VAE, its structure has become more flexible, and derivative models based on variational autoencoder-based models have emerged. For example, the conditional variational autoencoder was proposed by Makhzani et al.[309] in 2015. In 2017, Bao et al.[310] proposed a model conditional variational autoencoder-GAN combining VAE and GAN for synthesizing images in fine-

grained categories, such as faces of a specific person or objects in a category. Other examples include a variational loss autoencoder,[311] multistage variational autoencoder,[312] and Wasserstein autoencoder.[313]

In the field of molecular generation, various generative models based on variational autoencoders have been proposed, such as Graph Flow-Variational Autoencoder (GF-VAE),[314] which combines VAE and the normalized model to generate molecular maps at once. Through the use of variant models of the variational autoencoder, it is evident that VAE has been extensively applied in many research fields. While traditional autoencoders describe the difference of the latent space using numerical methods, VAE models the difference of the latent space using probabilistic distributions. It models the relationship between latent variables and input data from a probabilistic perspective to complete the task of data generation and solve the problem with very few training samples. The model structure of VAE is mainly composed of two parts: the inference network (i.e., encoder) and the generation network (i.e., decoder). The basic process is to map the samples to the latent variables of the low-dimensional space through the encoding process and then restore the hidden variables to the reconstructed samples through the decoding process. The following section summarizes the applications of (variational) autoencoders to small data sets in scientific research.

In 2019, Ohno et al.[64] used variable self-sorting encoders as a generative model for data augmentation to address the problem of small data volumes in regression tasks. The study utilized seven small data sets of regression type. First, the original data were divided into training and test data, and the generated model was trained on the training data. Next, sampling was carried out using generative models based on ratios. The generated samples were then trained on the regression model along with the original training data. Finally, the RMSE of the test data on the regression model was calculated. Several models were set up for comparison, including kernel density estimation using Gaussian kernel function (KDE), Variational autoencoder (single task learning (VAE)), VAE with linear regression (multitask learning), VAE with nonlinear regression (multitask learning), and denoising autoencoder with MCMC. In evaluating the test data, samples generated by the model were used, with the sample size to training data size ratio ranging from 0 to 1. Changes in the RMSE were evaluated according to the increase in the size of the training data. The experimental results for the ION data set were used as an example in the paper. The RMSE values of the KDE, VAE with linear regression, VAE with linear regression, VAE with nonlinear regression, and DAE-A models were 0.83080, 0.86738, 0.86181, 0.86258, and 1.07335, respectively. The RMSE values for the four models improved with the increasing ratio, except for DAE-A, which may be because the generated samples were highly similar to those in the training data.

In chemistry and biophysics, an issue frequently encountered is the imbalance between the number of available training and test samples. In 2022, Wei et al.[41] proposed a solution that combined a variable self-division encoder and GAN algorithm for data augmentation to address this problem. The authors demonstrated that the R^2 values of several models including ANN, VAE+ANN, GAN+ANN, RF, VAE+RF, and GAN+RF were 0.57, 0.71, 0.59, 0.89, 0.94, and 0.59, respectively. These results suggest that the proposed approach can

improve the performance of the task by enhancing the balance between training and test data sets.

In 2021, Feng et al.[315] developed a network analysis of cocaine dependence targets that involved more than 450 proteins, including dopamine (DAT), serotonin (SERT), and norepinephrine (NET) transporters. However, the available ligand binding data sets for many of these targets were limited. To improve the accuracy of their ML/DL models, the authors constructed autoencoder-assisted multitask ANN models, as depicted in Figure 12. This method was employed to facilitate drug repositioning and side effect analysis.

Nowadays, variational autoencoders are widely used not only in the field of generative models for sample design of chemical molecules[316–318] but also in other areas such as imaging[319] and text generation.[320,321] Variational self-encoders are commonly used in image and biomolecular research to generate new molecular samples. However, there are still some issues with VAEs, such as the generation of noisy data. Additionally, most VAE structures struggle with generating high-resolution image samples, making them less effective in this area compared with GAN-based generative models. As a result, VAEs are often used as feature extractors in image and molecular science.[315] However, in NLP, VAE-like models are capable of generating more coherent language samples than GANs and require only simple structures to produce fluent language, highlighting the advantages of VAEs in this field.

3.9. Transformers

SSL[322] is a type of unsupervised learning that extracts supervised signals from unlabeled data, which can be used to learn intrinsic constitutional rules and obtain desirable representations using neural networks. Because the supervised information in SSL is not manually annotated, it can be considered a branch of unsupervised learning. SSL is often the first choice for researchers to avoid the high cost of data annotation and the poor performance of traditional unsupervised learning. SSL was initially applied in computer vision and NLP[323,324] that requires large data sets for accurate representation learning. As SSL advances, it has been utilized to predict molecular properties.[325,326] For example, it can extract features from unlabeled molecular data.[28,327] Likewise, SSL can also extract features from genome data[328] to predict genome function. Recently, many studies have shown that SSL can alleviate the problem of few samples or insufficient supervised information, making it widely applicable in image classification,[329,330] recommender systems,[331] protein analysis and design,[332] speech recognition,[333] and other fields. For small data sets of biochemical molecules, many researchers have proposed SSL methods to process them.[334] In the following discussion, we will focus on the applications of SSL to small data sets of biochemical molecules.

In recent years, SSL methods have gained popularity in drug discovery.[336,337] In 2020, Shen et al.[335] proposed an SSL method called Motif Learning GNN (MoLGNN), which was trained on unlabeled chemical data to improve drug screening performance. The method was tested on three data sets, JAK1, JAK2, and JAK3, and compared to the results of three other methods in Table 5. “Non-MoLGNN” referred to a network trained using standard supervised classification methods without pretraining, ”GINVAE” indicated a procedure that

pretrained the network using GINVAE and then fine-tuned it, and ”Motif Only” meant a procedure that pretrained and then fine-tuned the network using a Motif learning network. From the table, it was found that the MoLGNN method produced superior results compared with the other methods. The results also suggested that MoLGNN can be applied to a range of machine learning tasks in chemistry, even in scenarios where high-quality labeled data was limited.

In 2021, Chen et al.[334] proposed an algebraic graph-assisted bidirectional transformer model for predicting molecular properties. The model was composed of four modules: an AG-FP generator (represented by the blue rectangle), a BT-FP generator (represented by the orange rectangle), a feature combination module using RF (represented by the green rectangle), and a downstream ML module (represented by the pink rectangle), as shown in Figure 13. The creation of BT-FPs involved two steps: training based on SSL (with a large amount of unlabeled input data) and task-specific fine-tuning. The RF algorithm was utilized to compute and rank the importance of the combined features, providing optimal features for the downstream ML algorithms. Experimental results demonstrated that the model proposed obtained the best predictions on the data sets LD50, LC50 and FDA, compared to existing advanced models, like ESTDS,[338] MACCS,[339] FP2,[339] HybridModel,[340] BTAMDL2,[341] ESTD-1,[342] Daylight-MTDNN,[339] XLOGP3,[343] and Estate2,[339] as indicated by the squared value of the Pearson correlation coefficient (R^2).

In 2022, Yang et al.[344] proposed a multitask SSL framework called SSLDR to tackle the label sparsity problem in computational drug repositioning and accelerate the drug development process. The experiments were conducted on three real-world data sets, namely, Gottlieb, Cdata set, and DNdatase.[345] The prediction results demonstrate that SSLDR not only enhances the generalization performance of the ”drug-disease association prediction” task but also leverages a multi-input decoder to improve the autoencoder’s capability to discover potential factors of drugs or diseases. Additionally, the results reveal that the SSLDR model outperforms other methods on all three data sets.

SSL has gained popularity not only in the prediction of chemical molecule properties, as evidenced by several studies,[346–348] but also in other fields such as protein[349,350] and drug design.[351] Uncovering valuable information from unlabeled data has been a vital research area, and SSL has played a critical role in this endeavor. The most significant advantage of SSL is its ability to achieve good performance without a vast number of labeled samples, which reduces labeling costs and saves time. However, SSL often requires significant memory resources during training and demands high hardware requirements. Moreover, SSL is faced with numerous challenges, such as extracting intrinsic representations from large quantities of unlabeled data and evaluating the accuracy of such representations, which are essential directions for future SSL research.

3.10. Reinforcement Learning

ML has become a ubiquitous computational method in research, and reinforcement learning (RL)[352,353] plays a significant role in it. RL studies the way natural and artificial systems can learn to predict the consequences of and optimize their behavior in environments where actions lead them from one state or situation to the next and can lead to rewards and

punishments. A common model for RL is the standard Markov Decision Process.[354,355] RL can be divided into model-based RL[356] and model-free RL,[357,358] as well as active RL[359] and passive RL.[360] DL models can also be used in RL to form deep RL (DRL).[361,362] These methods have been applied to diverse fields in biology and chemistry, including drug discovery,[363,364] protein design,[365,366] and chemical engineering.[367,368] They have also been used in image recognition[369,370] and financial markets.[371] Next, we will summarize how RL can be applied to the study of biochemical molecules in the context of small data sets.

In the field of drug discovery, evaluating compounds from libraries is one of the most time-consuming tasks. In 2022, Dou et al.[372] proposed a ML model suitable for small data sets to predict the inhibition constant (K_i) and half-maximal inhibitory concentration (IC_{50}) of compounds. The prediction task was first transformed into a simple binary classification task, and then the training data set was expanded as the original sample size was small. The paper also employed the reinforcement learning method for feature selection, as illustrated in Figure 14. Lastly, the authors used a particle swarm optimized SVM for the binary classification task, denoted as SVM+. The sample size of the K_i-related data set was 44, and that of the IC_{50}-related data set was 36. Among the classification results, the accuracy of SVM+ on the K_i data set was 0.8074, while the accuracy of traditional SVM, Gaussian NB (GNB), KNN, and RF were 0.7942, 0.7150, 0.7467, and 0.7309, respectively. Moreover, the accuracy of SVM+ on the IC_{50} data set was 0.8262, while the accuracy of traditional SVM, GNB, KNN, and RF were 0.7943, 0.7411, 0.7731, and 0.7304, respectively. Based on the experimental results, the proposed model outperformed other comparison methods.

In the field of RNA research, it is important to determine the relationship between MicroRNA and diseases to improve the treatment of complex diseases. In 2021, Cui et al.[373] presented the RFLMDA model by combining the Q-learning algorithm[374] and RL. The RFLMDA model fused three submodels, namely CMF,[375] NRLMF,[376] and LapRLS,[377] together by the Q-learning algorithm to obtain the optimal weights S. The data sets used in the experiments included MicroRNAs (with 495 samples), diseases (with 383 samples), and MiRNA–disease associations (with 5430 samples). The performance of RFLMDA was evaluated using 5-fold cross-validation and local validation in the experiments. Finally, the RFLMDA model was compared with other methods using the evaluation indexes AUC and AUPR. The experimental results showed that the AUC value of RFLMDA reached 0.9416, while the AUC values of CMF, NRLMF, and LapRLS were 0.9091, 0.9315, and 0.9367, respectively. These results demonstrate that the RL-based approach can achieve good performance on small data sets.

In 2021, Pereira et al.[378] introduced a new approach to optimize the generation of compounds that considered their biological properties and bioavailability through a DRL framework. The framework, illustrated in Figure 15, integrated several technologies, such as DL, multiobjective selection, and RL, with RL being the cornerstone. The RL algorithm updated the properties of the generated molecules by maximizing the reward function. A blood–brain barrier predictor was trained with a data set of 4534 molecules collected from various sources, and canonicalized SMILES were used to represent the molecules. Two descriptors were combined with two different oversampling methods to evaluate

the performance of the model's performance. The accuracy of SMILES+ADASYN and SMILES+SMOTE was 0.924 and 0.913, respectively, while the accuracy of extended-connectivity fingerprint (ECFP)+ADA-SYN and ECFP+SMOTE was 0.935 and 0.944, respectively. The experiments demonstrated that this approach can achieve excellent performance, even with a small number of samples.

In addition, in the field of cancer research, molecular-based cancer classification has become a hot research topic. In 2022, Prathik et al.[379] proposed a DRL model for efficient analysis of gene expression data to identify cancer types. The DRL model can easily predict cancer types from gene data sets, even with multiple classification labels. Each class was identified by the deep neural network and continuous estimation using the Q-learning method in RL. Three gene expression data sets were used in this study: glioblastoma data set (with 50 samples), brain tumor data set (with 40 samples), and lung cancer data set (with 34 samples). The principal component analysis algorithm[380] was used to analyze the data sets and extract the features, and then the DRL model was used for classification experiments. The DRL model was also compared with other classifiers such as ANN, RF, and SVM. The accuracy of the DRL model in the breast cancer, glioblastoma, and Llung cancer data sets was 98.3%, 99.2%, and 97.34%, respectively, which outperformed the other classifiers. The experimental results are summarized in Table 6, indicating that the DRL model is a useful tool in cancer classification tasks.

RL is a versatile method with applications in various fields, including ethics,[381] drug design,[382,383] psychology,[384,385] and control theory.[386,387] RL has garnered interest from researchers due to its ability to solve complex scenarios that cannot be tackled by traditional methods, as many problems can be converted to a Markov decision process and solved by using RL. However, RL also has some drawbacks. First, the learning efficiency of RL can be low, as seen in algorithms such as OpenAI Five[388] and AlphaZero.[389] These issues can be addressed using transfer learning[390] or replay buffers (also known as experiential replay).[391] Second, RL often requires high-quality data and involves a large number of computational processes. Lastly, RL's greatest feature is its generality, with a generic algorithm capable of learning almost anything. Despite the challenges associated with RL, continued research and development in this field will ensure its widespread use in various research domains.

3.11. Transfer Learning

One possible solution for scarce training data is transfer learning. This technique can address the problem of difficult label acquisition. The term "transfer learning" was formally introduced by the U.S. Department of Defense Advanced Research Projects Agency in 2005 and has been used earlier under different names in various research areas. Yang et al.[392] later provided a detailed introduction to the development, definition, classification, and application of transfer learning. Overall, transfer learning is the application of knowledge, patterns, or distributions learned on one task to different but related tasks,[393] which includes two important concepts: domain and task. The domain could be seen as a particular field at a given moment in time, and the task is to determine what needs to be done. Transfer learning is usually suitable for situations where the source domain has a relatively large amount of data, and the target domain has a small amount of data. The stronger the

correlation between the source and target domains, the better the predictive performance will be obtained. This technique reduces the need and effort to (re)collect a large training set, thus mitigating the limitation of small data sizes. In 2019, Jang et al.[394] addressed the question of what content to migrate and where to migrate it for transfer learning. As the application of transfer learning continues to expand, it is being used in many areas such as computer vision,[395–397] human–computer interaction,[398] text classification,[399–401] target recognition,[402,403] protein analysis,[404,405] and others. Transfer learning has proven to be a powerful technique in biology and chemistry, with applications in gene expression data analysis[406,407] and neuroscience research.[408] Researchers use their data sets to train transfer learning models and study their structure and function. In addition, drug molecular data can be used to predict the properties and activities of drug molecules, which can lead to the discovery of new drugs.[409,410] Fields that suffer from insufficient data or inadequate data annotation include research areas for rare diseases such as acute promyelocytic leukemia and acromegaly.[411] Below, we summarize how transfer learning can be applied to predict biochemical molecules with small data sets.

In 2019, Ye et al.[412] proposed a method that combines transfer learning and multitask learning (DeepPharm) to enhance the generalization ability of the model for scarce training data sets. The authors utilized four different data sets with small sample sizes: oral bioavailability (sample size: 410), plasma protein binding rate (sample size: 769), apparent volume of distribution at steady-state (VDss) (sample size: 412), and elimination half-life (sample size: 969). The predictive performance of DeepPharm was compared with other methods, such as SVM. The SVM method achieved an accuracy of 23% and a mean absolute error (MAE) value of 0.34 for the bioavailability data set. In contrast, DeepPharm increased the accuracy to 28% and decreased the MAE value to 0.31, which suggests that DeepPharm can be further employed in drug discovery and development.

In 2020, Sharifi-Noghabi et al.[413] introduced an adversarial inductive transfer learning technique, which combined adversarial training with inductive transfer learning, for solving problems in pharmacogenomics applications that required adaptation in both input and output spaces. The GDSC and GSE28796 data sets used in this study had small sample sizes of 829 and 12, respectively. Impressively, the adversarial inductive transfer learning technique improved the AUROC value up to 51% and 45% compared to the ProtoNet[414] and ADDA method,[415] respectively.

In addition, Bai et al.[416] developed a sequence-to-sequence (seq2seq) transfer learning method that introduced transfer learning into reverse synthesis analysis, as illustrated in Figure 16. The method utilized an unclassified large data set, USPTO 380K, for pretraining the model, followed by continuous training and reverse synthesis testing on the small data set USPTO-50K. Then, the transfer learning was combined with the seq2seq or transformer model for validation. The accuracy value obtained using the seq2seq-transfer learning method was 72.1%, which was higher than 65.9% obtained using the seq2seq baseline.[417] The experimental results demonstrated the feasibility of transferring learning between models that operated with different chemical data sets.

In 2021, Chen et al.[418] presented a neural network-based predictor of log *P*, named MRLogP, for predicting the lipophilicity of small molecules using transfer learning techniques. MRLogP achieved an average RMSE of 0.988 and 0.715 when tested on druglike molecules from Reaxys and PHYSPROP, respectively. This work demonstrates that the application of transfer learning techniques enables accurate log *P* prediction even with small experimental training data sets.

In 2017, Cang and Wei built an algebraic topology-based multitask and multichannel CNN model for predicting protein stability changes upon mutations.[419] As shown in Figure 17, this model shared and transformed algebraic topological invariants for transfer learning to the impact of mutations on protein stability. The large globular protein data set of 2648 samples was shared and simultaneously trained with a small membrane protein data set of 223 samples, which improved the prediction correlation from 0.52 to 0.57. The performances also indicated that the proposed model holds significant potential for predicting protein–ligand binding affinities and mutation-induced protein stability changes.

Transfer learning is an effective strategy for dealing with small data sets, as it can improve the accuracy of models for specific tasks. Transfer learning has been used in various fields, including activity prediction,[420] protein domain,[421,422] drug prediction,[412,423,424] image classification,[425–427] text sentiment classification,[428–430] and multilingual text classification. Although transfer learning has been extensively applied in many fields, its application in small molecule data sets is still in the early stages. Further research is needed to investigate related theoretical aspects such as the issue of transferability and the importance of data similarity and task correlation in achieving success.

Additionally, quantifying the correlation between different tasks is a challenge for transfer learning. The migration performance may depend on the source and target tasks, where the correlation of the tasks is often more important than the size of the data. Transfer learning faces various challenges, such as transfer boundaries, even though it is mainly applied in small and less fluctuating data sets.[431] Effectively transferring knowledge from one task to another simply and clearly is a significant challenge. Furthermore, using the theory of transfer learning in multitask or multidomain situations is also a question, as there may be multiple domains that differ from the target domain. Although knowledge in multiple fields can be transferred, there may be problems in transferring multiple fields that need to be resolved. Overall, most transfer learning techniques that handle small training samples achieve good experimental results and thus the issues of efficiency, experimental cost, etc., should also be considered.

3.12. Active Learning

In the industrial and scientific communities, data must be annotated to be used in ML algorithms, but this process is typically time-consuming and expensive, in terms of human and material resources. To mitigate these costs, AL methods were proposed in the ML domain. The AL concept was first introduced by Lewis in 1994.[432] The basic idea is to iteratively query an information source to obtain desirable labels, which is also termed optimal experimental design.

In general, AL can be divided into two categories: stream-based AL[433,434] and pool-based AL.[435] The stream-based AL framework requires all of the training data to be passed to the algorithm as a data stream. Each data point is sent to the algorithm separately for training, and the algorithm must decide immediately whether to label the data or not. Additionally, the training data are selected from the data pool for labeling, and the label of the current training data should be sent to the algorithm immediately before the next data point is trained. In contrast, the pool-based AL process is less complex than the stream-based approach. The training data come from an unlabeled data pool, and then the data are selected from the pool for labeling. Overall, AL is usually applied to scenarios with a large amount of unlabeled data in order to achieve the desired performance of the model with a few labeled samples.

Additionally, AL consists of five core components: the unlabeled pool, select queries, human annotator, labeled training set, and ML model. AL is mainly used in scenarios where data labeling is scarce or expensive, proactively requesting labeling and submitting the filtered data to experts for labeling to obtain a better model with fewer training samples. In the field of bioinformatics research, some chemical molecular data sets are typically small, making them ideal candidates for AL application. Recently, AL has played an essential role in predicting the biological and physical activities of small molecules in the fields of biology and chemistry. This includes predicting the structure of proteins,[436–438] as well as the toxicity of compounds.[439,440] However, how does AL deal with data sets that have a small number of labeled elements? Numerous works have been proposed to address this issue, which are outlined below.

In 2019, Zhang et al.[441] proposed a semi-supervised method using SVGD (stein variational gradient descent), called semi-supervised with SVGD, to quantify uncertainty in molecular properties. The method combined the algorithm SVGD[442] with semi-supervised learning and used AL to overcome the problem of data set bias in the training set, demonstrating that it can be robust to the uncertainty of molecules. The experiments used both small data sets, such as FreeSolv (sample size: 643), ESOL (sample size: 1128), and CatS (sample size: 595), and relatively large data sets, including MeltingPoint (sample size: 3025), p450 (sample size: 8817), and malaria (sample size: 13417). Two other methods, graph convolution with dropout and semi-supervised with dropout, were compared experimentally with semi-supervised with SVGD, where the first two methods were used in combination with dropout (dropout variational inference). The experiments were evaluated using the Spearman correlation coefficient, and the results are shown in Table 7. The experimental results demonstrated that semi-supervised with SVGD outperformed the other two methods on all six data sets.

Peptides are a popular target for biomaterials design, and their data are often scarce. In 2021, Rainier et al.[443] applied AL with CNN to binary classification of peptides, including two standard AL methods, query by committee and uncertainty minimization. The framework of the model is shown in Figure 18. The authors presented a multitask benchmark database of peptides designed to advance these methods for experimental design, and found that neither AL method tested to be better than random choice and combing meta-learning and AL could

offer inconsistent benefits. Their findings validate that AL could be used an extension to design of experiments through the selection of optimal experiments on limited resources.

In addition to the recent studies discussed above, AL has been applied not only in the field of biochemistry,[444] but also in medical imaging,[445] unmanned systems,[446] and Internet big data,[447] among others. AL has the potential to reduce the amount of annotated data required, as obtaining labeled data is a time-consuming and labor-intensive part of building ML models. As a relatively new ML method, AL aims to optimize operational resources and reduce the number of training samples.[448,449] The key idea of AL is to choose the appropriate annotation set and then manually annotate the data with the method selection depending on whether a single ML model or multiple ML models are used. Overall, AL strives to reduce annotation cost and increase model performance, enhancing the prosperity of applications in various scenarios, including imaging, NLP,[450] safety risk control, and time series anomaly detection,[451] among others. AL has the potential to be applied to more scientific tasks in the future, and effective AL strategies for optimizing repeated training in continuous data acquisition remain an important research topic.

3.13. Graph-Based Semi-Supervised Learning

Traditional ML tasks can be broadly categorized into unsupervised and supervised learning. Semi-supervised learning is a hybrid approach that addresses learning tasks where only a portion of the data is labeled and the amount of labeled data is much smaller than the unlabeled data. This approach combines the strengths of both supervised and unsupervised learning. In many practical scenarios, manually labeling samples can be expensive, which leads to very sparse labeled data. However, unlabeled data are often easily obtainable. Semi-supervised learning leverages a large amount of unlabeled data along with a small amount of labeled data to train the model, thus addressing the problem of insufficiently labeled samples. For example, the Merriman–Bence–Osher (MBO) algorithm is a popular method used in semi-supervised learning tasks. The first step is to construct a graph with a specified number of nearest neighbors, denoted as *Ne*. Then, the Laplacian and a specified number of eigenvalues and eigenvectors, again denoted as *Ne*, are calculated from the graph. A subset of the input data is selected as the labeled set for training, while the remaining data are used as unlabeled data for testing. This approach has been shown to yield good results, particularly for small data sets in the fields of biology and chemistry. For instance, it can be applied to predict biochemical molecular interactions, such as interactions between proteins and drug molecules,[452,453] or interactions between proteins.[454,455] In addition, the Nyström technique enables MBO to be used effectively with very large data sets. This section outlines recently developed techniques for applying this method to ML prediction with small data sets.

In the field of molecular and biological sciences, small or insufficiently labeled data sets are a common challenge due to the high costs of experiments. In 2022, Hayes et al.[456] proposed three new ML models, namely, an autoencoder coupled with an MBO scheme (AE-MBO), a bidirectional encoder transformer coupled with an MBO scheme (BT-MBO), and an ECFP[457] coupled with an MBO scheme. The proposed models were validated with experiments on five data sets, and their performance was compared with other methods,

such as SVMs, RFs, and gradient-boosted decision trees. Comparative experiments were conducted to test the effectiveness of the proposed models on a small amount of labeled data, using 1%, 2%, 5%, and 90% of labeled data from the data set in different models. The results for 1% labeled data for the five molecular classification data sets are presented in Figure 19. The proposed model in this article demonstrated strong predictive power in the presence of sparse marker data.

In 2021, Merkurjev et al.[458] proposed two MBO-based approaches for ML tasks with limited samples or small data sets. The first, called multikernel manifold learning (MML), integrated manifold learning with multikernel information. The second, called multiscale MBO (MMBO),[459] introduced multiscale Laplacians to a modification of the MBO scheme. These approaches were tested on various types of data sets, including α and β-protein (sample size: 900). The experimental results demonstrated that the proposed MMBO method consistently outperformed other methods and emerged as the top performer in most experiments, with the MML method closely following. Recently, the Poisson equation was used for graph based semi-supervised learning at very low label rates.[460] This approach replaced the assignment of label values at training points with the placement of sources and sinks in the Poisson equation. The resulting Poisson learning was compared with traditional Laplacian learning.

Semi-supervised learning has found wide application in various fields to solve problems encountered in real life. These fields include image classification,[461] sentiment analysis,[462] speech recognition,[463] bioinformatics,[464,465] and many others. Classification-based semi-supervised learning methods are similar to supervised methods in that they require a large amount of training data to classify the test data and thus obtain a superior classification system.[466]

The field of semi-supervised learning is aimed at building efficient learning methods and improving learning performance by leveraging the information contained in unlabeled samples.[59] Semi-supervised clustering is a specific type of clustering that uses both labeled and unlabeled data with auxiliary information to help group data patterns.[467] Additionally, reducing the dimensionality of high-dimensional data is a crucial technique in semi-supervised learning that often incorporates knowledge from the field of paired constraints.[468] However, semi-supervised learning still poses a significant challenge, particularly in addressing the problem of lack of robustness. While increasing the amount of labeled data has been proposed to counter this issue, many algorithms used in semi-supervised scenarios struggle to obtain sufficient labeled data, making this a pressing open problem.

4. PERSPECTIVES FOR MOLECULAR SCIENCE

4.1. Combining Deep Learning with Traditional Machine Learning

DL is a crucial tool in the fields of computer vision,[469] drug discovery,[470,471] and NLP,[472] where experiments often require a relatively large amount of data. DL has been widely used in the fields of chemistry and biology.[473,474] For instance, Yang and Li developed an interpretable uncertainty quantification method for DL-based molecular

property prediction.[475] Yang et al. explored the space of low-toxic chemicals through DL-based molecular generation.[476] Additionally, Pandey et al. introduced state-of-the-art DL architectures for accelerating molecular docking, evaluating off-target effects, and predicting pharmacological properties.[477]

However, when using very small data sets, DL models may struggle to establish a reliable distribution or determine their coefficients, leading to high prediction errors despite good training performance. In contrast, traditional ML methods such as KNN,[478] Bayesian network,[479] SVM,[480] and GBDT[481] tend to perform better on small data sets. While the DL is rapidly advancing, it is unlikely that it will fully replace ML algorithms. Instead, researchers have been exploring ways to combine the strengths of DL on large data sets with the strengths of ML on small data sets. This has led to an increasing amount of research on integrating DL with traditional ML algorithms for small data sets, as discussed below.

In 2022, Jiang et al. proposed a novel framework called boosting tree-assisted multitask DL (BTAMDL) for predicting chemical molecular properties.[341] The model consisted of multitask deep transfer learning and Gradient Boosting Decision Tree (GBDT). The BTAMDL model used small data sets in conjunction with related large data sets to learn the target and source tasks (involving small and large data sets, respectively) via multitask deep transfer learning and transferring knowledge from the source task to the target task. To validate the proposed method, the authors selected four types of data sets, including toxicity, log *P*, log *S*, and solvation. The toxicity data set consisted of four subsets with different sample sizes: LD_{50} (7413), IGC_{50} (1792), LC_{50} (823), and LC_{50}-DM (353). The performance of BTAMDL was compared with that of other methods in the literature, and the results are presented in Table 8, where the first four methods are those involved in the paper, and the rest can be found in the literature. The results showed that the proposed BTAMDL framework can improve the prediction performance of small data sets.

In 2021, Qiu et al.[483] presented a GBDT-based model called Bag-of-Words (BOW -GBDT for predicting drug interactions, as depicted in Figure 20. The framework consisted of three steps. First, features were obtained from the GPCR (G-proteincoupled receptor) module and combined with molecular fingerprint features. Second, the final features were generated through SMOTE (synthetic minority oversampling technique)[484] and ANN. Finally, GBDT was used to predict drug interactions. The data set D92 M used in the study was a cross-validation data set with 1860 samples, and the Check390 data set was a test data set with 390 samples. The accuracy of BOW-GBDT was reported as 86.7%, which outperformed the accuracy of 82.8% achieved by the DWKNN (ensemble) method proposed by Xiao[485] et al. in 2013.

In 2022, Yu et al.[42] developed a powerful model called SVM +GCN, which used GCNs and SVMs to classify drug data sets. The data set used in the study was a small compound data set, and the SVM+GCN model was compared with SVM, RF, and GCN methods. Two validation methods, namely, random-split validation and fingerprint-split validation, were employed to evaluate the performance of the models. The results of the experiment showed that the SVM+GCN model achieved the highest accuracy at 95.8%, while the GCN and RF models obtained accuracies of 91.6% and 87.4%, respectively.

In 2021, Deng et al.[486] proposed an integrated framework model, XGraphBoost, which combines the features of GNN with XGBoost[487] for accurate molecular property prediction, as shown in Figure 21. The original molecular data was first formatted into a graph structure, and then the molecular features were extracted through GCN, GGNN (gated GNN), and directed message passing neural network. Finally, the XGBoost classifier was used to obtain accurate predictions of the molecular properties. The data sets used in the experiments were essentially small sample data sets, including ESOL (sample size: 1128), FreeSolv (sample size: 642), Lipophilicity (sample size: 4200), HIV (sample size: 41127), BACE (sample size: 1513), BBBP (sample size: 2039), Tox21 (sample size: 7831), ToxCast (sample size: 8575), SIDER (sample size: 1427), and Clintox (sample size: 1478). The experimental results demonstrated the advantages of combining the DL with traditional ML methods.

The combination of DL methods and traditional ML algorithms is not only widely used in the field of biochemistry, but it is also prevalent in other research fields, such as cytotoxicity classification,[488] disease research,[489] and imaging.[490–492] While DL can improve prediction performance as the number of data increases, it falls short in its performance when small data sets are involved. Although DL has strong learning ability and portability, the number of model parameters will increase and their hardware requirements are also relatively high. Moreover, the model design can be relatively cumbersome. In contrast, the performance of traditional ML algorithms has an advantage in processing small data sets, and there is an increasing focus on developing methods that combine the advantages of the two. This direction represents an important area for future research.

4.2. Physical Model-Based Data Augmentation

The widespread use of ML and DL in fields such as imaging and text processing is heavily dependent on the quality and standardization of data. However, in the field of molecular science, due to the intricate complexity of molecular structures, especially macromolecules, it is challenging to obtain standardized features with the same dimensions. Many different molecule representations have been proposed,[339,493] and the field is still evolving. Adding to the challenge, there is often an insufficient amount of molecular samples available to build accurate and reliable ML models. To overcome this, researchers utilize traditional theoretical methods and physical models based on fundamental laws of physics to generate important parameters of molecular properties, which are used as labels to build and/or expand data sets. These labels are then used in downstream ML/DL procedures to predict molecular functions. Common theoretical methods and physical models used for this include molecular mechanics (MM),[494] molecular dynamics (MD),[495,496] quantum mechanics (QM),[497] quantum chemistry (QC),[498] density functional theory (DFT),[499] Monte Carlo method,[500] and finite element analysis.[501] By using theoretical calculations and simulations, researchers can generate high-quality, diverse, and large-scale training data sets, which can significantly improve the predictive accuracy of ML models. For instance, to improve the accuracy of predicting chemical reactions for small data sets, physical theory and transfer learning can be employed, as shown in ref 502. In another study, Jian et al. extended the training data set by physically modeling T-cell receptors and peptide pairs.[503] Additionally, Xie et al. presented an application of single-molecule ligation in monitoring molecular physical and chemical processes.[504]

MM, MD, and QM are all used to describe molecular interactions, including protein–protein, protein–nucleic acid, and protein–drug complexes.[505] MD, which is based on force fields, can be used to depict conformational changes in proteins or nucleic acids, the impact of mutations on protein folding stability, and the binding energies between small molecules and proteins. QM, on the other hand, based on the Schrödinger equation, can predict electronic structures involved in chemical reactions and describe polarization effects, especially with dimension-reduced DFT. While the number of atoms in the system that can be handled by MM can be as large as a million, the system suitable for QM and DFT may have only a few dozen atoms. Therefore, QM/MM methods have been developed to combine the advantages of MM and QM while taking into account the computing cost of QM and the size of most biological systems.[506] In QM/MM methods, the active site is handled by QM, while the rest of the system is considered by MM.

MM is a method used to calculate molecular structures and energies based on classical mechanical theory, using empirical and semiempirical parameters. The approach considers molecules as collections of atoms held together by elastic, van der Waals, and electrostatic forces, which reach equilibrium in the whole molecular system to determine its structure. The first use of MM dates back to 1927 when Born and Oppenheimer utilized it in their work, and it has since been widely used to calculate the conformations and energies of molecules. MM has been instrumental in determining and understanding the structure and properties of molecules since the 1950s.[507,508]

MD is the most extensively researched method in MM. Its early simulations were focused on rigid spherical systems and gradually expanded to include molten salts, metals, alloys, semiconductors, and silicates.[509–513] Various useful algorithms, such as the truncation and modification algorithm of Lennard-Jones potential function, Coulomb interaction algorithm, Verlet nearest neighbor list algorithm, and lattice index algorithm, were developed during the evolution of MD simulations.[514–516] These algorithms have greatly influenced the application of MD simulations. In recent years, MM methods have expanded beyond the study of small- and medium-sized molecules and have become capable of handling macromolecular systems.[517] These methods are implemented in popular software packages, such as AMBER,[518] and have wide distributions and applications. In various fields, such as biophysics, biochemistry, coordination chemistry, materials, and physics,[519,520] as well as in drug design,[521,522] MM methods have been extensively utilized in conjunction with lattice dynamics, energy band theory, and many other approaches.

QM is a foundational theory that investigates the electronic structure and properties of atoms, molecules, condensed matter, atomic nuclei, and basic particles. One of the most popular QM methods used since the 1990s is DFT, which has become widely applied in the study of biomolecules and materials.[523–525] DFT provides accuracy levels similar to those of semiempirical methods but at a lower computational cost. It is commonly used in condensed matter physics, computational materials, and computational chemistry, and its high efficiency allows it to handle larger and more complex systems, expanding the range of applications and the predictive power of electronic structure theory. This has also fostered greater collaboration between modelers and experimentalists.[526] However, due to the computational cost of QM methods and the large size of most biological

systems, QM/MM methods have been developed to enable electronic structure calculations of biological systems.[518,527]

As shown in Figure 22, Tavakoli et al.[528] employed DFT to determine methyl cation affinities and methyl anion affinities for over 2400 organic molecules. This work resulted in a large data set of chemical reactivity scores, which is now available to the scientific community. The authors used this data set to train several DNN, each with different representations, in order to predict reactivity. Their findings revealed that graph attention neural networks outperformed other methods and representations, achieving a 10-fold cross-validation accuracy of 92%. This work highlights the power of combining QM and ML methods to enhance the scientific understanding and promote technological progress.

Qiao et al.[529] proposed OrbNet, a framework that combined symmetry-adapted atomic orbitals features with a GNN to predict energy solutions. The experimental flowchart of OrbNet is presented in Figure 23. The authors demonstrated that OrbNet achieved prediction accuracy similar to that of DFT, but at a computational cost at least 3 orders of magnitude lower than DFT. OrbNet has been trained on approximately 100 000 molecules, and the training set can be further expanded to include more data.

Bennett et al.[530] developed 3D-CNN and spatial graph CNN models using atomic and molecular features based on atomistic MD simulations that calculated transfer free energies of 15 000 small molecules from water to cyclohexane. The DL models were trained to predict the transfer free energies based on MD-simulated data. The spatial graph CNN model showed higher accuracy than the 3D-CNN model, achieving a MAE of 4 kJ/mol when compared with MD calculations. This study suggests that the DL model can be a cost-effective alternative to expensive free energy calculations while providing similar accuracy to MD calculations. The experimental workflow is presented in Figure 24.

In the field of molecular activity prediction research, combining two-dimensional or three-dimensional descriptors with ML, can be effective for identifying active compounds. However, training ML models on data generated by MD is still being explored. In 2019, Jamal et al.[531] obtained MD descriptors using simulations and combined them with 2D and 3D descriptors. They conducted experiments using two models: ANN and RF. The final results showed that the MD descriptor outperformed both the 2D and 3D descriptors, indicating a significant improvement in the classification performance of the obtained MD descriptor.

During computer-assisted drug design, the quantitative structure/property relationships model combines experimental descriptors with those generated by MD or QM to expand data sets, which improves the prediction of molecular properties. However, the applicable conditions of each computational simulation method are limited. For example, DFT can simulate only small molecular systems, and errors in the simulation structure under high-temperature, high-pressure, and strong magnetic field environments are often significant. In addition, MD simulations are usually dependent on the accuracy of the potential function. For some applications, such as inferring force fields by ML, access to a large and diverse high-quality training data set obtained from QM calculations is essential to capture reliable

results for general applications.[532] However, there are still no known criteria for sufficiency, such as the question of how many molecular descriptors are required to explain ligand binding satisfactorily, or the question of how many large noncoding RNAs are diverse enough to represent the universe of RNA folds for these systems.[533]

Combining the simulation of these physical models with various ML algorithms could benefit the improvement of the QSAR model.[534]

4.3. Spatial and Temporal Pattern Extractions for Molecules

In recent times, there has been a significant increase in the availability of spatio-temporal data. Spatial pattern extraction, as demonstrated in ref 535, is commonly utilized to identify patterns or relationships in data that are associated with the spatial arrangement or position of data points. This is particularly useful in image classification or object detection tasks, where the accurate prediction of spatial relationships between pixels or points is crucial. Similarly, temporal pattern extraction is applied to identify patterns or relationships in data that are related to the sequence or timing of the data points. This technique is often employed in speech recognition[536] or NLP,[537] where the order of words or sequence holds significant importance for the data. With the development of molecular science, there has been growing interest in applying spatial and temporal pattern extraction to chemical and biological molecules. ML algorithms have made significant contributions to this field, as evidenced by numerous achievements.[538] In the study by Roth et al.,[539] it was found that material platforms like nanoparticles, hydrogels, and microneedles can be designed to control the interaction of vaccine components with immune cells spatially and temporally. Similarly, Goel et al.[540] explored an avenue to go beyond the space of known drug-like chemistry to benefit drug design.

A wide range of ML algorithms are available for spatial and temporal pattern extraction, including CNNs, RNNs, LSTM, GraphCNN, Autoencoders (AEs)/Stacked Autoencoders (SAEs), and Sequence-to-Sequence (Seq2Seq) models.

CNNs are primarily used to process spatial maps and are often applied to tasks such as image classification and object detection, as demonstrated in ref 144. GraphCNN is designed to handle graph data and can be further categorized into spatial maps, as shown in ref 541. RNN models, including LSTM and GRU, are particularly effective in dealing with trajectories, time series, and the sequences of spatial maps, as discussed in refs 542 and 543. ConvLSTM, a hybrid model that combines RNN and CNN, is typically used for handling spatial maps, as described in ref 544. AEs and SAEs are well-suited for extracting features from time series, trajectories, and spatial maps, as detailed in refs 308 and 545. Lastly, Seq2Seq models are generally designed for sequential data and are used for cases involving time series and trajectories, as explained in ref 416.

4.4. Natural Language Processing for Molecular Sequences

NLP, as discussed in ref 546, is the ability of a computer program to understand, interpret, and generate human language, both spoken and written, which is known as natural language. As a component of AI, NLP has various real-world applications such as language translation,[546] text classification,[547] text generation,[548] spam detection,[549] virtual agents and

chatbots,[546,550] and social media sentiment analysis.[551] Now, with more research, NLP is also being applied to chemical and biological molecules, showing powerful effects.[552,553]

NLP is also being increasingly applied to chemical and biological molecules, with promising results.[552,553] For example, Winter et al. used NLP to predict limiting activity coefficients from SMILES codes.[554] They developed a SMILES-to-properties transformer, an NLP network that accurately predicted binary limit activity coefficients from SMILES codes alone. Similarly, Lu and Zhang developed a unified DL model called T5Chem that used the Text-to-Text Transfer Transformer (T5) framework in NLP to predict various chemical reaction tasks.[555] They found that models trained with multiple tasks were more robust and can benefit from the mutual learning of related tasks. In addition, NLP can be used to predict the physiological effects of chemicals, as demonstrated by Mukherjee et al. who developed models for predicting physiological effects of chemicals based on their molecular structures using NLP methods.[556] They achieved high prediction accuracy using standard chemical data sets.

The process of NLP, as described in ref [546], can be divided into two main steps: data preprocessing and algorithm development. Data preprocessing involves preparing and cleaning text data, putting it in a workable form, and highlighting features in the text that can be analyzed by an ML algorithm. After the data have been preprocessed, an algorithm is developed to process it. There are two main types of algorithms: rules-based and ML-based. Rules-based algorithms are early NLP algorithms that use designed linguistic rules, while ML-based algorithms are used for tasks based on fed training data and can adjust their methods as more data is processed. There are various ML-based algorithms that can be used for NLP tasks, including the BOW algorithm, N-gram algorithm, word-embedding algorithm, RNN, and transformers, as explained in refs [557–561].

Syntax and semantic analysis are two primary techniques used in NLP, as explained in ref [562]. Syntax refers to the arrangement of words in a sentence to create grammatical sense, and NLP utilizes syntax to extract meaning from language based on grammatical rules. On the other hand, semantics is concerned with the meaning behind words. NLP uses various algorithms to comprehend the meaning and structure of sentences.

4.5. Generative AI for Molecular Generation

Generative AI or generative models[563] are a branch of unsupervised learning techniques in ML that are able to generate new data samples similar to a training data set, which are often used for tasks such as image generation, text generation, and data augmentation. They can also be effective in cases such as anomaly detection, where the goal is to identify examples that do not fit with the rest of the data. Generative networks or generative models are becoming increasingly popular in the field of chemical and biological molecules. According to Bilodeau et al., generative models can offer a new approach to molecular discovery by reframing molecular design as an inverse design problem.[564] Similarly, Tong et al. stated that generative models have received a lot of attention in recent years, with researchers applying them to new drug design.[565] They listed a number of publicly available generative-model-based molecular design tools that can be used to directly generate molecules. Additionally, in the study by Yakubovich et al., a computational workflow based on quantum

chemical calculations and a DNN-based generative model was proposed for the discovery of novel materials.[566] Here are four popular examples of generative network approaches as following.

GANs treat the training process as a game between two separate neural networks: a generator network and a discriminative network.[299] The generator network is trained to generate new data samples, however, the discriminative network is trained to classify samples as either coming from the true distribution or the model distribution. Every time the discriminator notices a difference between the two distributions and the generator adjusts its parameters slightly to make it go away, until at the end, the generator exactly reproduces the true data distribution and the discriminator cannot find a difference between the two distributions.

Variational autoencoders (VAEs)[567] are neural networks designed to learn an identity function in an unsupervised way to reconstruct the original input while compressing the data in the process so as to discover a more efficient and compressed representation. VAEs usually consist of an encoder network and a decoder network. The encoder network is trained to map input data samples to a latent space, while the decoder network is used to map points in the latent space back to the original data space. VAEs can be utilized to generate new data samples by sampling points in the latent space and passing them through the decoder network.

In drug discovery, it remains a challenge to create novel compounds that are not only druggable but also cheaply available. Gao et al.[568] proposed a generative network complex (GNC) model to enable the design of optimal lead compounds with desired chemical properties. The framework of the GNC model is shown in Figure 25 and GNC generated new drug-like molecules based on the multiproperty optimization in the latent space of an autoencoder. Both Monte Carlo-like random diffusion algorithm and gradient descent were used to create new molecules in the latent space. The resulting compounds were translated into SMILES strings by a decoder and further evaluated by the real space ML models.

Autoregressive models like PixelRNN[569] generate new data samples by predicting each data point in the sample based on the previous data points, which are commonly used for cases such as language modeling with the goal of predicting the next word in a sentence based on the previous words.

Generative pretraining (GPT)[570] is one of the pioneers in language understanding and modeling, and essentially proposes the concept of pretraining a language model on a huge corpus of data, and then fine-tuning the model for downstream tasks. The core ideas of GPT are attention mechanism and unsupervised pretraining. The reason for unsupervised learning is the shortage of massive labeled data sets. GPT and its extensions GPT-2 and GPT-3 are well-known for their impressive performance on small data or zero-shot learning which is a scenario wherein at test time the samples provided are not observed while training, and have successfully applied to a variety of tasks, such as machine translation,[571] question-answering,[570] reading of conceptual works, scripting of poems and elementary mathematics, etc., ChatGPT has gained a lot of popularity recently due to its impressive

strengths, such as increased efficiency and precision in NLP-related tasks. It is capable of providing answers to a wide range of issues promptly and accurately, making it invaluable in assisting with routine tasks, generating algorithms for computing tasks, and much more. However, its potential applications in molecular science, especially, in small molecular data sets, have yet to be fully explored because ChatGPT is trained on an extensive corpus of data. It is likely that ChatGPT will be proven useful in studying chemical and biological molecules in the future, but further research is still needed to confirm this.

4.6. Material Science

In recent years, machine learning methods have been successfully applied to predict chemical and material properties, particularly in material science. However, due to restrictions or limitations, collecting large labeled training samples is typically difficult in this field, which significantly reduces the predictive power of sophisticated deep learning models like convolutional neural networks and recurrent neural network. To address these small data challenges, simple regression models can be used by creating linear combinations of nonlinear basis functions.[580] For instance, when predicting the properties of elpasolite crystals, deep learning with a black box model may not be the optimal option for exploring the elpasolite universe and predicting the spin states of transition metal complexes. In such cases, the nature of the variables present in the linear model and the knowledge of the physics of the underlying problem can facilitate the identification of when simplistic linear solutions will bring comparable performance. Linear solutions can not only accurately predict material properties such as the bandgap and formation energy of transparent conducting oxides, the spin states for transition metal complexes, and the formation energy for elpasolite structures but also offer an excellent approach for interpretable predictions in the material science community.

5. OUTLOOK

In this review, we examine recent progress in addressing the challenge of working with small scientific data sets in machine learning and deep learning. Due to various constraints and limitations in data acquisition, small data sets are ubiquitous in scientific fields. The small data challenge in machine learning can be just as severe, if not more so, as the big data challenge. One of the most immediate problems posed by small scientific data sets is overfitting, which can occur not only during training but also during testing, ultimately leading to less accurate and reliable machine learning models. Additionally, small data sets are often associated with data imbalance. For example, in drug discovery, only a few drug candidates may be active, whereas for machine learning modeling, active and inactive samples should be well-balanced. Data imbalance can result in inaccurate, unreliable, and unstable machine learning and deep learning models. Moreover, augmenting small data sets using computational approaches can easily introduce noise or nonuniform data, which also presents a challenging issue in machine learning and deep learning. As summarized in Table 9, this paper reviews several approaches to address the challenges posed by small data sets, including transfer learning or multitask learning, combining traditional ML algorithms with deep learning, self-supervised learning, Generative Adversarial Networks, variational autoencoders, transformers, long short-term memory, active learning, semi-supervised

learning, and physical model-based data augmentation. While many of these approaches have been proposed in the past decade and are still in the early stages of development, there have been tremendous advances in recent years. However, the small data challenge remains a pressing issue in machine learning and deep learning, calling for innovative strategies.

Given the widespread need for machine learning techniques to handle large-scale training samples coupled with the increasing progress of small data studies, the concepts and methods of small data research are now being applied to diverse applications. In this regard, we highlight a few forefronts of the development and application of the machine learning methods for small data challenges in molecular science, particularly in molecular properties discovery, multilinear models in material science, machine learning force fields,[574,575] protein folding,[576] catalyst design,[577] and retrosynthetic pathways.[578]

5.1. Machine Learning Force Fields

Machine learning force fields[574,579] are applied to overcome the size limitations of accurate ab initio methods, by learning the energies and interactions in atomic-scale systems directly from, for example, density functional theory calculations. Unlike conventional force fields, Machine learning force fields are built on mathematical structures with limited underlying physics and chemistry concepts. Therefore, it is crucial to train the machine learning force field on relevant density functional theory data, such as energies, forces, and stress, to obtain a robust Machine learning force field for particular systems and applications. During training, the atomic environments in a configuration are transformed into a set of features that are then used to predict the energies of the atomic configuration for downstream tasks. Once the training is complete, the machine learning force field model can be used for atomic-scale simulations, much like any other conventional force field.

5.2. Biomolecular Properties Discovery

One of the most challenging issues in drug design and substance discovery is predicting molecular properties. Traditional methods based on density functional theory have explicit physical images but are time-consuming for processing large numbers of molecules. In recent years, data-driven machine learning models have successfully learned the relationship between the structure and properties of a molecule and can perform low-cost predictions instead of costly and time-consuming processes involving human expertise, computer simulation, and subsequent experimental synthesis. However, due to the complexity, cost, and time required to obtain molecular information experimentally, it is often difficult to obtain large labeled molecular data sets. Several approaches have been developed to address this challenge. Hayes et al. introduced three graph-based MBO models for molecular classification prediction with scarcely labeled data, including Ames, Bace, BBBP, Beet, and ClinTox data sets.[456] Jiang et al. built a BTAMDL architecture that integrates GBDT and multitask deep learning to achieve near-optimal predictions for small molecular properties such as partition coefficient, solubility, toxicity, and solvation.[341]

5.3. Protein Folding Prediction

Protein folding plays a decisive role in the biological functions of proteins. Predicting protein folding modes is crucial in expressing their spatial topological features and can be

solved as a classification problem with ML methods. Generally, machine learning algorithms take amino acid sequences as input and predict folding patterns by extracting features, which are then fed into a classifier for prediction and performance evaluation. Alphafold2 is currently the most popular tool for protein folding, which combines knowledge of protein structure with deep learning.[581] While Alphafold2 has achieved significant success in protein folding prediction, its predictive accuracy is lower compared to experimental techniques such as X-ray crystallography.[582] Additionally, running Alphafold2 requires substantial computational resources. Other advanced methods such as DeepSVM fold have also been proposed, which achieved a prediction accuracy of 67.3% and outperformed other methods.[583]

5.4. Catalyst Design

To understand catalyst catalysis, it is necessary to accurately identify descriptors of catalytic activity.[584] However, traditional methods often lack predictability and accuracy, leading researchers in catalyst design to focus on improving the accuracy of identifying catalyst descriptors and predicting rates using machine learning. In a recent study by Wenjie Liao et al.,[585] an enhanced method for accurately identifying descriptors was proposed using a machine learning surrogate model derived from a kinetic data set, which outperformed traditional derivative-based methods. Density functional theory is a commonly used computational chemistry tool for studying and predicting the geometric structure, mechanical properties, electronic structure, and reaction energies of materials. Xuhao Wan et al.[586] introduced a DFT-based machine learning approach (DMCP) and used transition metal phthalocyanine diatomic catalysts as electrocatalysts for carbon reduction reactions.

5.5. Retrosynthetic Pathways

Retrosynthesis, which was proposed by Corey in the 1960s, describes the iterative process of reducing a complex target molecule to a simple precursor by breaking bonds.[587] It summarizes the reverse work that organic chemists need to do when building new molecules, and these chemists have identified a series of chemical transformations that can be achieved through the simpler chemical structure of oil or other resources.[588] Currently, retrosynthetic programs are mainly divided into logic-based heuristic programs and detailed retrosynthetic route prediction programs.[589] Moreover, Badowski et al.[590] have shown that synergy between expert and machine learning approaches can lead to improved retrosynthetic planning. In the future, high-quality databases will accelerate further developments in retrosynthesis.[591]

5.6. Computational Chemistry

Recently, computational chemistry and machine learning have increasingly been combined to enhance the understanding and prediction of chemical and physical properties and behavior. More and more machine learning and deep learning techniques are borrowed in computational chemistry to generate models and algorithms to extract patterns and relationships from large data sets and to make predictions about chemical systems. For some chemical systems with limited data, there has been great progress made in recent years. For instance, Lilienfeld et al. developed quantum machine learning models to predict various molecular properties, such as energy, electronic structure, and spectroscopic data.[592] Ceriotti

et al. explored different machine learning methods to understand the behavior of molecules and materials in systems with various size at the atomic scale.[593] Moreover, Csanyi et al. provided a unified framework to predict atomic-scale properties based on local description of chemical environments and Bayesian statistical learning.[594] More related references can be found in refs 595–600.

We conclude our review with several reminders of challenges that need to be addressed when dealing with small data sets in machine learning and deep learning.

5.7. Modelability Metrics

It is essential to develop metrics for measuring the modelability of small data sets, which can be used to evaluate all methods, including transfer learning, where the data similarity index is closely related to the modelability.

5.8. Small and Diverse Data Sets

Developing machine learning and deep learning methods for handling small and diverse data sets is particularly challenging. Data diversity is closely related to data modelability, especially for small data sets.

5.9. Small and High-Dimensional Data Sets

Developing machine learning/deep learning methods for tackling small and high-dimensional data sets, especially for single-cell RNA sequencing (sc-RNA-seq) and transcriptomic data analysis, is another important task. Traditional approaches such as principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and uniform manifold approximation and projection (UMAP) have limited success, and machine learning and deep learning methods are expected to address challenges from the spatiotemporal entanglement of cells, genes, and tissues.

5.10. Small and Noisy Data Sets

Moreover, addressing truly small and noisy data sets is one of the most challenging tasks in machine learning and deep learning. Currently, there is limited feasibility and few results for this problem in the literature.

5.11. Small and Imbalanced Data Sets

The modeling of small and imbalanced data is a difficult issue that needs to be addressed. Imbalanced data sets naturally occur in experimental settings where successful results are reported while unsuccessful ones are ignored. On the other hand, in drug discovery, most drug candidates are unsuccessful.

5.12. Data Imputation in Small Data Sets

Treating concurrent small data and data imputation can be very challenging. This treatment is often a needed preprocessing in machine learning studies. It will be an important research topic.

5.13. Data Representability

The quantitative analysis of data representability will be an interesting issue. The construction of effective descriptors will continue to be an important area of research, particularly for data with intrinsically complex internal structures, such as biomolecules, macromolecules, and functional materials.

5.14. Machine/Deep Learning Complexes

The construction of sophisticated machine learning complexes that integrate different ML methods to deal with small data sets, such as using migration learning in combination with Generative Adversarial Networks, while optimizing the data and the model framework to obtain the desired results, will be both challenging and important. It is expected that such complexes will become common in molecular sciences.

5.15. Data Understanding

Finally, one cannot overemphasize the role of physical/chemical/biological understanding of data in the machine learning method design, development, or selection and machine learning result interpretation. It is important to utilize prior domain knowledge about small data sets to design machine learning and deep learning methods and improve their predictability.

We apologize that we could not review all related concepts and issues and cover all important references about machine learning and deep learning approaches for dealing with small scientific data sets. We hope that the reader can benefit from the perspective presented in this review and find a way to tackle the small data challenge.

ACKNOWLEDGMENTS

This work was supported in part by NIH grants R01GM126189 and R01AI164266, NSF grants DMS-2052983, DMS-1761320, and IIS-1900473, NASA grant 80NSSC21M0023, MSU Foundation, Bristol-Myers Squibb 65109, and Pfizer. The work of Jian Jiang and Bengong Zhang was supported by the National Natural Science Foundation of China under grant no. 11971367, no.12271416, and no.11972266.

Biographies

Bozheng Dou obtained his B.S. degree in Geomatics Engineering in 2021 from Nanjing University of Information Science and Technology. He is currently a M.S. candidate at Wuhan Textile University under Dr. Jie Liu and Dr. Jian Jiang.

Zailiang Zhu obtained his B.S. degree in Software Engineering in 2019 from Zhongyuan University of Technology and received his M.S. degree in Software Engineering in 2023 from Wuhan Textile University under Dr. Bengong Zhang and Dr. Jian Jiang.

Ekaterina Merkurjev is an Assistant Professor in the Departments of Mathematics and CMSE at Michigan State University. She received her PhD in Applied Mathematics from the University of California, Los Angeles, in 2015 under the supervision of Prof. Andrea Bertozzi. She also obtained her Bachelors degree from the University of California, Los Angeles. Her research interests include graph-based methods, semi-supervised learning, and image processing.

Lu Ke obtained her B.S. degree in Applied Mathematics in 2022 from Wuhan Textile University. She is currently a M.S. candidate at Wuhan Textile University under Dr. Bengong Zhang and Dr. Jian Jiang.

Long Chen obtained his B.S. degree in Information and Computing Science in 2022 from Wuhan Textile University. He is currently a M.S. candidate at Wuhan Textile University under Dr. Yueying Zhu and Dr. Jian Jiang.

Jian Jiang received his B.S and M.S. degrees in Theoretical Physics at Central China Normal University in 2005 and 2007, respectively. He received his Ph.D. degree in Theoretical Physics from University of Le Mans in France in 2011. He is currently a professor in the School of Mathematical and Physical at Wuhan Textile University in China. His research interest includes topological data analysis on molecular science, drug design and discovery, data mining, and modelling and analysis of complex networks.

Yueying Zhu obtained her Ph.D. degree in Physics from Le Mans University and Central China Normal University under the mentorship of Profs. Qiuping Alexandre Wang, Xu Cai, and Wei Li. Her Ph.D. study focused on the uncertainty and sensitivity analysis of nonlinear dynamical systems, and the modeling and simulations of spreading dynamics on complex network. Now, she is an associate professor in Prof. Jie Liu's group at Wuhan Textile University. Her current research concerns the application of uncertainty and sensitivity analysis to spreading dynamics, especially epidemic and opinion spreading on a complex network.

Jie Liu received his B.S. degree in Mathematics from Hubei Normal University, China, in 1997, and received his M.S. degree and Ph.D. degree in Application Mathematics and Computation Mathematics from Wuhan University, China, in 2003 and 2006, respectively. He is currently a professor in School of Mathematical and Physical at Wuhan Textile University in China. His research interest includes the analysis of dynamical nonlinear systems, control of complex systems, and modelling and analysis of complex networks. He is the author of more than 80 journal and conference papers.

Bengong Zhang received his Ph.D. degree in Applied Mathematics from South China University of Technology in 2010 and completed his postdoctoral studies at The University of Tokyo in 2013 under the guidance of Prof. Kazuyuki Aihara and Luonan Chen. His postdoctoral studies focused on computational systems biology. Now he is the professor of Wuhan Textile University, Wuhan, China. His current research concerns scRNA-seq data analysis and machine learning.

Guo-Wei Wei earned his Ph.D. degree from the University of British Columbia in 1996. He was awarded a fellowship from the NSERC of Canada to pursue his postdoctoral work at the University of Houston. In 1998, he joined the faculty of the National University of Singapore and was promoted to Associate Professor in 2001. In 2002, he relocated to Michigan State University, where he is an MSU Foundation Professor of Mathematics, Electrical and Computer Engineering, and Biochemistry and Molecular Biology. His research explores the mathematical foundations of biological science and data science,

including deep learning, drug discovery, and computational geometry, topology, and algebra. Dr. Wei has served extensively on a wide variety of national and international panels, committees, and journal editorships.

ABBREVIATIONS

AI	artificial intelligence
AL	active learning
ANN	artificial neural network
AUC	area under the curve
BLSTM	bidirectional LSTM
BOW	bag-of-words
BTAMDL	boosting tree-assisted multitask deep learning
CNNs	convolutional neural networks
DFT	density functional theory
DL	deep learning
DNN	deep neural networks
Deep RL	DRL
ECFP	extended-connectivity fingerprint
F-measure	F-measure represents the harmonic mean of precision and recall
GAN	Generative Adversarial Network
GBDT	Gradient Boosting Decision Tree
GF-VAE	Graph Flow-Variational AutoEncoder
GNC	generative network complex
GNN	graph neural network
GCNs	graph convolutional networks
GPCR	G protein-coupled receptor
KDE	kernel density estimation using Gaussian kernel function
KNN	K-nearest neighbor
LSTM	long short-term memory
log *P*	partition coefficient

log *S*	solubility
MBO	Merriman–Bence–Osher
MAE	mean absolute error
MD	molecular dynamics
ML	machine learning
MM	molecular mechanics
MMBO	multiscale MBO
MML	multikernel manifold learning
MoLGNN	Motif Learning Graph Neural Network
NB	naive Bayes
NLP	natural language processing
ProGAN	Protein Solubility Generative Adversarial Net
QC	quantum chemistry
QM	quantum mechanics
RF	random forest
RNNs	recurrent neural networks
R^2	coefficient of determination
RMSE	root mean square error
SSL	self-supervised learning
SVM	support vector machine
VAE	variational auto-encoder

REFERENCES

(1). Jordan MI; Mitchell TM Machine learning: Trends, perspectives, and prospects. Science 2015, 349, 255–260. [PubMed: 26185243]

(2). Campbell C Springer Handbook of Bio-/Neuroinformatics; Springer, 2014; pp 185–206.

(3). Lutnick B; Ginley B; Govind D; McGarry SD; LaViolette PS; Yacoub R; Jain S; Tomaszewski JE; Jen K-Y; Sarder P Iterative annotation to ease neural network training: Specialized machine learning in medical image analysis. arXiv 2018, arXiv.1812.07509.

(4). Keith JA; Vassilev-Galindo V; Cheng B; Chmiela S; Gastegger M; Muller K-R; Tkatchenko A Combining machine learning and computational chemistry for predictive insights into chemical systems. Chem. Rev. 2021, 121, 9816–9872. [PubMed: 34232033]

(5). Chen M; Wei Z; Huang Z; Ding B; Li Y Simple and deep graph convolutional networks. In Proceedings of the 37th International Conference on Machine Learning, 2020; Vol. 110, pp 1725–1735

(6). O'Shea K; Nash R An introduction to convolutional neural networks. arXiv 2015, arXiv.1511.08458.

(7). Mandic D; Chambers J Recurrent Neural Networks for Prediction: Learning Algorithms, Architectures and Stability; Wiley, 2001.

(8). Creswell A; White T; Dumoulin V; Arulkumaran K; Sengupta B; Bharath AA Generative adversarial networks: An overview. IEEE Signal Process Mag 2018, 35, 53–65.

(9). He K; Zhang X; Ren S; Sun J Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; IEEE, 2016; pp 770–778.

(10). Läubli S; Castilho S; Neubig G; Sennrich R; Shen Q; Toral A A set of recommendations for assessing human–machine parity in language translation. J. Artif. Intell. Res. 2020, 67, 653–672.

(11). Hinton G; Deng L; Yu D; Dahl GE; Mohamed A.-r.; Jaitly N; Senior A; Vanhoucke V; Nguyen P; Sainath TN; et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Process Mag 2012, 29, 82–97.

(12). Silver D; Huang A; Maddison CJ; Guez A; Sifre L; Van Den Driessche G; Schrittwieser J; Antonoglou I; Panneershelvam V; Lanctot M; et al. Mastering the game of Go with deep neural networks and tree search. nature 2016, 529, 484–489. [PubMed: 26819042]

(13). Altae-Tran H; Ramsundar B; Pappu AS; Pande V Low data drug discovery with one-shot learning. ACS Cent. Sci. 2017, 3, 283–293. [PubMed: 28470045]

(14). Hariono M; Wijaya DB; Chandra T; Frederick N; Putri AB; Herawati E; Warastika LA; Permatasari M; Putri AD; Ardyantoro S A Decade of Indonesian Atmosphere in Computer-Aided Drug Design. J. Chem. Inf. Model. 2022, 62, 5276–5288. [PubMed: 36373286]

(15). Wang W; Zheng VW; Yu H; Miao C A survey of zero-shot learning: Settings, methods, and applications. ACM T INTEL SYST TEC. 2019, 10, 1–37.

(16). Eloff R; Engelbrecht HA; Kamper H Multimodal one-shot learning of speech and images. ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing; ICASSP. 2019; pp 8623–8627.

(17). Prabhu V; Kannan A; Ravuri M; Chaplain M; Sontag D; Amatriain X Few-shot learning for dermatological disease diagnosis. In Machine Learning for Healthcare Conference, 2019; pp 532–552.

(18). Wang Y; Yao Q; Kwok JT; Ni LM Generalizing from a few examples: A survey on few-shot learning. ACM Comput. Surv. 2021, 53, 1–34.

(19). Pham HNA; Triantaphyllou E Soft Computing for Knowledge Discovery and Data Mining; Springer, 2008; pp 391–431.

(20). Barman R; Deshpande S; Agarwal S; Inamdar U; Devare M; Patil A Transfer learning for small dataset. In Proceedings of the National Conference on Machine Learning, Mumbai, India, 2019; pp 132–137.

(21). Li X; Fourches D Inductive transfer learning for molecular activity prediction: Next-Gen QSAR Models with MolPMoFiT. J. Cheminform. 2020, 12, 27. [PubMed: 33430978]

(22). Kumar SA; Ananda Kumar TD; Beeraka NM; Pujar GV; Singh M; Narayana Akshatha HS; Bhagyalalitha M Machine learning and deep learning in data-driven decision making of drug discovery and challenges in high-quality data acquisition in the pharmaceutical industry. Future Med. Chem. 2022, 14, 245–270. [PubMed: 34939433]

(23). Chato L; Latifi S Machine learning and deep learning techniques to predict overall survival of brain tumor patients using MRI images. In 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE); IEEE, 2017; pp 9–14.

(24). Li J; Topaloglu RO; Ghosh S Quantum generative models for small molecule drug discovery. IEEE Trans. Quantum Eng. 2021, 2, 1–8.

(25). Xu Y; Zhang Z; You L; Liu J; Fan Z; Zhou X scIGANs: single-cell RNA-seq imputation using generative adversarial networks. Nucleic Acids Res. 2020, 48, e85–e85. [PubMed: 32588900]

(26). Hadipour H; Liu C; Davis R; Cardona ST; Hu P Deep clustering of small molecules at large-scale via variational autoencoder embedding and K-means. BMC Bioinform. 2022, 23, 132.

(27). Armitage J; Spalek LJ; Nguyen M; Nikolka M; Jacobs IE; Marañón L; Nasrallah I; Schweicher G; Dimov I; Simatos D, et al. Fragment graphical variational autoencoding for screening molecules with small data. arXiv 2019, arXiv.1910.13325.

(28). Zhang Z; Liu Q; Wang H; Lu C; Lee C-K Motif-based graph self-supervised learning for molecular property prediction. arXiv 2021, arXiv, 2110.00987.

(29). Liu H; HaoChen JZ; Gaidon A; Ma T Self-supervised learning is more robust to dataset imbalance. arXiv 2021, arXiv.2110.05025.

(30). Wang Y-B; You Z-H; Yang S; Yi H-C; Chen Z-H; Zheng K A deep learning-based method for drug-target interaction prediction based on long short-term memory neural network. BMC Med. Inf. Decis. Making 2020, 20, 49.

(31). Chakravarti SK; Alla SRM Descriptor free QSAR modeling using deep learning with long short-term memory neural networks. Front. Artif. Intell. 2019, 2, 17. [PubMed: 33733106]

(32). Rodriguez Serrano AF; Hsing I-M Prediction of Aptamer–Small-Molecule Interactions Using Metastable States from Multiple Independent Molecular Dynamics Simulations. J. Chem. Inf. Model. 2022, 62, 4799–4809. [PubMed: 36134737]

(33). Kumar A; Purohit R Use of long term molecular dynamics simulation in predicting cancer associated SNPs. PLoS Comput. Biol. 2014, 10, No. e1003318. [PubMed: 24722014]

(34). Azzimonti D; Rottondi C; Giusti A; Tornatore M; Bianco A Comparison of domain adaptation and active learning techniques for quality of transmission estimation with small-sized training datasets. J. Opt. Commun. Networking 2021, 13, A56–A66.

(35). Quteineh H; Samothrakis S; Sutcliffe R Textual data augmentation for efficient active learning on tiny datasets. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020; pp 7400–7410.

(36). Inés A; Domínguez C; Heras J; Mata E; Pascual V Biomedical image classification made easier thanks to transfer and semi-supervised learning. Comput. Methods Programs Biomed. 2021, 198, 105782. [PubMed: 33065493]

(37). Hyun M; Jeong J; Kwak N Class-imbalanced semi-supervised learning. arXiv 2020, arXiv.2002.06815.

(38). Young SI; Balbastre Y; Dalca AV; Wells WM; Iglesias JE; Fischl B SuperWarp: Supervised Learning and Warping on U-Net for Invariant Subvoxel-Precise Registration. arXiv 2022, arXiv.2205.07399.

(39). Farasin A; Colomba L; Garza P Double-step u-net: A deep learning-based approach for the estimation of wildfire damage severity through sentinel-2 satellite data. Appl. Sci. 2020, 10, 4332.

(40). Chen J; Wang J; Wang X; Du Y; Chang H Predicting drug target interactions based on GBDT. In International Conference on Machine Learning and Data Mining in Pattern Recognition. 2018; pp 202–212.

(41). Wei S; Chen Z; Arumugasamy SK; Chew IML Data augmentation and machine learning techniques for control strategy development in bio-polymerization process. Environ. Sci. Ecotechnol. 2022, 11, 100172. [PubMed: 36158757]

(42). Yu T-H; Su B-H; Battalora LC; Liu S; Tseng YJ Ensemble modeling with machine learning and deep learning to provide interpretable generalized rules for classifying CNS drugs with high prediction power. Briefings Bioinf. 2022, 23, bbab377.

(43). Lazarovits J; Sindhwani S; Tavares AJ; Zhang Y; Song F; Audet J; Krieger JR; Syed AM; Stordy B; Chan WC Supervised learning and mass spectrometry predicts the in vivo fate of nanomaterials. ACS Nano 2019, 13, 8023–8034. [PubMed: 31268684]

(44). Sandfort F; Strieth-Kalthoff F; Kühnemund M; Beecks C; Glorius F. A structure-based platform for predicting chemical reactivity. Chem. 2020, 6, 1379–1390.

(45). Das P; Mazumder DH An extensive survey on the use of supervised machine learning techniques in the past two decades for prediction of drug side effects. Artif. Intell. Rev. 2023, 2023, 10413–7.

(46). Muñoz E; Nováček V; Vandenbussche P-Y. Facilitating prediction of adverse drug reactions by using knowledge graphs and multi-label learning models. Briefings Bioinf. 2019, 20, 190–202.

(47). Zhou H; Cao H; Matyunina L; Shelby M; Cassels L; McDonald JF; Skolnick J MEDICASCY: a machine learning approach for predicting small-molecule drug side effects, indications, efficacy, and modes of action. Mol. Pharmaceutics 2020, 17, 1558–1574.

(48). Zhou L; Kaess M Windowed bundle adjustment framework for unsupervised learning of monocular depth estimation with u-net extension and clip loss. IEEE Rob. Autom. Lett. 2020, 5, 3283–3290.

(49). Khan Z; Yang J Bottom-up unsupervised image segmentation using FC-Dense u-net based deep representation clustering and multidimensional feature fusion based region merging. Image Vision Comput. 2020, 94, 103871.

(50). Pena JM; Lozano JA; Larranaga P; Inza I Dimensionality reduction in unsupervised learning of conditional Gaussian networks. IEEE Trans. Geosci. Electron. 2001, 23, 590–603.

(51). Glielmo A; Husic BE; Rodriguez A; Clementi C; Noé F; Laio A Unsupervised learning methods for molecular simulation data. Chem. Rev. 2021, 121, 9722–9758. [PubMed: 33945269]

(52). Basdogan Y; Groenenboom MC; Henderson E; De S; Rempe SB; Keith JA Machine learning-guided approach for studying solvation environments. J. Chem. Theory Comput. 2020, 16, 633–642. [PubMed: 31809056]

(53). Chen D; Ao Y; Liu S Semi-supervised learning method of u-net deep learning network for blood vessel segmentation in retinal images. Symmetry 2020, 12, 1067.

(54). Oymak S; Gulcu TC Statistical and algorithmic insights for semi-supervised learning with self-training. arXiv 2020, arXiv.2006.11006.

(55). Xia Y; Liu F; Yang D; Cai J; Yu L; Zhu Z; Xu D; Yuille A; Roth H 3D semi-supervised learning with uncertainty-aware multi-view co-training. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision; IEEE/CVF, 2020; pp 3646–3655.

(56). Li S; Li W-T; Wang W Co-gcn for multi-view semi-supervised learning. In Proceedings of the AAAI Conference on Artificial Intelligence; AAAI. 2020; pp 4691–4698.

(57). Ji C; Wang Y; Gao Z; Li L; Ni J; Zheng C A semi-supervised learning method for MiRNA-disease association prediction based on variational autoencoder. IEEE/ACM Trans. Comput. Biol. Bioinf. 2022, 19, 2049–2059.

(58). Yin C; Chen Z Developing sustainable classification of diseases via deep learning and semi-supervised learning. Healthcare 2020, 8, 291. [PubMed: 32846941]

(59). Kostopoulos G; Karlos S; Kotsiantis S; Ragos O Semi-supervised regression: A recent review. J. Intell. Fuzzy Syst. 2018, 35, 1483–1500.

(60). Salvador A; Gundogdu E; Bazzani L; Donoser M Revamping cross-modal recipe retrieval with hierarchical transformers and self-supervised learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; IEEE, 2021; pp 15475–15484.

(61). Huang H; Wang T; Cheng J; Xiong Y; Wang C; Geng J Self-Supervised Deep Learning to Reconstruct Seismic Data With Consecutively Missing Traces. IEEE Trans. Geosci. Electron. 2022, 60, 5911514.

(62). Wang Y; Chen X; Min Y; Wu J Molcloze: a unified cloze-style self-supervised molecular structure learning model for chemical property prediction. In 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); IEEE, 2021; pp 2896–2903.

(63). Zhang S; Yan Z; Huang Y; Liu L; He D; Wang W; Fang X; Zhang X; Wang F; Wu H; wang H HelixADMET: a robust and endpoint extensible ADMET system incorporating self-supervised knowledge transfer. Bioinformatics 2022, 38, 3444–3453. [PubMed: 35604079]

(64). Ohno H Auto-encoder-based generative models for data augmentation on regression problems. Soft Comput. 2020, 24, 7999–8009.

(65). Pekel E Estimation of soil moisture using decision tree regression. Theor. Appl. Climatol. 2020, 139, 1111–1119.

(66). Massy WF Principal components regression in exploratory statistical research. J. Am. Stat. Assoc. 1965, 60, 234–256.

(67). Segal MR Machine Learning Benchmarks and Random Forest Regression; Center for Bioinformatics & Molecular, 2004.

(68). Smola AJ; Schölkopf B A tutorial on support vector regression. Stat. Comput. 2004, 14, 199–222.

(69). Ostertagová E Modelling using polynomial regression. Procedia Eng. 2012, 48, 500–506.

(70). Rahman MM; Saha T; Islam KJ; Suman RH; Biswas S; Rahat EU; Hossen MR; Islam R; Hossain MN; Mamun AA; et al. Virtual screening, molecular dynamics and structure–activity relationship studies to identify potent approved drugs for Covid-19 treatment. J. Biomol. Struct. Dyn. 2021, 39, 6231–6241. [PubMed: 32692306]

(71). Yan A; Chong Y; Wang L; Hu X; Wang K Prediction of biological activity of Aurora-A kinase inhibitors by multilinear regression analysis and support vector machine. Bioorg. Med. Chem. Lett. 2011, 21, 2238–2243. [PubMed: 21421314]

(72). Ye Q; Li Q; Gao A; Ying H; Cheng G; Chen J; Che J; Li J; Dong X; Zhou Y Discovery of novel indoleaminopyrimidine NIK inhibitors based on molecular docking-based support vector regression (SVR) model. Chem. Phys. Lett. 2019, 718, 38–45.

(73). Chen Y; Liu Y; Podimata C Learning strategy-aware linear classifiers. Adv. Neural Inf. Process. Syst. 2020, 33, 15265–15276.

(74). Wei K; Li T; Huang F; Chen J; He Z Cancer classification with data augmentation based on generative adversarial networks. Front. Comput. Sci. 2022, 16, 162601.

(75). Arian R; Hariri A; Mehridehnavi A; Fassihi A; Ghasemi F Protein kinase inhibitors classification using K-Nearest neighbor algorithm. Comput. Biol. Chem. 2020, 86, 107269. [PubMed: 32413830]

(76). Madhulatha TS An overview on clustering methods. arXiv 2012, arXiv.1205.1117.

(77). Duran BS; Odell PL Cluster Analysis: A Survey; Springer Science & Business Media, 2013; Vol. 100.

(78). Xu Q; Zhang Q; Liu J; Luo B Efficient synthetical clustering validity indexes for hierarchical clustering. Expert Syst. Appl. 2020, 151, 113367.

(79). Uppada SK Centroid based clustering algorithmsA clarion study. Int. J. Comput. Sci. Inform. Technol. 2014, 5, 7309–7313.

(80). Xu X; Ester M; Kriegel H-P; Sander J A distribution-based clustering algorithm for mining in large spatial databases. In Proceedings of the 14th International Conference on Data Engineering. 1998; pp 324–331.

(81). Kriegel H-P; Kröger P; Sander J; Zimek A Density-based clustering. Wiley Interdiscip. Rev.: Data Min. Knowl. Discovery 2011, 1, 231–240.

(82). Park NH; Lee WS Statistical grid-based clustering over data streams. Acm Sigmod Record 2004, 33, 32–37.

(83). Ferro S; Bottigliengo D; Gregori D; Fabricio AS; Gion M; Baldi I Phenomapping of patients with primary breast cancer using machine learning-based unsupervised cluster analysis. J. Pers. Med. 2021, 11, 272. [PubMed: 33916398]

(84). Yansari RT; Mirzarezaee M; Sadeghi M; Araabi BN A new survival analysis model in adjuvant Tamoxifen-treated breast cancer patients using manifold-based semi-supervised learning. J. Comput. Sci. 2022, 61, 101645.

(85). Sorzano COS; Vargas J; Montano AP A survey of dimensionality reduction techniques. arXiv, 2014, arXiv.1403.2877.

(86). Pearson K LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 1901, 2, 559–572.

(87). Mulaik SA Foundations of Factor Analysis; CRC Press, 2009.

(88). Gisbrecht A; Schulz A; Hammer B Parametric nonlinear dimensionality reduction using kernel t-SNE. Neurocomputing 2015, 147, 71–82.

(89). McInnes L; Healy J; Melville J Umap: Uniform manifold approximation and projection for dimension reduction. arXiv 2018, arXiv.1802.03426.

(90). Hozumi Y; Wang R; Wei G-W CCP: Correlated Clustering and Projection for Dimensionality Reduction. arXiv 2022, arXiv.2206.04189.

(91). Karnati KR; Wang Y Structural and binding insights into HIV-1 protease and P2-ligand interactions through molecular dynamics simulations, binding free energy and principal component analysis. J. Mol. Graphics Modell. 2019, 92, 112–122.

(92). Bort W; Baskin II; Gimadiev T; Mukanov A; Nugmanov R; Sidorov P; Marcou G; Horvath D; Klimchuk O; Madzhidov T; et al. Discovery of novel chemical reactions by deep generative recurrent neural network. Sci. Rep. 2021, 11, 3178. [PubMed: 33542271]

(93). Samuel AL Some studies in machine learning using the game of checkers. IBM J. Res. Dev. 1959, 3, 210–229.

(94). Cover T Estimation by the nearest neighbor rule. IEEE Trans. Inf. Theory 1968, 14, 50–55.

(95). Cortes C; Vapnik V Support-vector networks. Mach. Learn. 1995, 20, 273–297.

(96). Ho TK Random decision forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, 1995; pp 278–282.

(97). Helma C; Cramer T; Kramer S; De Raedt L Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of non-congeneric compounds. J. Chem. Inf. Comput. Sci. 2004, 44, 1402–1411. [PubMed: 15272848]

(98). Kavakiotis I; Tsave O; Salifoglou A; Maglaveras N; Vlahavas I; Chouvarda I Machine learning and data mining methods in diabetes research. Comput. Struct. Biotechnol. J. 2017, 15, 104–116. [PubMed: 28138367]

(99). Ball NM; Brunner RJ Data mining and machine learning in astronomy. Int. J. Mod. Phys. D 2010, 19, 1049–1106.

(100). Iniesta R; Stahl D; McGuffin P Machine learning, statistical learning and the future of biological research in psychiatry. Psychol. Med. 2016, 46, 2455–2465. [PubMed: 27406289]

(101). Hothorn T CRAN Task View; Machine Learning & Statistical Learning. 2022,

(102). Khan AI; Al-Habsi S Machine learning in computer vision. Procedia Comput. Sci. 2020, 167, 1444–1451.

(103). Huang M; Ninić J; Zhang Q BIM, machine learning and computer vision techniques in underground construction: Current status and future perspectives. Tunnelling Underground Space Technol. 2021, 108, 103677.

(104). Silahtaroğlu G; Yılmaztürk N. Data analysis in health and big data: a machine learning medical diagnosis model based on patients complaints. Commun. Stat.-Theory Methods 2021, 50, 1547–1556.

(105). Alakus TB; Turkoglu I Comparison of deep learning approaches to predict COVID-19 infection. Chaos, Solitons Fractals 2020, 140, 110120. [PubMed: 33519109]

(106). Shorten C; Khoshgoftaar TM; Furht B Deep Learning applications for COVID-19. J. Big Data 2021, 8, 18. [PubMed: 33457181]

(107). Yu K; Tan L; Lin L; Cheng X; Yi Z; Sato T Deeplearning-empowered breast cancer auxiliary diagnosis for 5GB remote E-health. IEEE Wireless Commun. 2021, 28, 54–61.

(108). Harmon SA; Sanford TH; Xu S; Turkbey EB; Roth H; Xu Z; Yang D; Myronenko A; Anderson V; Amalou A; et al. Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets. Nat. Commun. 2020, 11, 4080. [PubMed: 32796848]

(109). Oh Y; Park S; Ye JC Deep learning COVID-19 features on CXR using limited training data sets. IEEE Trans. Med. Imaging 2020, 39, 2688–2700. [PubMed: 32396075]

(110). Ardakani AA; Kanafi AR; Acharya UR; Khadem N; Mohammadi A Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks. Comput. Biol. Med. 2020, 121, 103795. [PubMed: 32568676]

(111). Zhou S; Li G-B; Huang L-Y; Xie H-Z; Zhao Y-L; Chen Y-Z; Li L-L; Yang S-Y A prediction model of drug-induced ototoxicity developed by an optimal support vector machine (SVM) method. Comput. Biol. Med. 2014, 51, 122–127. [PubMed: 24907415]

(112). Rish I An empirical study of the naive Bayes classifier. IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence; IJCAI, 2001; pp 41–46.

(113). Zhang H; Singer BH Recursive Partitioning and Applications; Springer Science & Business Media, 2010.

(114). Siemers FM; Feldmann C; Bajorath J Minimal data requirements for accurate compound activity prediction using machine learning methods of different complexity. Cell Rep. Phys. Sci. 2022, 3, 101113.

(115). Yu F; Li B; Sun J; Qi J; De Wilde RL; Torres-de la Roche LA; Li C; Ahmad S; Shi W; Li X; et al. PSRR: A Web Server for Predicting the Regulation of miRNAs Expression by Small Molecules. Front. Mol. Biosci. 2022, 9, 817294. [PubMed: 35386297]

(116). Albuquerque M; Gerassis S; Sierra C; Taboada J; Martín J; Antunes IMHR; Gallego J Developing a new Bayesian Risk Index for risk evaluation of soil contamination. Sci. Total Environ. 2017, 603, 167–177. [PubMed: 28624637]

(117). James G; Witten D; Hastie T; Tibshirani R An introduction to statistical learning; Springer, 2021; pp 367–402.

(118). Oliveira J; Nogueira D; Ferreira C; Jorge AM; Coimbra M The robustness of Random Forest and Support Vector Machine Algorithms to a Faulty Heart Sound Segmentation. In 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); IEEE, 2022; pp 1989–1992.

(119). Zakariah M Classification of large datasets using Random Forest Algorithm in various applications: Survey. Int. J. Eng. Innov. Technol. 2014, 4, 189–198.

(120). Oshiro TM; Perez PS; Baranauskas JA How many trees in a random forest? In International Workshop on Machine Learning and Data Mining in Pattern Recognition, 2012; pp 154–168.

(121). McCulloch WS; Pitts W Bull. Math. Biophys. Bull. Math. Biophys. 1943, 5, 115–133.

(122). Le TH Applying artificial neural networks for face recognition. Adv. Artif. Neural Syst. 2011, 2011, 673016.

(123). Zhang M; Fulcher J Face recognition using artificial neural network group-based adaptive tolerance (GAT) trees. IEEE Trans. Neural Networks 1996, 7, 555–567. [PubMed: 18263454]

(124). Nazeer SA; Omar N; Khalid M Face recognition system using artificial neural networks approach. In 2007 International Conference on Signal Processing, Communications and Networking, 2007; pp 420–425.

(125). Amato F; López A; Peña-Méndez EM; Vaňhara P; Hampl A; Havel J. Artificial neural networks in medical diagnosis. J. Appl. Biomed. 2013, 11, 47–58.

(126). Zhou Z-H; Jiang Y Medical diagnosis with C4. 5 rule preceded by artificial neural network ensemble. IEEE Trans. Inf. Technol. Biomed. 2003, 7, 37–42. [PubMed: 12670017]

(127). Tourassi GD; Floyd CE The effect of data sampling on the performance evaluation of artificial neural networks in medical diagnosis. Med. Decis. Making 1997, 17, 186–192. [PubMed: 9107614]

(128). Dede G; Sazlı MH Speech recognition with artificial neural networks. Digital Signal Process 2010, 20, 763–768.

(129). Lim CP; Woo SC; Loh AS; Osman R Speech recognition using artificial neural networks. In Proceedings of the First International Conference on Web Information Systems Engineering, 2000; pp 419–423.

(130). Olson M; Wyner A; Berk R Modern neural networks generalize on small data sets. In Advances in Neural Information Processing Systems, 2018; Vol. 31, p 3623–3632.

(131). Chen Y-K; Shave S; Auer M Mrlogp: transfer learning enables accurate logp prediction using small experimental training datasets. Processes 2021, 9, 2029.

(132). Hoseini Ahari SMM; Mirzaei M The artificial neural network-based QSPR and DFT prediction of lipophilicity for thioguanine. Main Group Chem. 2022, 21, 1091–1103.

(133). Dadfar E; Shafiei F; Isfahani TM Structural Relationship Study of Octanol-Water Partition Coefficient of Some Sulfa Drugs Using GA-MLR and GA-ANN Methods. Curr. Comput.-Aided Drug Des. 2020, 16, 207–221. [PubMed: 32507103]

(134). Mamada H; Iwamoto K; Nomura Y; Uesawa Y Predicting blood-to-plasma concentration ratios of drugs from chemical structures and volumes of distribution in humans. Mol. Diversity 2021, 25, 1261–1270.

(135). Mayer AE; Krasnikov VS; Pogorelko VV Homogeneous nucleation of dislocations in copper: Theory and approximate description based on molecular dynamics and artificial neural networks. Comput. Mater. Sci. 2022, 206, 111266.

(136). Mahmood A; Irfan A; Wang J-L Developing efficient small molecule acceptors with sp2-hybridized nitrogen at different positions by density functional theory calculations, molecular

dynamics simulations and machine learning. Chem.-Eur. J. 2022, 28, No. e202103712. [PubMed: 34767281]

(137). Yang GR; Wang X-J Artificial neural networks for neuroscientists: A primer. Neuron 2020, 107, 1048–1070. [PubMed: 32970997]

(138). Tabbussum R; Dar AQ Performance evaluation of artificial intelligence paradigmsartificial neural networks, fuzzy logic, and adaptive neuro-fuzzy inference system for flood prediction. Environ. Sci. Pollut. Res. 2021, 28, 25265–25282.

(139). Wu J; Wang Z A hybrid model for water quality prediction based on an artificial neural network, wavelet transform, and long short-term memory. Water 2022, 14, 610.

(140). Huang Y Advances in artificial neural networks–methodological development and application. Algorithms 2009, 2, 973–1007.

(141). Abbass HA An evolutionary artificial neural networks approach for breast cancer diagnosis. Artif. Intell. Med. 2002, 25, 265–281. [PubMed: 12069763]

(142). Benítez JM; Castro JL; Requena I Are artificial neural networks black boxes? IEEE Trans. Neural Networks 1997, 8, 1156–1164. [PubMed: 18255717]

(143). Vaz JM; Balaji S Convolutional neural networks (CNNs): Concepts and applications in pharmacogenomics. Mol. Diversity 2021, 25, 1569–1584.

(144). Hubel DH; Wiesel TN Shape and arrangement of columns in cat's striate cortex. J Physiol. 1963, 165, 559. [PubMed: 13955384]

(145). LeCun Y; Bottou L; Bengio Y; Haffner P Gradient-based learning applied to document recognition. Proc. IEEE 1998, 86, 2278–2324.

(146). Krizhevsky A; Sutskever I; Hinton GE Imagenet classification with deep convolutional neural networks. Commun. ACM 2017, 60, 84–90.

(147). Simonyan K; Zisserman A Very deep convolutional networks for large-scale image recognition. arXiv 2014, arXiv.1409.1556 DOI: 10.48550/arXiv.1409.1556.

(148). Szegedy C; Liu W; Jia Y; Sermanet P; Reed S; Anguelov D; Erhan D; Vanhoucke V; Rabinovich A Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition; IEEE, 2015; pp 1–9.

(149). Huang G; Liu Z; Van Der Maaten L; Weinberger KQ Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; IEEE, 2017; pp 4700–4708.

(150). Xue L; Tang B; Chen W; Luo J Prediction of CRISPR sgRNA activity using a deep convolutional neural network. J. Chem. Inf. Model. 2019, 59, 615–624. [PubMed: 30485088]

(151). Chen L; Li S; Bai Q; Yang J; Jiang S; Miao Y Review of image classification algorithms based on convolutional neural networks. Remote Sens. 2021, 13, 4712.

(152). Naranjo-Torres J; Mora M; Hernández-García R; Barrientos RJ; Fredes C; Valenzuela A A review of convolutional neural network applied to fruit image processing. Appl. Sci. 2020, 10, 3443.

(153). Anwar SM; Majid M; Qayyum A; Awais M; Alnowami M; Khan MK Medical image analysis using convolutional neural networks: a review. J. Med. Syst. 2018, 42, 226. [PubMed: 30298337]

(154). Gatys LA; Ecker AS; Bethge M Image style transfer using convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition; IEEE, 2016; pp 2414–2423.

(155). Meglouli H; Bentabet L; Airouche M A new technique based on 3D convolutional neural networks and filtering optical flow maps for action classification in infrared video. J. Control Eng. Appl. Inform. 2019, 21, 43–50.

(156). Yao G; Lei T; Zhong J A review of convolutional-neural-network-based action recognition. Pattern Recognit. Lett. 2019, 118, 14–22.

(157). Liu Z; Zhang C; Tian Y 3D-based deep convolutional neural network for action recognition with depth sequences. Image Vision Comput. 2016, 55, 93–100.

(158). Song P; Geng C; Li Z Research on text classification based on convolutional neural network. In 2019 International Conference on Computer Network, Electronic and Automation (ICCNEA), 2019; pp 229–232.

(159). Giménez M; Palanca J; Botti V Semantic-based padding in convolutional neural networks for improving the performance in natural language processing. A case of study in sentiment analysis. Neurocomputing 2020, 378, 315–323.

(160). Albawi S; Mohammed TA; Al-Zawi S Understanding of a convolutional neural network. In 2017 International Conference on Engineering and Technology (ICET), 2017; pp 1–6.

(161). Collobert R; Weston J A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th International Conference on Machine Learning, 2008; pp 160–167.

(162). Sadoughi M; Hu C Physics-based convolutional neural network for fault diagnosis of rolling element bearings. IEEE Sens. J. 2019, 19, 4181–4192.

(163). Zhao X; Gong Z; Zhang Y; Yao W; Chen X Physicsinformed convolutional neural networks for temperature field prediction of heat source layout without labeled data. Eng. Appl. Artif. Intell. 2023, 117, 105516.

(164). Madrazo CF; Heredia I; Lloret L; de Lucas JM Application of a Convolutional Neural Network for Image Classification for the Analysis of Collisions in High Energy Physics; EPJ Web of Conferences, 2019; p 06017.

(165). Hu S; Chen P; Gu P; Wang B A deep learning-based chemical system for QSAR prediction. IEEE J. Biomed. Health. Inf. 2020, 24, 3020–3028.

(166). Karpov P; Godin G; Tetko IV Transformer-CNN: Swiss knife for QSAR modeling and interpretation. J. Cheminform. 2020, 12, 17. [PubMed: 33431004]

(167). Hamza H; Nasser M; Salim N; Saeed F Bioactivity prediction using convolutional neural network. InEmerging Trends in Intelligent Computing and Informatics: Data Science, Intelligent Information Systems and Smart Computing; Advances in Intelligent Systems and Computing; Springer International: Cham, 2020; Vol. 4, pp 341–351.

(168). Nguyen-Vo T-H; Nguyen L; Do N; Le PH; Nguyen T-N; Nguyen BP; Le L Predicting drug-induced liver injury using convolutional neural network and molecular fingerprint-embedded features. ACS Omega 2020, 5, 25432–25439. [PubMed: 33043223]

(169). Zhong S; Hu J; Yu X; Zhang H Molecular image-convolutional neural network (CNN) assisted QSAR models for predicting contaminant reactivity toward OH radicals: Transfer learning, data augmentation and model interpretation. Chem. Eng. J. 2021, 408, 127998.

(170). Zhong S; Zhang K; Wang D; Zhang H Shedding light on Black Box machine learning models for predicting the reactivity of HO radicals toward organic compounds. Chem. Eng. J. 2021, 405, 126627.

(171). Hammes-Schiffer S; Tully JC Proton transfer in solution: Molecular dynamics with quantum transitions. J. Chem. Phys. 1994, 101, 4657–4667.

(172). Li G; Guo Y; Mabuchi T; Surblys D; Ohara T; Tokumasu T Prediction of the adsorption properties of liquid at solid surfaces with molecular scale surface roughness via encoding-decoding convolutional neural networks. J. Mol. Liq. 2022, 349, 118489.

(173). Sun X; Ma L; Du X; Feng J; Dong K Deep convolution neural networks for drug-drug interaction extraction. In IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2018; IEEE, 2018; pp 1662–1668.

(174). Han R; Yang Y; Li X; Ouyang D Predicting oral disintegrating tablet formulations by neural network techniques. Asian J. Pharm. Sci. 2018, 13, 336–342. [PubMed: 32104407]

(175). Meyer JG; Liu S; Miller IJ; Coon JJ; Gitter A Learning drug functions from chemical structures with convolutional neural networks and random forests. J. Chem. Inf. Model. 2019, 59, 4438–4449. [PubMed: 31518132]

(176). Senior AW; Evans R; Jumper J; Kirkpatrick J; Sifre L; Green T; Qin C; Žídek A; Nelson AW; Bridgland A; et al. Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). Proteins: Struct., Funct., Bioinf. 2019, 87, 1141–1148.

(177). Hernández-García A; König P Further advantages of data augmentation on convolutional neural networks. International Conference on Artificial Neural Networks 2018, 11139, 95–103.

(178). Yamashita R; Nishio M; Do RKG; Togashi K Convolutional neural networks: an overview and application in radiology. Insights into imaging 2018, 9, 611–629. [PubMed: 29934920]

(179). Ma S; Zhang Z OmicsMapNet: Transforming omics data to take advantage of Deep Convolutional Neural Network for discovery. arXiv 2018, arXiv.1804.05283

(180). Liu S; Deng W Very deep convolutional neural network based image classification using small training sample size. In 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR); IAPR, 2015; pp 730–734.

(181). Kamilaris A; Prenafeta-Boldu FX. A review of the use of convolutional neural networks in agriculture. J Agric Sci. 2018, 156, 312–322.

(182). Islam MA; Jia S; Bruce ND How Much Position Information Do Convolutional Neural Networks Encode? arXiv 2020, arXiv.2001.08248.

(183). Mohakud R; Dash R Intelligent and Cloud Computing; Springer, 2021; pp 737–744.

(184). Feng D; Haase-Schütz C; Rosenbaum L; Hertlein H; Glaeser C; Timm F; Wiesbeck W; Dietmayer K. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. IEEE Trans. Intell. Transp. Syst. 2021, 22, 1341–1360.

(185). Li J; Jiang F; Yang J; Kong B; Gogate M; Dashtipour K; Hussain A Lane-deeplab: Lane semantic segmentation in automatic driving scenarios for high-definition maps. Neurocomputing 2021, 465, 15–25.

(186). Asad MH; Bais A Weed density estimation using semantic segmentation. In Pacific-Rim Symposium on Image and Video Technology, 2020; pp 162–171.

(187). Long J; Shelhamer E; Darrell T Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; IEEE, 2015; pp 3431–3440.

(188). Ronneberger O; Fischer P; Brox T U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, 2015; pp 234–241.

(189). Badrinarayanan V; Handa A; Cipolla R Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. arXiv 2015, arXiv.1505.07293.

(190). Yu F; Koltun V Multi-scale context aggregation by dilated convolutions. arXiv 2015, arXiv.1511.07122.

(191). Chen L-C; Papandreou G; Kokkinos I; Murphy K; Yuille AL Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. 2018, 40, 834–848. [PubMed: 28463186]

(192). Zhou Z; Rahman Siddiquee MM; Tajbakhsh N; Liang J Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support; Springer, 2018; pp 3–11.

(193). Yan X; Lu Y; Li Z; Wei Q; Gao X; Wang S; Wu S; Cui S PointSite: A Point Cloud Segmentation Tool for Identification of Protein Ligand Binding Atoms. J. Chem. Inf. Model. 2022, 62, 2835–2845. [PubMed: 35621730]

(194). Ibtehaz N; Rahman MS MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. Neural networks 2020, 121, 74–87. [PubMed: 31536901]

(195). Al-Shaebi Z; Uysal Ciloglu F; Nasser M; Aydin O Highly Accurate Identification of Bacterias Antibiotic Resistance Based on Raman Spectroscopy and U-Net Deep Learning Algorithms. ACS omega 2022, 7, 29443–29451. [PubMed: 36033656]

(196). Pfab J; Phan NM; Si D DeepTracer for fast de novo cryo-EM protein structure modeling and special studies on CoV-related complexes. Proc. Natl. Acad. Sci. 2021, 118, No. e2017525118. [PubMed: 33361332]

(197). Zhang X; Zhang B; Freddolino PL; Zhang Y CR-I-TASSER: assemble protein structures from cryo-EM density maps using deep convolutional neural networks. Nat. Methods 2022, 19, 195–204. [PubMed: 35132244]

(198). Pan Z; Xu J; Guo Y; Hu Y; Wang G Deep learning segmentation and classification for urban village using a worldview satellite image based on U-Net. Remote Sens. 2020, 12, 1574.

(199). Lin D; Li Y; Prasad S; Nwe TL; Dong S; Oo ZM CAM-guided Multi-Path Decoding U-Net with Triplet Feature Regularization for defect detection and segmentation. Knowledge-Based Syst. 2021, 228, 107272.

(200). Nazem F; Ghasemi F; Fassihi A; Dehnavi AM 3D U-Net: A voxel-based method in binding site prediction of protein structure. J. Bioinf. Comput. Biol. 2021, 19, 2150006.

(201). Kotowski K; Smolarczyk T; Roterman-Konieczna I; Stapor K ProteinUnetAn efficient alternative to SPIDER3-single for sequence-based prediction of protein secondary structures. J. Comput. Chem. 2021, 42, 50–59. [PubMed: 33058261]

(202). Prasad PJR; Elle OJ; Lindseth F; Albregtsen F; Kumar RP Modifying U-Net for small dataset: a simplified U-Net version for liver parenchyma segmentation. In Medical Imaging 2021: Computer-Aided Diagnosis, 2021; pp 396–405.

(203). Isensee F; Petersen J; Klein A; Zimmerer D; Jaeger PF; Kohl S; Wasserthal J; Koehler G; Norajitra T; Wirkert S, et al. nnU-Net: Self-adapting framework for u-net-based medical image segmentation. arXiv 2018, arXiv.1809.10486.

(204). Zhang J; Jin Y; Xu J; Xu X; Zhang Y MDU-Net: Multi-scale densely connected u-net for biomedical image segmentation. arXiv 2018, arXiv.1812.00352 DOI: 10.48550/arXiv.1812.00352.

(205). Tong G; Li Y; Chen H; Zhang Q; Jiang H Improved U-NET network for pulmonary nodules segmentation. Optik 2018, 174, 460–469.

(206). Wu Z; Lu T; Zhang Y; Wang B; Zhao X Crack detecting by recursive attention U-Net. In 2020 3rd International Conference on Robotics, Control and Automation Engineering (RCAE0, 2020; pp 103–107.

(207). Wang W; Yu K; Hugonot J; Fua P; Salzmann M Recurrent U-Net for resource-constrained segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision; IEEE, 2019; pp 2142–2151.

(208). Du G; Cao X; Liang J; Chen X; Zhan Y Medical image segmentation based on u-net: A review. J. Imaging Sci. Technol. 2020, 64, 20508–1.

(209). Wang JL; Farooq H; Zhuang H; Ibrahim AK Segmentation of intracranial hemorrhage using semi-supervised multi-task attention-based U-net. Appl. Sci. 2020, 10, 3297.

(210). Ryu SM; Shin K; Shin SW; Lee S; Kim N Enhancement of evaluating flatfoot on a weight-bearing lateral radiograph of the foot with U-Net based semantic segmentation on the long axis of tarsal and metatarsal bones in an active learning manner. Comput. Biol. Med. 2022, 145, 105400. [PubMed: 35358752]

(211). Dan H-C; Zeng H-F; Zhu Z-H; Bai G-W; Cao W Methodology for Interactive Labeling of Patched Asphalt Pavement Images Based on U-Net Convolutional Neural Network. Sustainability 2022, 14, 861.

(212). Gori M; Monfardini G; Scarselli F A new model for learning in graph domains. In Proceedings of the 2005 IEEE International Joint Conference on Neural Networks; IEEE, 2005; pp 729–734.

(213). Scarselli F; Gori M; Tsoi AC; Hagenbuchner M; Monfardini G The graph neural network model. IEEE Trans. Neural Networks 2009, 20, 61–80. [PubMed: 19068426]

(214). Micheli A Neural network for graphs: A contextual constructive approach. IEEE Trans. Neural Networks 2009, 20, 498–511. [PubMed: 19193509]

(215). Wu S; Sun F; Zhang W; Xie X; Cui B Graph neural networks in recommender systems: a survey. ACM. Comput. Surv. 2023, 55, 97.

(216). Ying R; He R; Chen K; Eksombatchai P; Hamilton WL; Leskovec J Graph convolutional neural networks for web-scale recommender systems. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2018; pp 974–983.

(217). Pradhyumna P; Shreya G Graph neural network (GNN) in image and video understanding using deep learning for computer vision applications. In 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), 2021; pp 1183–1189.

(218). Hwang D; Yang S; Kwon Y; Lee KH; Lee G; Jo H; Yoon S; Ryu S Comprehensive study on molecular supervised learning with graph neural networks. J. Chem. Inf. Model. 2020, 60, 5936–5945. [PubMed: 33164522]

(219). Wu L; Chen Y; Shen K; Guo X; Gao H; Li S; Pei J; Long B Graph neural networks for natural language processing: A survey. arXiv 2021, arXiv.2106.06090 DOI: 10.48550/arXiv.2106.06090.

(220). Liu W; Zhang Y; Wang J; He Y; Caverlee J; Chan PP; Yeung DS; Heng P-A Item relationship graph neural networks for e-commerce. IEEE Trans. Neural Networks Learn. Syst. 2022, 33, 4785–4799.

(221). Li Z; Shen X; Jiao Y; Pan X; Zou P; Meng X; Yao C; Bu J Hierarchical bipartite graph neural networks: Towards large-scale e-commerce applications. 2020 IEEE 36th International Conference on Data Engineering (ICDE); IEEE, 2020; pp 1677–1688.

(222). Low K; Coote ML; Izgorodina EI Explainable Solvation Free Energy Prediction Combining Graph Neural Networks with Chemical Intuition. J. Chem. Inf. Model. 2022, 62, 5457–5470. [PubMed: 36317829]

(223). Coley CW; Jin W; Rogers L; Jamison TF; Jaakkola TS; Green WH; Barzilay R; Jensen KF A graph-convolutional neural network model for the prediction of chemical reactivity. Chem. Sci. 2019, 10, 370–377. [PubMed: 30746086]

(224). Holm AN; Plank B; Wright D; Augenstein I Longitudinal citation prediction using temporal graph neural networks. arXiv 2020, arXiv.2012.05742 DOI: 10.48550/arXiv.2012.05742.

(225). Gong L; Cheng Q Exploiting edge features for graph neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; IEEE, 2019; pp 9211–9219.

(226). Kipf TN; Welling M Semi-supervised classification with graph convolutional networks. arXiv 2016, arXiv.1609.02907.

(227). Kipf TN; Welling M Variational graph auto-encoders. arXiv 2016, arXiv.1611.07308.

(228). Wu Z; Pan S; Chen F; Long G; Zhang C; Yu PS A comprehensive survey on graph neural networks. IEEE Trans. Neural Networks Learn. Syst. 2021, 32, 4–24.

(229). Wang Y; Abuduweili A; Yao Q; Dou D Property-aware relation networks for few-shot molecular property prediction arXiv 2021, arXiv:2107.07994.

(230). Wu Z; Ramsundar B; Feinberg EN; Gomes J; Geniesse C; Pappu AS; Leswing K; Pande V MoleculeNet: a benchmark for molecular machine learning. Chem. Sci. 2018, 9, 513–530. [PubMed: 29629118]

(231). Pappu A; Paige B Making graph neural networks worth it for low-data molecular machine learning. arXiv 2020, arXiv.2011.12203.

(232). Guo Z; Zhang C; Yu W; Herr J; Wiest O; Jiang M; Chawla NV Few-shot graph learning for molecular property prediction. In Proceedings of the Web Conference 2021; Vol. 2021, pp 2559–2567.

(233). Maddhuri Venkata Subramaniya SR; Terashi G; Jain A; Kagaya Y; Kihara D Protein contact map refinement for improving structure prediction using generative adversarial networks. Bioinformatics 2021, 37, 3168–3174. [PubMed: 33787852]

(234). Balogh OM; Benczik B; Horváth A; Pétervári M; Csermely P; Ferdinandy P; Ágg B Efficient link prediction in the protein–protein interaction network using topological information in a generative adversarial network machine learning model. BMC Bioinform. 2022, 23, 78.

(235). Ishida S; Miyazaki T; Sugaya Y; Omachi S Graph neural networks with multiple feature extraction paths for chemical property estimation. Molecules 2021, 26, 3125. [PubMed: 34073745]

(236). Almasan P; Suárez-Varela J; Rusek K; Barlet-Ros P; Cabellos-Aparicio A Deep reinforcement learning meets graph neural networks: Exploring a routing optimization use case. Comput. Commun. 2022, 196, 184–194.

(237). Chen S; Dong J; Ha P; Li Y; Labi S Graph neural network and reinforcement learning for multi-agent cooperative control of connected autonomous vehicles. Comput.-Aided Civ. Infrastruct. Eng. 2021, 36, 838–857.

(238). Wang Y; Jin W; Derr T Graph Neural Networks: Foundations, Frontiers, and Applications; Springer, 2022; pp 391–420.

(239). Feng W; Zhang J; Dong Y; Han Y; Luan H; Xu Q; Yang Q; Kharlamov E; Tang J Graph random neural networks for semi-supervised learning on graphs. arXiv 2020, arXiv:2005.11079.

(240). Xie Y; Xu Z; Zhang J; Wang Z; Ji S Self-supervised learning of graph neural networks: A unified review. IEEE Trans. Pattern Anal. Mach. Intell. 2023, 45, 2412–2429. [PubMed: 35476575]

(241). Zhu Y; Xu Y; Yu F; Wu S; Wang L CAGNN: Cluster-aware graph neural networks for unsupervised graph representation learning. arXiv 2020, arXiv.2009.01674.

(242). Geisler S; Schmidt T; Şirin H; Zügner D; Bojchevski A; Günnemann S. Robustness of graph neural networks at scale. arXiv2021 arXiv:2110.14038.

(243). Huang Q; Yamada M; Tian Y; Singh D; Chang Y Graphlime: Local interpretable model explanations for graph neural networks. IEEE Trans. Knowl. Data Eng. 2022, 35, 6968–6972.

(244). Ying Z; Bourgeois D; You J; Zitnik M; Leskovec J GNNExplainer: Generating Explanations for Graph Neural Networks. arXiv 2019, arXiv:1903.03894.

(245). Hu W; Liu B; Gomes J; Zitnik M; Liang P; Pande V; Leskovec J Strategies for pre-training graph neural networks. arXiv 2019, arXiv1905.12265.

(246). Loukas A What graph neural networks cannot learn: depth vs width. arXiv 2019, arXiv1907.03199.

(247). Mandal D; Medya S; Uzzi B; Aggarwal C MetaLearning with Graph Neural Networks: Methods and Applications. ACM SIGKDD Explorations Newsletter 2021, 23, 13–22.

(248). Hochreiter S; Schmidhuber J Long short-term memory. Neural Comput. 1997, 9, 1735–1780. [PubMed: 9377276]

(249). Mehryary F; Björne J; Salakoski T; Ginter F Potent pairing: ensemble of long short-term memory networks and support vector machine for chemical-protein relation extraction. Database 2018, 2018, bay120.

(250). Zhang J; Liu J; Luo Y; Fu Q; Bi J; Qiu S; Cao Y; Ding X Chemical substance classification using long short-term memory recurrent neural network. In 2017 IEEE 17th International Conference on Communication Technology (ICCT), 2017; pp 1994–1997.

(251). Awale M; Sirockin F; Stiefl N; Reymond J-L Drug analogs from fragment-based long short-term memory generative neural networks. J. Chem. Inf. Model. 2019, 59, 1347–1356. [PubMed: 30908913]

(252). Jia X; Gavves E; Fernando B; Tuytelaars T Guiding the long-short term memory model for image caption generation. In Proceedings of the IEEE International Conference on Computer Vision; IEEE, 2015; pp 2407–2415.

(253). Balderas D; Ponce P; Molina A Convolutional long short term memory deep neural networks for image sequence prediction. Expert Syst Appl. 2019, 122, 152–162.

(254). Sak H; Senior A; Beaufays F Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. arXiv 2014, arXiv.1402.1128.

(255). Li X; Wu X Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); IEEE, 2015; pp 4520–4524.

(256). Kasthuri E; Balaji S Natural language processing and deep learning chatbot using long short term memory algorithm. Mater. Today: Proc. 2023, 81, 690–693.

(257). Mukherjee A; Su A; Rajan K Deep learning model for identifying critical structural motifs in potential endocrine disruptors. J. Chem. Inf. Model. 2021, 61, 2187–2197. [PubMed: 33872000]

(258). Guo Y; Li W; Wang B; Liu H; Zhou D DeepACLSTM: deep asymmetric convolutional long short-term memory neural models for protein secondary structure prediction. BMC Bioinf. 2019, 20, 341.

(259). Liang D; Zhang Y AC-BLSTM: asymmetric convolutional bidirectional LSTM networks for text classification. arXiv 2016, arXiv.1611.01884.

(260). Pollastri G; Przybylski D; Rost B; Baldi P Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins: Struct., Funct., Bioinf. 2002, 47, 228–235.

(261). Wang S; Peng J; Ma J; Xu J Protein secondary structure prediction using deep convolutional neural fields. Sci. Rep. 2016, 6, 18962. [PubMed: 26752681]

(262). Guo Y; Wang B; Li W; Yang B Protein secondary structure prediction improved by recurrent neural networks integrated with two-dimensional convolutional neural networks. J. Bioinf. Comput. Biol. 2018, 16, 1850021.

(263). Yi H-C; You Z-H; Zhou X; Cheng L; Li X; Jiang T-H; Chen Z-H ACP-DL: a deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation. Mol. Ther.–Nucleic Acids 2019, 17, 1–9. [PubMed: 31173946]

(264). Li S; Chen J; Liu B Protein remote homology detection based on bidirectional long short-term memory. BMC Bioinf. 2017, 18, 443.

(265). Håndstad T; Hestnes AJ; Sætrom P Motif kernel generated by genetic programming improves remote homology and fold detection. BMC Bioinform. 2007, 8, 23.

(266). Liao L; Noble WS Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. J. Comput. Biol. 2003, 10, 857–868. [PubMed: 14980014]

(267). Ben-Hur A; Brutlag D Remote homology detection: a motif based approach. Bioinformatics 2003, 19, i26–i33. [PubMed: 12855434]

(268). Saigo H; Vert J-P; Ueda N; Akutsu T Protein homology detection using string alignment kernels. Bioinformatics 2004, 20, 1682–1689. [PubMed: 14988126]

(269). Altschul SF; Madden TL; Schäffer AA; Zhang J; Zhang Z; Miller W; Lipman DJ Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997, 25, 3389–3402. [PubMed: 9254694]

(270). Hochreiter S; Heusel M; Obermayer K Fast model-based protein homology detection without alignment. Bioinformatics 2007, 23, 1728–1736. [PubMed: 17488755]

(271). Gers FA; Schmidhuber J; Cummins F Learning to forget: Continual prediction with LSTM. Neural Comput. 2000, 12, 2451–2471. [PubMed: 11032042]

(272). Zhu L; Ye C; Hu X; Yang S; Zhu C ACP-check: An anticancer peptide prediction model based on bidirectional long short-term memory and multi-features fusion strategy. Comput. Biol. Med. 2022, 148, 105868. [PubMed: 35868046]

(273). Wang S; Wang X; Wang S; Wang D Bi-directional long short-term memory method based on attention mechanism and rolling update for short-term load forecasting. Int. J. Electr. Power Energy Syst. 2019, 109, 470–479.

(274). Goodfellow Ian J; Jean P-A; Mehdi M; Bing X; David W-F; Sherjil O; Courville Aaron C Generative adversarial nets. In Proceedings of the 27th International Conference on Neural Information Processing Systems, 2014; pp 2672–2680.

(275). Tolstikhin IO; Gelly S; Bousquet O; Simon-Gabriel C-J; Schölkopf B AdaGAN: Boosting generative models. In Advances in Neural Information Processing Systems 30 (NIPS 2017); Curran Associates, 2017; pp 5424–5433

(276). Ghosh A; Kulharia V; Namboodiri VP; Torr PH; Dokania PK Multi-agent diverse generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; Computer Vision Foundation, 2018; pp 8513–8521.

(277). Lin Z; Khetan A; Fanti G; Oh S Pacgan: The power of two samples in generative adversarial networks. In Conference on Neural Information Processing Systems, 2018; pp 1498–1507.

(278). Nguyen T; Le T; Vu H; Phung D Dual discriminator generative adversarial nets. In Advances in Neural Information Processing Systems 30 (NIPS 2017), 2017; pp 2667–2677.

(279). Chavdarova T; Fleuret FS: An alternative training of generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; IEEE, 2018; pp 9407–9415.

(280). Radford A; Metz L; Chintala S Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv 2015, arXiv.1511.06434.

(281). Han X; Zhang L; Zhou K; Wang X ProGAN: Protein solubility generative adversarial nets for data augmentation in DNN framework. Comput. Chem. Eng. 2019, 131, 106533.

(282). Niwa T; Ying B-W; Saito K; Jin W; Takada S; Ueda T; Taguchi H Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins. Proc. Natl. Acad. Sci. 2009, 106, 4201–4206. [PubMed: 19251648]

(283). Marreiros AC; Daunizeau J; Kiebel SJ; Friston KJ Population dynamics: variance and the sigmoid activation function. Neuroimage 2008, 42, 147–157. [PubMed: 18547818]

(284). Liu Y; Zhou Y; Liu X; Dong F; Wang C; Wang Z Wasserstein GAN-based small-sample augmentation for new-generation artificial intelligence: a case study of cancer-staging data in biology. Engineering 2019, 5, 156–163.

(285). Breiman L Random forests. Mach. Learn. 2001, 45, 5–32.

(286). Han H; Wang W-Y; Mao B-H Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. International conference on intelligent computing 2005, 3644, 878–887.

(287). Ribeiro e Sousa LR; Miranda T; Leal e Sousa RL; Tinoco J The use of data mining techniques in rockburst risk assessment. Engineering 2017, 3, 552–558.

(288). Sun Y; Kamel MS; Wong AK; Wang Y Cost-sensitive boosting for classification of imbalanced data. Pattern Recognit. 2007, 40, 3358–3378.

(289). Hsu T-C; Lin C Generative adversarial networks for robust breast cancer prognosis prediction with limited data size. 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); IEEE, 2020; pp 5669–5672.

(290). Li C; Wei Y; Chen X; Schönlieb C-B Deep Generative Models, and Data Augmentation, Labelling, and Imperfections; Springer, 2021; pp 103–111.

(291). Lin T-T; Sun Y-Y; Cheng W-C; Lu I-H; Chen S-H; Lin C-YDeveloping an Antiviral Peptides Predictor with Generative Adversarial Network Data Augmentation. bioRxiv 2021, bio-Rxiv.2021.11.29.470292.

(292). Lee YJ; Kahng H; Kim SB Generative adversarial networks for de novo molecular design. Mol. Inf. 2021, 40, 2100045.

(293). Dan Y; Zhao Y; Li X; Li S; Hu M; Hu J Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse design of inorganic materials. NPJ Comput. Mater. 2020, 6, 84.

(294). Sawada Y; Morikawa K; Fujii M Conditional Generative Adversarial Networks for Inorganic Chemical Compositions. Chem. Lett. 2021, 50, 623–626.

(295). Yi X; Walia E; Babyn P Generative adversarial network in medical imaging: A review. Med. Image Anal. 2019, 58, 101552. [PubMed: 31521965]

(296). Bing X; Zhang W; Zheng L; Zhang Y Medical image super resolution using improved generative adversarial networks. IEEE Access 2019, 7, 145030–145038.

(297). Zhang T; Cheng J; Fu H; Gu Z; Xiao Y; Zhou K; Gao S; Zheng R; Liu J Noise adaptation generative adversarial network for medical image analysis. IEEE Trans. Med. Imaging 2020, 39, 1149–1159. [PubMed: 31567075]

(298). Zhang H; Sindagi V; Patel VM Image de-raining using a conditional generative adversarial network. IEEE Trans. Circuits Syst. Video Technol. 2020, 30, 3943–3956.

(299). Zhu L; Chen Y; Ghamisi P; Benediktsson JA Generative adversarial networks for hyperspectral image classification. IEEE Trans. Geosci. Electron. 2018, 56, 5046–5063.

(300). Lin E; Lin C-H; Lane H-Y De Novo Peptide and Protein Design Using Generative Adversarial Networks: An Update. J. Chem. Inf. Model. 2022, 62, 761–774. [PubMed: 35128926]

(301). Kong J; Kim J; Bae J HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In Advances in Neural Information Processing Systems 33 (NeurIPS 2020)2020, Vol. 33, pp 17022–17033

(302). Mira R; Vougioukas K; Ma P; Petridis S; Schuller BW; Pantic M End-to-end video-to-speech synthesis using generative adversarial networks. IEEE Trans. Cybern. 2023, 53, 3454–3466. [PubMed: 35439155]

(303). Tian Q; Chen Y; Zhang Z; Lu H; Chen L; Xie L; Liu S TFGAN: Time and frequency domain based generative adversarial network for high-fidelity speech synthesis. arXiv 2020, arXiv.2011.12206.

(304). Cao Y-J; Jia L-L; Chen Y-X; Lin N; Yang C; Zhang B; Liu Z; Li X-X; Dai H-H Recent advances of generative adversarial networks in computer vision. IEEE Access 2019, 7, 14985–15006.

(305). Park S-W; Ko J-S; Huh J-H; Kim J-C Review on generative adversarial networks: focusing on computer vision and its applications. Electronics 2021, 10, 1216.

(306). Sampath V; Maurtua I; Aguilar Martín JJ; Gutierrez A A survey on generative adversarial networks for imbalance problems in computer vision tasks. J. Big Data 2021, 8, 27. [PubMed: 33552840]

(307). Mishra D; Prathosh AP; Jayendran A; Srivastava V; Chaudhury S Mode matching in GANs through latent space learning and inversion. arXiv 2018, arXiv.1811.03692.

(308). Kingma DP; Welling M Auto-encoding variational Bayes. arXiv 2013, arXiv.1312.6114.

(309). Makhzani A; Shlens J; Jaitly N; Goodfellow I; Frey B Adversarial autoencoders. arXiv 2015, arXiv.1511.05644.

(310). Bao J; Chen D; Wen F; Li H; Hua G CVAE-GAN: fine-grained image generation through asymmetric training. Proceedings of the IEEE International Conference on Computer Vision; IEEE, 2017; pp 2745–2754.

(311). Chen X; Kingma DP; Salimans T; Duan Y; Dhariwal P; Schulman J; Sutskever I; Abbeel P Variational lossy autoencoder. arXiv, 2016, arXiv.1611.02731.

(312). Cai L; Gao H; Ji S Multi-stage variational auto-encoders for coarse-to-fine image generation. Proceedings of the 2019 SIAM International Conference on Data Mining; SIAM, 2019; pp 630–638.

(313). Tolstikhin I; Bousquet O; Gelly S; Schoelkopf B Wasserstein auto-encoders. arXiv, 2017, arXiv.1711.01558.

(314). Ma C; Zhang X GF-VAE: A Flow-based Variational Autoencoder for Molecule Generation. Proceedings of the 30th ACM International Conference on Information & Knowledge Management; ACM, 2021; pp 1181–1190.

(315). Feng H; Gao K; Chen D; Shen L; Robison AJ; Ellsworth E; Wei G-W Machine learning analysis of cocaine addiction informed by DAT, SERT, and NET-based interactome networks. J. Chem. Theory Comput. 2022, 18, 2703–2719. [PubMed: 35294204]

(316). Gómez-Bombarelli R; Wei JN; Duvenaud D; Hernández-Lobato JM; Sánchez-Lengeling B; Sheberla D; Aguilera-Iparraguirre J; Hirzel TD; Adams RP; Aspuru-Guzik A Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent. Sci. 2018, 4, 268–276. [PubMed: 29532027]

(317). Glushkovsky A AI discovering a coordinate system of chemical elements: dual representation by variational autoencoders. arXiv 2020, arXiv.2011.12090.

(318). Gircha A; Boev AS; Avchaciov K; Fedichev P; Fedorov AK Training a discrete variational autoencoder for generative chemistry and drug design on a quantum annealer. arXiv 2021, arXiv.2108.11644.

(319). Gregor K; Danihelka I; Graves A; Rezende D; Wierstra D Draw: A recurrent neural network for image generation. International Conference on Machine Learning, 2015; pp 1462–1471.

(320). Bowman SR; Vilnis L; Vinyals O; Dai AM; Jozefowicz R; Bengio S Generating sentences from a continuous space. arXiv 2015, arXiv.1511.06349.

(321). Jang M; Seo S; Kang P Recurrent neural network-based semantic variational autoencoder for sequence-to-sequence learning. Inf. Sci. 2019, 490, 59–73.

(322). Liu X; Zhang F; Hou Z; Mian L; Wang Z; Zhang J; Tang J Self-supervised learning: Generative or contrastive. IEEE Trans. Knowl. Data Eng. 2023, 35, 857–876.

(323). Bachman P; Hjelm RD; Buchwalter W Learning representations by maximizing mutual information across views. Conference on Neural Information Processing Systems, 2019; pp 15535–15545.

(324). Devlin J; Chang M-W; Lee K; Toutanova K BERT: Pretraining of deep bidirectional transformers for language understanding. arXiv 2018, arXiv.1810.04805.

(325). Chithrananda S; Grand G; Ramsundar B ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. arXiv 2020, arXiv.2010.09885.

(326). Rong Y; Bian Y; Xu T; Xie W; Wei Y; Huang W; Huang J Self-supervised graph transformer on large-scale molecular data. InAdvances in Neural Information Processing Systems 33 (NeurIPS 2020), 2020; Vol. 33, pp 12559–12571

(327). Li P; Wang J; Qiao Y; Chen H; Yu Y; Yao X; Gao P; Xie G; Song S An effective self-supervised framework for learning expressive molecular global representations to drug discovery. Briefings Bioinform. 2021, 22, bbab109.

(328). Mo S; Fu X; Hong C; Chen Y; Zheng Y; Tang X; Shen Z; Xing EP; Lan Y Multi-modal Self-supervised Pre-training for Regulatory Genome Across Cell Types. arXiv 2021, arXiv.2110.05231.

(329). He K; Fan H; Wu Y; Xie S; Girshick R Momentum contrast for unsupervised visual representation learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; IEEE, 2020; pp 9729–9738.

(330). Chen T; Kornblith S; Norouzi M; Hinton G A simple framework for contrastive learning of visual representations. International Conference on Machine Learning. 2020; pp 1597–1607.

(331). Zhou K; Wang H; Zhao WX; Zhu Y; Wang S; Zhang F; Wang Z; Wen J-R S3-Rec: Self-supervised learning for sequential recommendation with mutual information maximization. Proceedings of the 29th ACM International Conference on Information & Knowledge Management; ACM, 2020; pp 1893–1902.

(332). Lopez-del Rio A; Picart-Armada S; Perera-Lluna A Balancing data on deep learning-based proteochemometric activity classification. J. Chem. Inf. Model. 2021, 61, 1657–1669. [PubMed: 33779173]

(333). Sermanet P; Lynch C; Chebotar Y; Hsu J; Jang E; Schaal S; Levine S; Brain G Time-contrastive networks: Self-supervised learning from video. 2018 IEEE International Conference on Robotics and Automation (ICRA); IEEE, 2018; pp 1134–1141.

(334). Chen D; Gao K; Nguyen DD; Chen X; Jiang Y; Wei G-W; Pan F Algebraic graph-assisted bidirectional transformers for molecular property prediction. Nat. Commun. 2021, 12, 3521. [PubMed: 34112777]

(335). Shen X; Liu Y; Wu Y; Xie L MoLGNN: Self-supervised motif learning graph neural network for drug discovery. Machine Learning for Molecules Workshop at NeurIPS, 2020; 1–8.

(336). Zheng J; Qian Y; He J; Kang Z; Deng L Graph Neural Network with Self-Supervised Learning for Noncoding RNA–Drug Resistance Association Prediction. J. Chem. Inf. Model. 2022, 62, 3676–3684. [PubMed: 35838124]

(337). Wu Z; Hruby VJ Backbone Alignment Modeling of the Structure–Activity Relationships of Opioid Ligands. J. Chem. Inf. Model. 2011, 51, 1151–1164. [PubMed: 21488692]

(338). Wu K; Wei G-W Quantitative toxicity prediction using topology based multitask deep neural networks. J. Chem. Inf. Model. 2018, 58, 520–531. [PubMed: 29314829]

(339). Gao K; Nguyen DD; Sresht V; Mathiowetz AM; Tu M; Wei G-W Are 2D fingerprints still valuable for drug discovery? Phys. Chem. Chem. Phys. 2020, 22, 8373–8390. [PubMed: 32266895]

(340). Karim A; Mishra A; Newton MH; Sattar A Efficient toxicity prediction via simple features using shallow neural networks and decision trees. Acs Omega 2019, 4, 1874–1888.

(341). Jiang J; Wang R; Wang M; Gao K; Nguyen DD; Wei G-W Boosting tree-assisted multitask deep learning for small scientific datasets. J. Chem. Inf. Model. 2020, 60, 1235–1244. [PubMed: 31977216]

(342). Wu K; Zhao Z; Wang R; Wei G-W TopP–S: Persistent homology-based multi-task deep neural networks for simultaneous predictions of partition coefficient and aqueous solubility. J. Comput. Chem. 2018, 39, 1444–1454. [PubMed: 29633287]

(343). Cheng T; Zhao Y; Li X; Lin F; Xu Y; Zhang X; Li Y; Wang R; Lai L Computation of octanol-water partition coefficients by guiding an additive model with knowledge. J. Chem. Inf. Model. 2007, 47, 2140–2148. [PubMed: 17985865]

(344). Yang X; Yang G; Chu J Self-supervised Learning for Label Sparsity in Computational Drug Repositioning. arXiv 2022, arXiv.2206.00262.

(345). Luo H; Wang J; Li M; Luo J; Peng X; Wu F-X; Pan Y Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm. Bioinformatics 2016, 32, 2664–2671. [PubMed: 27153662]

(346). Hirschfeld L; Swanson K; Yang K; Barzilay R; Coley CW Uncertainty quantification using neural networks for molecular property prediction. J. Chem. Inf. Model. 2020, 60, 3770–3780. [PubMed: 32702986]

(347). Li H; Zhao D; Zeng J KPGT: Knowledge-Guided Pretraining of Graph Transformer for Molecular Property Prediction. arXiv 2022, arXiv.2206.03364.

(348). Chithrananda S; Grand G; Ramsundar B ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. arXiv 2020, arXiv.2010.09885.

(349). Cai T; Lim H; Abbu KA; Qiu Y; Nussinov R; Xie L MSA-regularized protein sequence transformer toward predicting genome-wide chemical-protein interactions: application to GPCRome deorphanization. J. Chem. Inf. Model. 2021, 61, 1570–1582. [PubMed: 33757283]

(350). Chen C; Zhou J; Wang F; Liu X; Dou D Structure-aware protein self-supervised learning. arXiv 2022, arXiv.2204.04213 DOI: 10.48550/arXiv.2204.04213.

(351). Sanner MF; Dieguez L; Forli S; Lis E Improving Docking Power for Short Peptides Using Random Forest. J. Chem. Inf. Model. 2021, 61, 3074–3090. [PubMed: 34124893]

(352). Thrun S; Littman ML A Review of Reinforcement Learning. AI Mag. 2000, 21, 103–103.

(353). Szepesvári C Algorithms for reinforcement learning. In Synthesis Lectures on Artificial Intelligence and Machine Learning; Springer International: Cham, 2010; Vol. 4, pp 1–103.

(354). White CC A survey of solution techniques for the partially observed Markov decision process. Ann. Oper. Res. 1991, 32, 215–230.

(355). White DJ A survey of applications of Markov decision processes. J. Oper. Res. Soc. 1993, 44, 1073–1096.

(356). Moerland TM; Broekens J; Jonker CM Model-based reinforcement learning: A survey. arXiv 2020, arXiv.2006.16712.

(357). Çalışır S; Pehlivanoğlu MK. Model-free reinforcement learning algorithms: A survey. In 2019 27th Signal Processing and Communications Applications Conference; SIU, 2019; pp 1–4.

(358). Renaudo E; Girard B; Chatila R; Khamassi M Respective advantages and disadvantages of model-based and model-free reinforcement learning in a robotics neuro-inspired cognitive architecture. Procedia Comput. Sci. 2015, 71, 178–184.

(359). Epshteyn A; Vogel A; DeJong G Active reinforcement learning. Proceedings of the 25th International Conference on Machine Learning. 2008; pp 296–303.

(360). Mitchell TM Machine learning and data mining. Commun. ACM 1999, 42, 30–36.

(361). François-Lavet V; Henderson P; Islam R; Bellemare MG; Pineau J. An introduction to deep reinforcement learning. Found. Trends Mach. Learn. 2018, 11, 219–354.

(362). Lei C Deep Learning and Practice with MindSpore; Springer, 2021; pp 217–243.

(363). Gottipati SK; Pathak Y; Sattarov B; Nuttall R; Amini M; Taylor ME; Chandar S Towered actor critic for handling multiple action types in reinforcement learning for drug discovery. In Proceedings of the AAAI Conference on Artificial Intelligence; AAAI, 2021; pp 142–150.

(364). Padalkar GR; Patil SD; Hegadi MM; Jaybhaye NK Drug discovery using generative adversarial network with reinforcement learning. In 2021 International Conference on Computer Communication and Informatics; ICCCI, 2021; pp 1–3.

(365). Lutz ID; Wang S; Norn C; Borst AJ; Zhao YT; Dosey A; Cao L; Li Z; Baek M; King NP; Ruohola-Baker H; Baker D Top-down design of protein nanomaterials with reinforcement learning. bioRxiv 2022, 2022–09, 2022.09.25.509419.

(366). McNaughton AD; Bontha MS; Knutson CR; Pope JA; Kumar N De novo design of protein target specific scaffold-based Inhibitors via Reinforcement Learning. arXiv 2022, arXiv.2205.10473.

(367). Joy M; Kaisare NS Approximate dynamic programming-based control of distributed parameter systems. Asia-Pac. J. Chem. Eng. 2011, 6, 452–459.

(368). Lee JM; Lee JH Approximate dynamic programming-based approaches for input–output data-driven control of nonlinear processes. Automatica 2005, 41, 1281–1288.

(369). Mousavi HK; Nazari M; Takáč M; Motee N Multi-agent image classification via reinforcement learning. 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); IEEE, 2019; pp 5020–5027.

(370). Baker B; Gupta O; Naik N; Raskar R Designing neural network architectures using reinforcement learning. arXiv 2016, arXiv.1611.02167.

(371). Meng TL; Khushi M Reinforcement learning in financial markets. Data 2019, 4, 110.

(372). Dou H; Tan J; Wei H; Wang F; Yang J; Ma X-G; Wang J; Zhou T Transfer inhibitory potency prediction to binary classification: A model only needs a small training set. Comput. Methods Programs Biomed. 2022, 215, 106633. [PubMed: 35091229]

(373). Cui L; Lu Y; Sun J; Fu Q; Xu X; Wu H; Chen J Rflmda: a novel reinforcement learning-based computational model for human microRNA-disease association prediction. Biomolecules 2021, 11, 1835. [PubMed: 34944479]

(374). Clifton J; Laber E Q-learning: theory and applications. Annu. Rev. Stat. Appl. 2020, 7, 279–301.

(375). Zheng X; Ding H; Mamitsuka H; Zhu S Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Association for Computing Machinery, 2013; pp 1025–1033.

(376). Liu Y; Wu M; Miao C; Zhao P; Li X-L Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. PLoS Comput. Biol. 2016, 12, No. e1004760. [PubMed: 26872142]

(377). Xia Z; Wu L-Y; Zhou X; Wong STC Semi-supervised drug-protein interaction prediction from heterogeneous spaces. BMC Syst. Biol. 2010, 4, 56. [PubMed: 20438628]

(378). Pereira T; Abbasi M; Oliveira JL; Ribeiro B; Arrais J Optimizing blood–brain barrier permeation through deep reinforcement learning for de novo drug design. Bioinformatics 2021, 37, i84–i92. [PubMed: 34252946]

(379). Prathik A; Vinodhini M; Karthik N; Ebenezer V Intelligent Data Communication Technologies and Internet of Things; Springer, 2022; pp 541–552.

(380). Rokhlin V; Szlam A; Tygert M A randomized algorithm for principal component analysis. SIAM J. Matrix Anal. Appl. 2010, 31, 1100–1124.

(381). Wu Y-H; Lin S-D A low-cost ethics shaping approach for designing reinforcement learning agents. Thirty-Second AAAI Conference on Artificial Intelligence 2018, 32, 1687–1694.

(382). Ståhl N; Falkman G; Karlsson A; Mathiason G; Bostrom J Deep reinforcement learning for multiparameter optimization in de novo drug design. J. Chem. Inf. Model. 2019, 59, 3166–3176. [PubMed: 31273995]

(383). Popova M; Isayev O; Tropsha A Deep reinforcement learning for de novo drug design. Sci. Adv. 2018, 4, No. eaap7885. [PubMed: 30050984]

(384). Leibo JZ; d'Autume C. d. M.; Zoran D; Amos D; Beattie C; Anderson K; Castañeda AG; Sanchez M; Green S; Gruslys A, et al. Psychlab: a psychology laboratory for deep reinforcement learning agents. arXiv 2018, arXiv.1801.08116.

(385). Subramanian A; Chitlangia S; Baths V Reinforcement learning and its connections with neuroscience and psychology. Neural Networks 2022, 145, 271–287. [PubMed: 34781215]

(386). Kappen HJ An introduction to stochastic control theory, path integrals and reinforcement learning. In AIP Conference Proceedings, 2007; pp 149–181.

(387). Kretchmar RM A Synthesis of Reinforcement Learning and Robust Control Theory. Ph.D. Thesis. Colorado State University, 2000.

(388). Pachocki J; Brockman G; Raiman J; Zhang S; Pondé H; Tang J; Wolski F; Dennison C; Jozefowicz R; Debiak P, et al. Openai Five, 2018; https://openai.com/research/openai-five.

(389). Silver D; Hubert T; Schrittwieser J; Hassabis D AlphaZero: Shedding New Light on the Grand Games of Chess, Shogi and Go. DeepMind blog 2018; https://www.deepmind.com/blog/alphazero-shedding-new-light-on-chess-shogi-and-go.

(390). Taylor ME; Stone P Transfer learning for reinforcement learning domains: a survey. J. Mach. Learn. Res. 2009, 10, 1633–1685.

(391). Lin L-J Self-improving reactive agents based on reinforcement learning, planning and teaching. Mach. Learn. 1992, 8, 293–321.

(392). Pan SJ; Yang Q A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 2010, 22, 1345–1359.

(393). Liu X; LUAN X; Xie Y; Huang M Transfer learning research and algorithm review. J. Changsha Univ. 2018, 32, 29–36.

(394). Jang Y; Lee H; Hwang SJ; Shin J Learning What and Where to Transfer; International Conference on Machine Learning. 2019; pp 3030–3039.

(395). Li X; Grandvalet Y; Davoine F; Cheng J; Cui Y; Zhang H; Belongie S; Tsai Y-H; Yang M-H Transfer learning in computer vision tasks: Remember where you come from. Image Vision Comput. 2020, 93, 103853.

(396). Brodzicki A; Piekarski M; Kucharski D; Jaworek-Korjakowska J; Gorgon M Transfer learning methods as a new approach in computer vision tasks with small datasets. Found. Comput. Decis. Sci. 2020, 45, 179–193.

(397). Shao L; Zhu F; Li X Transfer learning for visual categorization: A survey. IEEE Trans. Neural Networks Learn. Syst. 2015, 26, 1019–1034.

(398). Liu R; Liu Q; Zhu H; Cao H Multistage Deep Transfer Learning for EmIoT-Enabled Human-Computer Interaction. IEEE Internet Things J. 2022, 9, 15128–15137.

(399). Xiao Z; Wang L; Du J Improving the performance of sentiment classification on imbalanced datasets with transfer learning. IEEE Access. 2019, 7, 28281–28290.

(400). Liu B; Xiao Y; Hao Z A selective multiple instance transfer learning method for text categorization problems. Knowledge-Based Syst. 2018, 141, 178–187.

(401). Zheng D; Zhang C; Fei G; Zhao T Research on text categorization based on a weakly-supervised transfer learning method. International Conference on Intelligent Text Processing and Computational Linguistics, 2012; pp 144–156.

(402). Malmgren-Hansen D; Kusk A; Dall J; Nielsen AA; Engholm R; Skriver H Improving SAR automatic target recognition models with transfer learning from simulated data. IEEE Geosci. Remote Sens. Lett. 2017, 14, 1484–1488.

(403). Wang Z; Du L; Mao J; Liu B; Yang D SAR target detection based on SSD with data augmentation and transfer learning. IEEE Geosci. Remote Sens. Lett. 2019, 16, 150–154.

(404). Du X; Sun S; Hu C; Yao Y; Yan Y; Zhang Y DeepPPI: boosting prediction of protein–protein interactions with deep neural networks. J. Chem. Inf. Model. 2017, 57, 1499–1510. [PubMed: 28514151]

(405). Kozlovskii I; Popov P Protein–peptide binding site detection using 3D convolutional neural networks. J. Chem. Inf. Model. 2021, 61, 3814–3823. [PubMed: 34292750]

(406). Pio G; Mignone P; Magazzù G; Zampieri G; Ceci M; Angione C Integrating genome-scale metabolic modelling and transfer learning for human gene regulatory network reconstruction. Bioinformatics 2022, 38, 487–493. [PubMed: 34499112]

(407). Lopez-Garcia G; Jerez JM; Franco L; Veredas FJ Transfer learning with convolutional neural networks for cancer survival prediction using gene-expression data. PloS one 2020, 15, No. e0230536. [PubMed: 32214348]

(408). Aldayel MS; Ykhlef M; Al-Nafjan AN Electroencephalogram-based preference prediction using deep transfer learning. IEEE Access 2020, 8, 176818–176829.

(409). Kim Y; Zheng S; Tang J; Jim Zheng W; Li Z; Jiang X Anticancer drug synergy prediction in understudied tissues using transfer learning. J. Am. Med. Inf. Assoc. 2021, 28, 42–51.

(410). El-allaly E.-d.; Sarrouti M; En-Nahnahi N; El Alaoui SO. MTTLADE: A multi-task transfer learning-based method for adverse drug events extraction. s. 2021, 58, 102473.

(411). Taroni JN; Grayson PC; Hu Q; Eddy S; Kretzler M; Merkel PA; Greene CS MultiPLIER: a transfer learning framework for transcriptomics reveals systemic features of rare disease. Cell Syst. 2019, 8, 380–394. [PubMed: 31121115]

(412). Ye Z; Yang Y; Li X; Cao D; Ouyang D An integrated transfer learning and multitask learning approach for pharmacokinetic parameter prediction. Mol. Pharmaceutics 2019, 16, 533–541.

(413). Sharifi-Noghabi H; Peng S; Zolotareva O; Collins CC; Ester M AITL: Adversarial Inductive Transfer Learning with input and output space adaptation for pharmacogenomics. Bioinformatics 2020, 36, i380–i388. [PubMed: 32657371]

(414). Snell J; Swersky K; Zemel R Prototypical networks for few-shot learning. In Advances in Neural Information Processing Systems 30 (NIPS 2017), 2017; Vol. 30.

(415). Tzeng E; Hoffman J; Saenko K; Darrell T Adversarial discriminative domain adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017; pp 7167–7176.

(416). Bai R; Zhang C; Wang L; Yao C; Ge J; Duan H Transfer learning: making retrosynthetic predictions based on a small chemical reaction dataset scale to a new level. Molecules 2020, 25, 2357. [PubMed: 32438572]

(417). Liu B; Ramsundar B; Kawthekar P; Shi J; Gomes J; Luu Nguyen Q; Ho S; Sloane J; Wender P; Pande V Retrosynthetic reaction prediction using neural sequence-to-sequence models. ACS Cent. Sci. 2017, 3, 1103–1113. [PubMed: 29104927]

(418). Chen Y-K; Shave S; Auer M MRlogP: transfer learning enables accurate logP prediction using small experimental training datasets. Processes 2021, 9, 2029.

(419). Cang Z; Wei G-W TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions. PLoS Comput. Biol. 2017, 13, No. e1005690. [PubMed: 28749969]

(420). Sakkiah S; Leggett C; Pan B; Guo W; Valerio LG Jr; Hong H. Development of a nicotinic acetylcholine receptor nAChR α7 binding activity prediction model. J. Chem. Inf. Model. 2020, 60, 2396–2404. [PubMed: 32159345]

(421). Imrie F; Bradley AR; van der Schaar M; Deane CM Protein family-specific models using deep neural networks and transfer learning improve virtual screening and highlight the need for more data. J. Chem. Inf. Model. 2018, 58, 2319–2330. [PubMed: 30273487]

(422). Hurtado DM; Uziela K; Elofsson A Deep transfer learning in the assessment of the quality of protein models. arXiv 2018, arXiv.1804.06281.

(423). Turki T; Wei Z; Wang JT Transfer learning approaches to improve drug sensitivity prediction in multiple myeloma patients. IEEE Access 2017, 5, 7381–7393.

(424). Cai C; Wang S; Xu Y; Zhang W; Tang K; Ouyang Q; Lai L; Pei J Transfer learning for drug discovery. J. Med. Chem. 2020, 63, 8683–8694. [PubMed: 32672961]

(425). Kulis B; Saenko K; Darrell T What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. CVPR 2011 2011, 1785–1792.

(426). Duan L; Xu D; Tsang I Learning with augmented features for heterogeneous domain adaptation. arXiv 2012, arXiv.1206.4660.

(427). Zhu Y; Chen Y; Lu Z; Pan S; Xue G-R; Yu Y; Yang Q Heterogeneous transfer learning for image classification. Proceedings of the AAAI Conference on Artificial Intelligence; AAAI, 2011; pp 1304–1309.

(428). Wang C; Mahadevan S Heterogeneous domain adaptation using manifold alignment. IJCAI Proceedings-International Joint Conference on Artificial Intelligence; IJCAI, 2011; p 1541.

(429). Cao Z; Zhou Y; Yang A; Peng S Deep transfer learning mechanism for fine-grained cross-domain sentiment classification. Connect. Sci. 2021, 33, 911–928.

(430). Liu R; Shi Y; Ji C; Jia M A survey of sentiment analysis based on transfer learning. IEEE Access 2019, 7, 85401–85412.

(431). Mahmud M; Ray S Transfer learning using Kolmogorov complexity: Basic theory and empirical evaluations. In Advances in Neural Information Processing Systems 20 (NIPS 2007) 2007; Vol. 20, pp 985–992.

(432). Lewis DD A sequential algorithm for training text classifiers: Corrigendum and additional data. Acm Sigir Forum 1995, 29, 13–19.

(433). Dagan I; Engelson SP Machine Learning Proceedings 1995; Elsevier, 1995; pp 150–157.

(434). Krishnamurthy V Algorithms for optimal scheduling and management of hidden Markov model sensors. IEEE Trans. Signal Process. 2002, 50, 1382–1397.

(435). Zhan X; Liu H; Li Q; Chan AB A Comparative Survey: Benchmarking for Pool-based Active Learning; IJCAI, 2021; pp 4679–4686.

(436). Kelz JI; Takahashi GR; Safizadeh F; Farahmand V; Crosby MG; Uribe JL; Kim SH; Sprague-Piercy MA; Diessner EM; Norton-Baker B; et al. Active Learning Module for Protein Structure Analysis Using Novel Enzymes. The Biophysicist 2022, 3, 49–63.

(437). Kleiman DE; Shukla D Active Learning of the Conformational Ensemble of Proteins using Maximum Entropy VAMPNets. J. Chem. Theory Comput. 2023, DOI: 10.1021/acs.jctc.3c00040.

(438). Shmilovich K; Mansbach RA; Sidky H; Dunne OE; Panda SS; Tovar JD; Ferguson AL Discovery of self-assembling π-conjugated peptides by active learning-directed coarse-grained molecular simulation. J. Phys. Chem. B 2020, 124, 3873–3891. [PubMed: 32180410]

(439). Polash AH; Nakano T; Rakers C; Takeda S; Brown J Active learning efficiently converges on rational limits of toxicity prediction and identifies patterns for molecule design. Comput. Toxicol. 2020, 15, 100129.

(440). Morger A; Garcia de Lomana M; Norinder U; Svensson F; Kirchmair J; Mathea M; Volkamer A Studying and mitigating the effects of data drifts on ML model performance at the example of chemical toxicity data. Sci. Rep. 2022, 12, 7244. [PubMed: 35508546]

(441). Zhang Y; Lee AA Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning. Chem. Sci. 2019, 10, 8154–8163. [PubMed: 31857882]

(442). Liu Q; Wang D Stein variational gradient descent: A general purpose bayesian inference algorithm. In 30th Conference on Neural Information Processing Systems (NIPS 2016), 2016, Vol. 30, pp 2378–2386.

(443). Barrett R; White AD Investigating Active Learning and Meta-Learning for Iterative Peptide Design. J. Chem. Inf. Model. 2021, 61, 95–105. [PubMed: 33350829]

(444). Cicuto CAT; Torres BB Implementing an active learning environment to influence students motivation in biochemistry. J. Chem. Educ. 2016, 93, 1020–1026.

(445). Budd S; Robinson EC; Kainz B A survey on active learning and human-in-the-loop deep learning for medical image analysis. Med. Image Anal. 2021, 71, 102062. [PubMed: 33901992]

(446). Taylor AT; Berrueta TA; Murphey TD Active learning in robotics: A review of control principles. Mechatronics 2021, 77, 102576.

(447). Qiu J; Wu Q; Ding G; Xu Y; Feng S A survey of machine learning for big data processing. EURASIP J. Adv. Signal Process. 2016, 2016, 67.

(448). Ienco D; Pensa RG Positive and unlabeled learning in categorical data. Neurocomputing 2016, 196, 113–124.

(449). Hu R; Mac Namee B; Delany SJ Active learning for text classification with reusability. Expert. Syst. Appl. 2016, 45, 438–449.

(450). Zhang Z; Strubell E; Hovy E A Survey of Active Learning for Natural Language Processing. arXiv 2022, arXiv.2210.10109.

(451). Wu T; Ortiz J: RLAD: Time series anomaly detection through reinforcement learning and active learning. arXiv 2021, arXiv.2104.00543.

(452). de Aquino Afonso BK; Berton L Analysis of label noise in graph-based semi-supervised learning. In SAC '20: Proceedings of the 35th Annual ACM Symposium on Applied Computing; ACM, 2020; pp 1127–1134.

(453). Afonso B. K. d. A.; Berton L. Analysis of label noise in graph-based semi-supervised learning. arXiv 2020, arXiv.2009.12966

(454). Van Zyl G Graph-Based Semi-Supervised Learning for the Detection of Potential Disease Causing Genes. Ph.D. Thesis. Stellenbosch University: Stellenbosch, 2020.

(455). Chen C; Li Y; Qian H; Zheng Z; Hu Y Multi-view semi-supervised learning for classification on dynamic networks. Knowledge-Based Syst. 2020, 195, 105698.

(456). Hayes N; Rapinchuk E; Wei G-W Integrating transformer and autoencoder techniques with spectral graph algorithms for the prediction of scarcely labeled molecular data. Comput. Biol. Med. 2023, 153, 106479. [PubMed: 36610214]

(457). Morgan HL The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. J. Chem. Doc. 1965, 5, 107–113.

(458). Merkurjev E; Nguyen DD; Wei G-W Multiscale LaPlacian Learning. arXiv 2021, arXiv.2109.03718.

(459). Merriman B; Bence JK; Osher S Diffusion Generated Motion by Mean Curvature; Department of Mathematics, University of California: Los Angeles, 1992.

(460). Calder J; Cook B; Thorpe M; Slepcev D Poisson learning: Graph based semi-supervised learning at very low label rates. In International Conference on Machine Learning, 2020; pp 1306–1316.

(461). Guillaumin M; Verbeek J; Schmid C Multimodal semi-supervised learning for image classification. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010; pp 902–909.

(462). Han Y; Liu Y; Jin Z Sentiment analysis via semi-supervised learning: a model based on dynamic threshold and multi-classifiers. Neural Comput. Appl. 2020, 32, 5117–5129.

(463). Zhang Y; Park DS; Han W; Qin J; Gulati A; Shor J; Jansen A; Xu Y; Huang Y; Wang S; et al. Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition. IEEE J. Sel. Top. Signal Process. 2022, 16, 1519–1532.

(464). Liu S Generalized Mahalanobis Depth in Point Process and Its Application in Neural Coding and Semi-Supervised Learning in Bioinformatics. Ph.D. thesis. The Florida State University, 2018.

(465). Sahoo P; Roy I; Wang Z; Mi F; Yu L; Balasubramani P; Khan L; Stoddart JF MultiCon: a semi-supervised approach for predicting drug function from chemical structure analysis. J. Chem. Inf. Model. 2020, 60, 5995–6006. [PubMed: 33140954]

(466). Shi S; Nie F; Wang R; Li X Semi-supervised learning based on intra-view heterogeneity and inter-view compatibility for image classification. Neurocomputing 2022, 488, 248–260.

(467). Bair E Semi-supervised clustering methods. Wiley Interdiscip. Rev. Comput. Stat. 2013, 5, 349–361. [PubMed: 24729830]

(468). Zhao M; Zhang Z; Chow TW; Li B A general soft label based linear discriminant analysis for semi-supervised dimensionality reduction. Neural Networks 2014, 55, 83–97. [PubMed: 24819874]

(469). Wu Q; Liu Y; Li Q; Jin S; Li F The application of deep learning in computer vision. In 2017 Chinese Automation Congress (CAC), 2017; pp 6522–6527.

(470). Leidner F; Kurt Yilmaz N; Schiffer CA Deciphering Antifungal Drug Resistance in Pneumocystis jirovecii DHFR with Molecular Dynamics and Machine Learning. J. Chem. Inf. Model. 2021, 61, 2537–2541. [PubMed: 34138546]

(471). Yilancioglu K; Weinstein ZB; Meydan C; Akhmetov A; Toprak I; Durmaz A; Iossifov I; Kazan H; Roth FP; Cokol M Target-independent prediction of drug synergies using only drug lipophilicity. J. Chem. Inf. Model. 2014, 54, 2286–2293. [PubMed: 25026390]

(472). Otter DW; Medina JR; Kalita JK A survey of the usages of deep learning for natural language processing. IEEE Trans. Neural Networks Learn. Syst. 2021, 32, 604–624.

(473). Wigh DS; Goodman JM; Lapkin AA A review of molecular representation in the age of machine learning. Wiley Interdiscip. Rev.: Comput. Mol. Sci. 2022, 12, No. e1603.

(474). Staszak M; Staszak K; Wieszczycka K; Bajek A; Roszkowski K; Tylkowski B Machine learning in drug design: Use of artificial intelligence to explore the chemical structure–biological activity relationship. Wiley Interdiscip. Rev.: Comput. Mol. Sci. 2022, 12, No. e1568.

(475). Yang C-I; Li Y-P Explainable uncertainty quantifications for deep learning-based molecular property prediction. J. Cheminform. 2023, 15, 13. [PubMed: 36737786]

(476). Yang Y; Wu Z; Yao X; Kang Y; Hou T; Hsieh C-Y; Liu H Exploring Low-Toxicity Chemical Space with Deep Learning for Molecular Generation. J. Chem. Inf. Model. 2022, 62, 3191–3199. [PubMed: 35713712]

(477). Pandey M; Fernandez M; Gentile F; Isayev O; Tropsha A; Stern AC; Cherkasov A The transformational role of GPU computing and deep learning in drug discovery. Nat. Mach. Intell. 2022, 4, 211–221.

(478). Cover T; Hart P Nearest neighbor pattern classification. IEEE Trans. Inf. Theory 1967, 13, 21–27.

(479). Pearl J Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference; Morgan Kaufmann, 1988.

(480). Vapnik V The Nature of Statistical Learning Theory; Springer Science & Business Media, 1999.

(481). Friedman JH Greedy function approximation: a gradient boosting machine. Ann. Statist. 2001, 29, 1189–1232.

(482). Martin T Users Guide for T.E.S.T. (version 4.2). In Toxicity Estimation Software Tool) A Program to Estimate Toxicity from Molecular Structure; EPA/600/R-16/058; U.S. EPA Office of Research and Development: Washington, DC, 2016.

(483). Qiu W; Lv Z; Hong Y; Jia J; Xiao X BOW-GBDT: a GBDT classifier combining with artificial neural network for identifying GPCR–drug interaction based on wordbook learning from sequences. Front. Cell Dev. Biol. 2021, 8, 623858. [PubMed: 33598456]

(484). Chawla NV; Bowyer KW; Hall LO; Kegelmeyer WP SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 2002, 16, 321–357.

(485). Xiao X; Min J-L; Wang P; Chou K-C iGPCR-Drug: A web server for predicting interaction between GPCRs and drugs in cellular networking. PLoS One 2013, 8, No. e72234. [PubMed: 24015221]

(486). Deng D; Chen X; Zhang R; Lei Z; Wang X; Zhou F XGraphBoost: extracting graph neural network-based features for a better prediction of molecular properties. J. Chem. Inf. Model. 2021, 61, 2697–2705. [PubMed: 34009965]

(487). Chen T; Guestrin C Xgboost: A scalable tree boosting system; Proceedings of the 22nd ACM SIGDD International Conference on Knowledge Discovery and Data Mining, 2016; p 785.

(488). Liu Q; He D; Wang J; Hou Y Intelligent Equipment, Robots, and Vehicles; Springer, 2021; pp 755–764.

(489). Parkinson J; Hard R; Ainsworth RI; Li N; Wang W Engineering a histone reader protein by combining directed evolution, sequencing, and neural network based ordinal regression. J. Chem. Inf. Model. 2020, 60, 3992–4004. [PubMed: 32786513]

(490). Gyires-Tóth BP; Gyires-Tóth M; Papp D; Szücs G. Deep learning and SVM classification for plant recognition in content-based large scale image retrieval. Cybernetics Information Technol. 2019, 19, 88–100.

(491). Chaganti SY; Nanda I; Pandi KR; Prudhvith TG; Kumar N Image Classification Using SVM and CNN. In 2020 International Conference on Computer Science, Engineering and Applications; ICCSEA, 2020; pp 1–5.

(492). Fu R; Li B; Gao Y; Wang P Content-based image retrieval based on CNN and SVM. In 2016 2nd IEEE International Conference on Computer and Communications (ICCC); IEEE, 2016; pp 638–642.

(493). Nguyen DD; Cang Z; Wei G-W A review of mathematical representations of biomolecular data. Phys. Chem. Chem. Phys. 2020, 22, 4343–4367. [PubMed: 32067019]

(494). Schneider J; Korshunova K; Si Chaib Z; Giorgetti A; Alfonso-Prieto M; Carloni P Ligand pose predictions for human G protein-coupled receptors: insights from the Amber-based hybrid Molecular Mechanics/Coarse-Grained approach. J. Chem. Inf. Model. 2020, 60, 5103–5116. [PubMed: 32786708]

(495). Bai Q; Liu S; Tian Y; Xu T; Banegas-Luna AJ; Pérez-Sánchez H; Huang J; Liu H; Yao X Application advances of deep learning methods for de novo drug design and molecular dynamics simulation. WIREs Comput. Mol. Sci. 2022, 12, No. e1581.

(496). Chmiela S; Sauceda HE; Müller K-R; Tkatchenko A. Towards exact molecular dynamics simulations with machine-learned force fields. Nat. Commun. 2018, 9, 3887. [PubMed: 30250077]

(497). Han Y; Ali I; Wang Z; Cai J; Wu S; Tang J; Zhang L; Ren J; Xiao R; Lu Q; et al. Machine learning accelerates quantum mechanics predictions of molecular crystals. Phys. Rep. 2021, 934, 1–71.

(498). Dral PO Quantum chemistry in the age of machine learning. J. Phys. Chem. Lett. 2020, 11, 2336–2347. [PubMed: 32125858]

(499). Parr RG Density functional theory. Annu. Rev. Phys. Chem. 1983, 34, 631–656.

(500). Metropolis N; Ulam S The monte carlo method. J. Am. Stat. Assoc. 1949, 44, 335–341. [PubMed: 18139350]

(501). Bhavikatti S Finite Element Analysis; New Age International, 2005.

(502). Zhang Y; Wang L; Wang X; Zhang C; Ge J; Tang J; Su A; Duan H Data augmentation and transfer learning strategies for reaction prediction in low chemical data regimes. Org. Chem. Front. 2021, 8, 1415–1423.

(503). Jian Y; Kruus E; Min MR T-Cell Receptor–Peptide Interaction Prediction with Physical Model Augmented Pseudo-Labeling. Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022; pp 3090–3097.

(504). Xie X; Li P; Xu Y; Zhou L; Yan Y; Xie L; Jia C; Guo X Single-molecule junction: A reliable platform for monitoring molecular physical and chemical processes. ACS Nano 2022, 16, 3476–3505. [PubMed: 35179354]

(505). Pogozheva ID; Armstrong GA; Kong L; Hartnagel TJ; Carpino CA; Gee SE; Picarello DM; Rubin AS; Lee J; Park S; et al. Comparative Molecular Dynamics Simulation Studies of Realistic Eukaryotic, Prokaryotic, and Archaeal Membranes. J. Chem. Inf. Model. 2022, 62, 1036–1051. [PubMed: 35167752]

(506). Li TE; Hammes-Schiffer S QM/MM Modeling of Vibrational Polariton Induced Energy Transfer and Chemical Dynamics. J. Am. Chem. Soc. 2023, 145, 377–384. [PubMed: 36574620]

(507). Mulliken RS; Roothaan CC Broken bottlenecks and the future of molecular quantum mechanics. Proc. Natl. Acad. Sci. 1959, 45, 394–398. [PubMed: 16590398]

(508). Hassan-Harrirou H; Zhang C; Lemmin T RosENet: improving binding affinity prediction by leveraging molecular mechanics energies with an ensemble of 3D convolutional neural networks. J. Chem. Inf. Model. 2020, 60, 2791–2802. [PubMed: 32392050]

(509). YazdanYar A; Aschauer U; Bowen P Interaction of biologically relevant ions and organic molecules with titanium oxide (rutile) surfaces: A review on molecular dynamics studies. Colloids Surf., B 2018, 161, 563–577.

(510). Bengtson A; Nam HO; Saha S; Sakidja R; Morgan D First-principles molecular dynamics modeling of the LiCl–KCl molten salt system. Comput. Mater. Sci. 2014, 83, 362–370.

(511). Zepeda-Ruiz LA; Stukowski A; Oppelstrup T; Bulatov VV Probing the limits of metal plasticity with molecular dynamics simulations. Nature 2017, 550, 492–495. [PubMed: 28953878]

(512). Yu W; Wang Z; Stroud D Empirical molecular-dynamics study of diffusion in liquid semiconductors. Phys. Rev. B 1996, 54, 13946.

(513). Bauchy M; Laubie H; Abdolhosseini Qomi MA; Hoover C; Ulm F-J; Pellenq R-M Fracture toughness of calcium–silicate–hydrate from molecular dynamics simulations. J. Non-Cryst. Solids 2015, 419, 58–64.

(514). Pasichnyk I; Dünweg, B. Coulomb interactions via local dynamics: A molecular-dynamics algorithm. J. Phys.: Condens. Matter 2004, 16, S3999.

(515). Plimpton S Fast parallel algorithms for short-range molecular dynamics. J. Comput. Phys. 1995, 117, 1–19.

(516). Soares TA; Hünenberger PH; Kastenholz MA.; Kräutler V; Lenz T; Lins RD; Oostenbrink C; van Gunsteren s. An improved nucleic acid parameter set for the GROMOS force field. J. Comput. Chem. 2005, 26, 725–737. [PubMed: 15770662]

(517). Shen L; Yang W Molecular dynamics simulations with quantum mechanics/molecular mechanics and adaptive neural networks. J. Chem. Theory Comput. 2018, 14, 1442–1455. [PubMed: 29438614]

(518). Case DA; Cheatham TE III; Darden T; Gohlke H; Luo R; Merz KM Jsr; Onufriev A; Simmerling C; Wang B; Woods RJ. The Amber biomolecular simulation programs. J. Comput. Chem. 2005, 26, 1668–1688. [PubMed: 16200636]

(519). Suárez D; Díaz N SARS-CoV-2 main protease: A molecular dynamics study. J. Chem. Inf. Model. 2020, 60, 5815–5831. [PubMed: 32678588]

(520). Guterres H; Im W Improving protein-ligand docking results with high-throughput molecular dynamics simulations. J. Chem. Inf. Model. 2020, 60, 2189–2198. [PubMed: 32227880]

(521). Homeyer N; Gohlke H Free energy calculations by the molecular mechanics Poisson-Boltzmann surface area method. Mol. Inf. 2012, 31, 114–122.

(522). Do P-C; Lee EH; Le L Steered molecular dynamics simulation in rational drug design. J. Chem. Inf. Model. 2018, 58, 1473–1482. [PubMed: 29975531]

(523). Hohenberg P; Kohn W Inhomogeneous electron gas. Phys. Rev. 1964, 136, B864.

(524). Kohn W; Sham LJ Self-consistent equations including exchange and correlation effects. Phys. Rev. 1965, 140, A1133.

(525). Rai BK; Sresht V; Yang Q; Unwalla R; Tu M; Mathiowetz AM; Bakken GA TorsionNet: A Deep Neural Network to Rapidly Predict Small-Molecule Torsional Energy Profiles with the Accuracy of Quantum Mechanics. J. Chem. Inf. Model. 2022, 62, 785–800. [PubMed: 35119861]

(526). Ban F; Rankin KN; Gauld JW; Boyd RJ Recent applications of density functional theory calculations to biomolecules. Theor. Chem. Acc. 2002, 108, 1–11.

(527). Senn HM; Thiel W QM/MM methods for biomolecular systems. Angew. Chem., Int. Ed. 2009, 48, 1198–1229.

(528). Tavakoli M; Mood A; Van Vranken D; Baldi P Quantum mechanics and machine learning synergies: graph attention neural networks to predict chemical reactivity. J. Chem. Inf. Model. 2022, 62, 2121–2132. [PubMed: 35020394]

(529). Qiao Z; Welborn M; Anandkumar A; Manby FR; Miller TF III OrbNet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features. J. Chem. Phys. 2020, 153, 124111. [PubMed: 33003742]

(530). Bennett WD; He S; Bilodeau CL; Jones D; Sun D; Kim H; Allen JE; Lightstone FC; Ingólfsson HI Predicting small molecule transfer free energies by combining molecular dynamics simulations and deep learning. J. Chem. Inf. Model. 2020, 60, 5375–5381. [PubMed: 32794768]

(531). Jamal S; Grover A; Grover S Machine learning from molecular dynamics trajectories to predict caspase-8 inhibitors against Alzheimers disease. Front. Pharmacol. 2019, 10, 780. [PubMed: 31354494]

(532). Botu V; Batra R; Chapman J; Ramprasad R Machine learning force fields: construction, validation, and outlook. J. Phys. Chem. C 2017, 121, 511–522.

(533). Schlick T; Portillo-Ledesma S Biomolecular modeling thrives in the age of technology. Nat. Comput. Sci. 2021, 1, 321–331. [PubMed: 34423314]

(534). Soares TA; Nunes-Alves A; Mazzolari A; Ruggiu F; Wei G-W; Merz K The (Re)-Evolution of Quantitative Structure–Activity Relationship (QSAR) Studies Propelled by the Surge of Machine Learning Methods. J. Chem. Inf. Model. 2022, 62, 5317–5320. [PubMed: 36437763]

(535). Liu D; Xu P; Ren L TPFlow: Progressive partition and multidimensional pattern extraction for large-scale spatio-temporal data analysis. IEEE Trans. Visual Comput. Graphics 2019, 25, 1–11.

(536). Trine A; Monson BB Extended high frequencies provide both spectral and temporal information to improve speech-in-speech recognition. Trends Hearing 2020, 24, 2331216520980299.

(537). Kormilitzin A; Vaci N; Liu Q; Nevado-Holgado A Med7: A transferable clinical natural language processing model for electronic health records. Artif. Intell. Med. 2021, 118, 102086. [PubMed: 34412834]

(538). Sridharan B; Goel M; Priyakumar UD Modern machine learning for tackling inverse problems in chemistry: molecular design to realization. Chem. Commun. 2022, 58, 5316–5331.

(539). Roth GA; Picece VC; Ou BS; Luo W; Pulendran B; Appel EA Designing spatial and temporal control of vaccine responses. Nat. Rev. Mater. 2022, 7, 174–195. [PubMed: 34603749]

(540). Goel M; Aggarwal R; Sridharan B; Pal PK; Priyakumar UD Efficient and enhanced sampling of drug-like chemical space for virtual screening and molecular design using modern machine learning methods. WIREs Comput. Mol. Sci. 2023, 13, No. e1637.

(541). Wang Y; Sun Y; Liu Z; Sarma SE; Bronstein MM; Solomon JM Dynamic graph cnn for learning on point clouds. ACM Trans. Graph. 2019, 38, 1–12.

(542). Yu Y; Si X; Hu C; Zhang J A review of recurrent neural networks: LSTM cells and network architectures. Neural Comput. 2019, 31, 1235–1270. [PubMed: 31113301]

(543). Fu R; Zhang Z; Li L Using LSTM and GRU Neural Network Methods for Traffic Flow Prediction. In 2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC), 2016; pp 324–328.

(544). Azad R; Asadi-Aghbolaghi M; Fathy M; Escalera S Bidirectional ConvLSTM U-Net with densley connected convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops; IEEE, 2019.

(545). Bao W; Yue J; Rao Y A deep learning framework for financial time series using stacked autoencoders and long-short term memory. PloS one 2017, 12, No. e0180944. [PubMed: 28708865]

(546). Chowdhary K Natural Language Processing; Fundamentals of Artificial Intelligence, 2020; pp 603–649.

(547). Minaee S; Kalchbrenner N; Cambria E; Nikzad N; Chenaghlu M; Gao J Deep learning–based text classification: a comprehensive review. ACM Comput. Surv. 2022, 54, 1–40.

(548). Liu P; Yuan W; Fu J; Jiang Z; Hayashi H; Neubig G Pretrain, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Comput. Surv. 2023, 55, 1–35.

(549). Khurana D; Koli A; Khatter K; Singh S Natural language processing: State of the art, current trends and challenges. Multimedia Tools Appl. 2023, 82, 3713.

(550). Suta P; Lan X; Wu B; Mongkolnam P; Chan JH An overview of machine learning in chatbots. Int. J. Mech. Eng. Robot. Res. 2020, 9, 502–510.

(551). Nemes L; Kiss A Social media sentiment analysis based on COVID-19. J. Inf. Telecommun. 2021, 5, 1–15.

(552). Karthikeyan A; Priyakumar UD Artificial intelligence: machine learning for chemical sciences. J. Chem. Sci. 2022, 134, 134.

(553). Singh S; Sunoj RB A transfer learning protocol for chemical catalysis using a recurrent neural network adapted from natural language processing. Digital Discovery 2022, 1, 303–312.

(554). Winter B; Winter C; Schilling J; Bardow A A smile is all you need: predicting limiting activity coefficients from SMILES with natural language processing. Digital Discovery 2022, 1, 859–869. [PubMed: 36561987]

(555). Lu J; Zhang Y Unified deep learning model for multitask reaction predictions with explanation. J. Chem. Inf. Model. 2022, 62, 1376–1387. [PubMed: 35266390]

(556). Mukherjee S; Ben-Joseph J; Campos M; Malla P; Nguyen H; Pham A; Oates T; Janarthanan V Predicting Physiological Effects of Chemical Substances Using Natural Language Processing. In 2021 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE); IEEE, 2021; pp 1–6.

(557). Xie Y; Le L; Zhou Y; Raghavan VV Handbook of Statistics; Elsevier, 2018; Vol. 38; pp 317–328.

(558). Brown PF; Della Pietra VJ; Desouza PV; Lai JC; Mercer RL Class-based n-gram models of natural language. Comput. Linguist. 1992, 18, 467–480.

(559). Li Y; Yang T Guide to Big Data Applications; Springer, 2018; pp 83–104.

(560). Yin W; Kann K; Yu M; Schütze H. Comparative study of CNN and RNN for natural language processing. arXiv 2017, arXiv.1702.01923.

(561). Wolf T; Debut L; Sanh V; Chaumond J; Delangue C; Moi A; Cistac P; Rault T; Louf R; Funtowicz M, et al. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020; pp 38–45.

(562). Aziz MVG; Prihatmanto AS; Henriyan D; Wijaya R Design and implementation of natural language processing with syntax and semantic analysis for extract traffic conditions from social media data. In 2015 5th IEEE International Conference on System Engineering and Technology (ICSET); IEEE, 2015; pp 43–48.

(563). G M H; Gourisaria MK; Pandey M; Rautaray SS A comprehensive survey and analysis of generative models in machine learning. Comput. Sci. Rev. 2020, 38, 100285.

(564). Bilodeau C; Jin W; Jaakkola T; Barzilay R; Jensen KF Generative models for molecular discovery: Recent advances and challenges. WIREs Comput Mol Sci. 2022, 12, No. e1608.

(565). Tong X; Liu X; Tan X; Li X; Jiang J; Xiong Z; Xu T; Jiang H; Qiao N; Zheng M Generative models for De Novo drug design. J. Med. Chem. 2021, 64, 14011–14027. [PubMed: 34533311]

(566). Yakubovich A; Odinokov A; Nikolenko S; Jung Y; Choi H Computational Discovery of TTF Molecules with Deep Generative Models. Front. Chem. 2021, 9, 800133. [PubMed: 35004615]

(567). Kingma DP; Welling M An introduction to variational autoencoders. Found. Trends Mach. Learn. 2019, 12, 307–392.

(568). Gao K; Nguyen DD; Tu M; Wei G-W Generative network complex for the automated generation of drug-like molecules. J. Chem. Inf. Model. 2020, 60, 5682–5698. [PubMed: 32686938]

(569). Van Den Oord A; Kalchbrenner N; Kavukcuoglu K Pixel recurrent neural networks. International Conference on Machine Learning, 2016; pp 1747–1756.

(570). Radford A; Narasimhan K; Salimans T; Sutskever I Improving Language Understanding by Generative Pre-Trainingt. 2018.

(571). Zhang Y; Sun S; Galley M; Chen Y-C; Brockett C; Gao X; Gao J; Liu J; Dolan B DialoGPT: Large-scale generative pretraining for conversational response generation. arXiv 2019, arXiv.1911.00536 DOI: 10.48550/arXiv.1911.00536.

(572). Zhang Y; Wang L; Wang X; Zhang C; Ge J; Tang J; Su A; Duan H Data augmentation and transfer learning strategies for reaction prediction in low chemical data regimes. Org. Chem. Front. 2021, 8, 1415–1423.

(573). Hao Z; Lu C; Huang Z; Wang H; Hu Z; Liu Q; Chen E; Lee C 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; pp 731–752.

(574). Unke OT; Chmiela S; Sauceda HE; Gastegger M; Poltavsky I; Schutt KT; Tkatchenko A; Muller K-R Machine learning force fields. Chem. Rev. 2021, 121, 10142–10186. [PubMed: 33705118]

(575). Poltavsky I; Tkatchenko A Machine Learning Force Fields: Recent Advances and Remaining Challenges. J. Phys. Chem. Lett. 2021, 12, 6551–6564. [PubMed: 34242032]

(576). Noé F; De Fabritiis G; Clementi C Machine learning for protein folding and dynamics. Curr. Opin. Struct. Biol. 2020, 60, 77–84. [PubMed: 31881449]

(577). dos Passos Gomes G; Pollice R; Aspuru-Guzik A Navigating through the maze of homogeneous catalyst design with machine learning. Trends Chem. 2021, 3, 96–110.

(578). Mo Y; Guan Y; Verma P; Guo J; Fortunato ME; Lu Z; Coley CW; Jensen KF Evaluating and clustering retrosynthesis pathways with learned strategy. Chem. Sci. 2021, 12, 1469–1478.

(579). Maldonado AM; Poltavsky I; Vassilev-Galindo V; Tkatchenko A; Keith JA Modeling molecular ensembles with gradient-domain machine learning force fields. Digital Discovery 2023, 2, 871–880.

(580). Allen AE; Tkatchenko A Machine learning of material properties: Predictive and interpretable multilinear models. Sci. Adv. 2022, 8, No. eabm7185. [PubMed: 35522750]

(581). Jumper J; Evans R; Pritzel A; Green T; Figurnov M; Ronneberger O; Tunyasuvunakool K; Bates R; Žídek A; Potapenko A; et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021, 596, 583–589. [PubMed: 34265844]

(582). Qiu Y; Wei G-W Persistent spectral theory-guided protein engineering. Nat. Comput. Sci. 2023, 3, 149–163. [PubMed: 37637776]

(583). Liu B; Li C-C; Yan K DeepSVM-fold: protein fold recognition by combining support vector machines and pairwise sequence similarity scores generated by deep learning networks. Briefings Bioinf. 2020, 21, 1733–1741.

(584). Benkovic SJ; Hammes-Schiffer S A perspective on enzyme catalysis. Science 2003, 301, 1196–1202. [PubMed: 12947189]

(585). Liao W; Liu P Enhanced descriptor identification and mechanism understanding for catalytic activity using a data-driven framework: revealing the importance of interactions between elementary steps. Catal. Sci. Technol. 2022, 12, 3836–3845.

(586). Wan X; Zhang Z; Yu W; Guo Y A density-functional-theory-based and machine-learning-accelerated hybrid method for intricate system catalysis. Mater. Rep.: Energy 2021, 1, 100046.

(587). Corey EJ; Wipke WT Computer-Assisted Design of Complex Organic Syntheses: Pathways for molecular synthesis can be devised with a computer and equipment for graphical communication. Science 1969, 166, 178–192. [PubMed: 17731475]

(588). Lin G-M; Warden-Rothman R; Voigt CA Retrosynthetic design of metabolic pathways to chemicals not found in nature. Curr. Opin. Syst. Biol. 2019, 14, 82–107.

(589). Shen Y; Borowski JE; Hardy MA; Sarpong R; Doyle AG; Cernak T Automation and computer-assisted planning for chemical synthesis. Nat. Rev. Methods Primers 2021, 1, 23.

(590). Badowski T; Gajewska EP; Molga K; Grzybowski BA Synergy between expert and machine-learning approaches allows for improved retrosynthetic planning. Angew. Chem., Int. Ed. 2020, 59, 725–730.

(591). Sun Y; Sahinidis NV Computer-aided retrosynthetic design: fundamentals, tools, and outlook. Curr. Opin. Chem. Eng. 2022, 35, 100721.

(592). Christensen AS; Bratholm LA; Faber FA; Anatole von Lilienfeld O FCHL revisited: Faster and more accurate quantum machine learning. J. Chem. Phys. 2020, 152, 044107. [PubMed: 32007071]

(593). Ceriotti M; Clementi C; Anatole von Lilienfeld O Introduction: machine learning at the atomic scale. Chem. Rev. 2021, 121, 9719–9721. [PubMed: 34428897]

(594). Bartók AP; De S; Poelking C; Bernstein N; Kermode JR; Csányi G; Ceriotti M Machine learning unifies the modeling of materials and molecules. Sci. Adv. 2017, 3, No. e1701816. [PubMed: 29242828]

(595). Paesani F; Bajaj P; Riera M Chemical accuracy in modeling halide ion hydration from many-body representations. Adv. Phys.: X 2019, 4, 1631212.

(596). Artrith N; Butler KT; Coudert F-X; Han S; Isayev O; Jain A; Walsh A Best practices in machine learning for chemistry. Nat. Chem. 2021, 13, 505–508. [PubMed: 34059804]

(597). Duan C; Nandy A; Meyer R; Arunachalam N; Kulik HJ A transferable recommender approach for selecting the best density functional approximations in chemical discovery. Nat. Comput. Sci. 2023, 3, 38–47. [PubMed: 38177951]

(598). Folmsbee D; Hutchison G Assessing conformer energies using electronic structure and machine learning methods. Int. J. Quantum Chem. 2021, 121, No. e26381.

(599). Kolluru A; Shuaibi M; Palizhati A; Shoghi N; Das A; Wood B; Zitnick CL; Kitchin JR; Ulissi ZW Open Challenges in Developing Generalizable Large-Scale Machine-Learning Models for Catalyst Discovery. ACS Catal 2022, 12, 8572–8581.

(600). Kitchin JR Machine learning in catalysis. Nat. Catal. 2018, 1, 230–232.

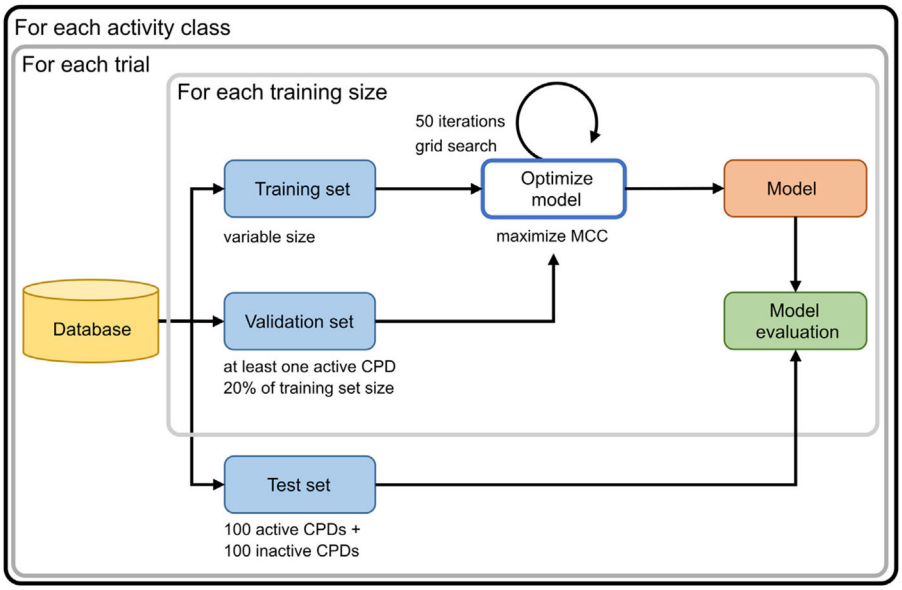


Figure 1.

Calculation protocol for molecular classification. For each activity class, eight independent trials with different seeds were carried out. For each trial, a test data set was randomly chosen containing 100 active and 100 inactive compounds. For each training set size, training and validation data sets were assembled. Reproduced with permission from ref [114]. Copyright 2022 Elsevier.

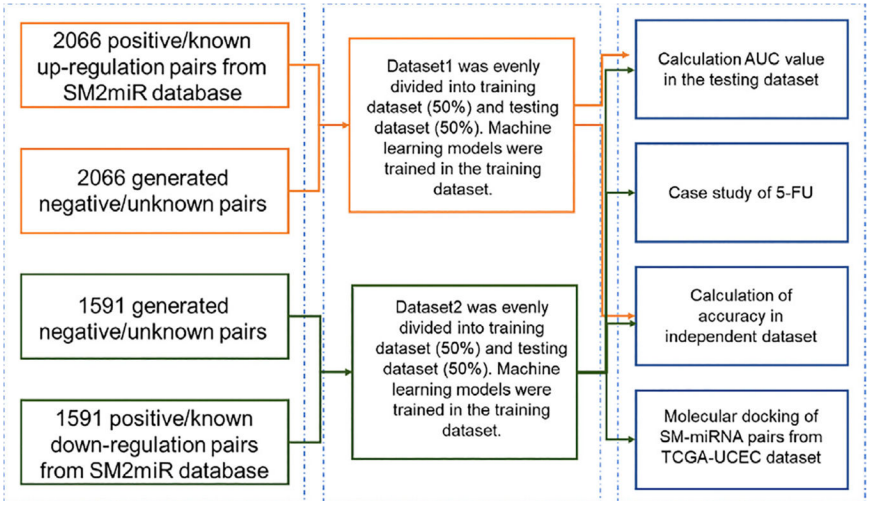


Figure 2.

Flowchart of developing models for SM-miRNA regulation prediction. Data set1 was used to construct models to predict the upregulation pairs of small molecules and miRNAs. Similarly, data set2 was used to construct models to predict down-regulation pairs. Reproduced with permission from ref [115]. Copyright 2022 Frontiers Media SA.

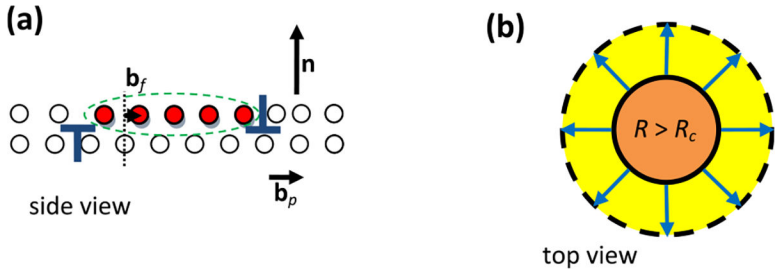


Figure 3.

MD assisted ANN prediction of the nucleation of dislocations in homogeneous lattices. (a) Nucleation of a dislocation loop by gradual displacement of a part of the atoms along the loop area. (b) The following mechanical growth of a supercritical dislocation loop by slip of dislocation lines. Reproduced with permission from ref [135]. Copyright 2022 Elsevier.

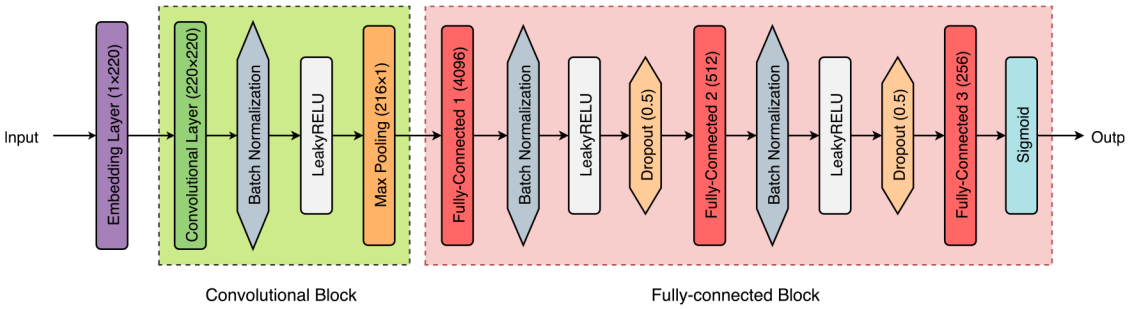


Figure 4.

Architecture of a DL model for screening of DILI compounds. The model consists of an embedding layer, a convolutional block and a fully connected block. The fully connected block consists of three fully connected layers. Except for the fully connected blocks in the last layer, the others are designed with batch normalization. Reproduced with permission from ref 168. Copyright 2020 American Chemical Society.

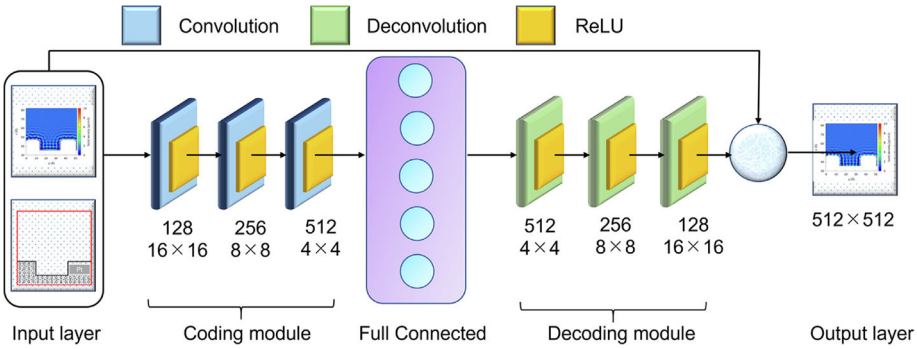


Figure 5.

Encoding–decoding CNN construction for the molecular adsorption density prediction. The proposed CNN mainly consisted of four parts: input layer, encoding module, decoding module and output layer. Reproduced with permission from ref 172. Copyright 2022 Elsevier.

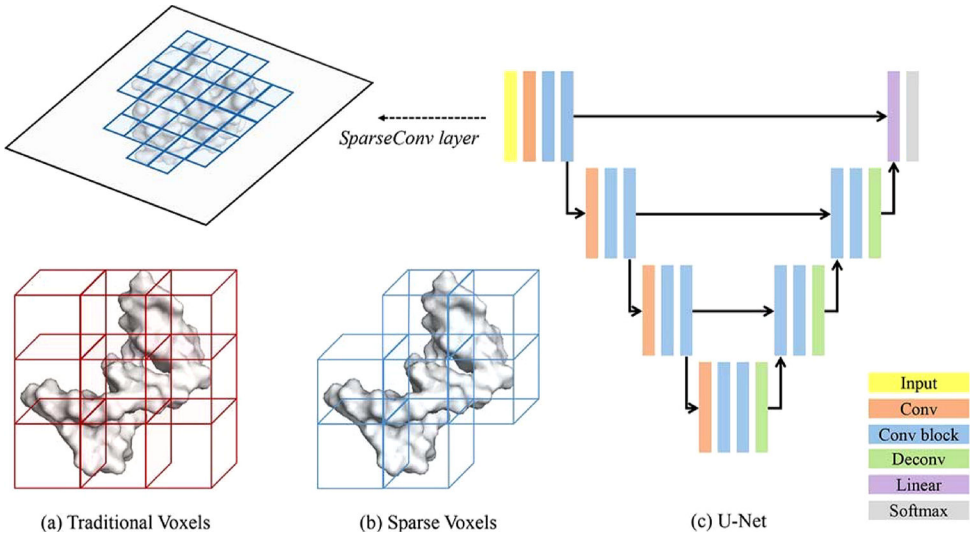


Figure 6.

Submanifold sparse convolution based U-Net in a 2D perspective. The difference between it and traditional 3D-CNN is illustrated in (a) and (b). In (c), we demonstrate the architecture of U-Net. Reproduced with permission from ref 193. Copyright 2022 American Chemical Society.

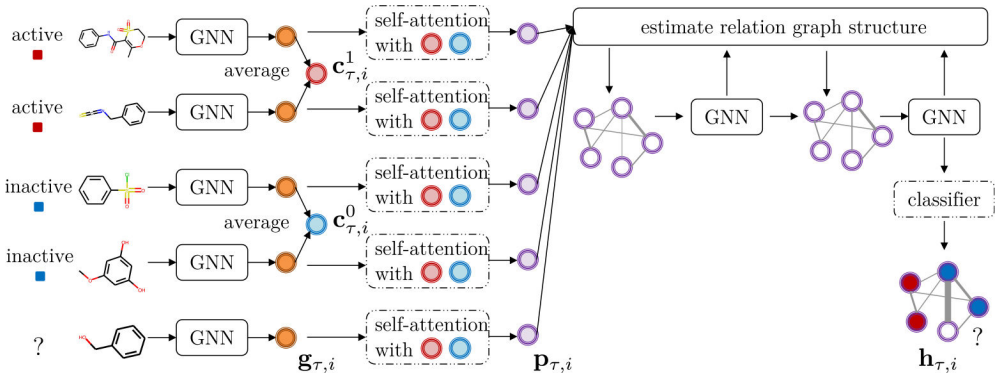


Figure 7.

Architecture of a GNN-based classifier for toxicity classification. PAR is optimized over a set of tasks. Within each task T_τ, the modules with dotted lines are fine-tuned on support set S_τ and those with solid lines are fixed. A query molecule $x_{\tau,i}$ will first be represented as $g_{\tau,i}$ using a graph-based molecular encoder, then transformed to $p_{\tau,i}$ by our property-aware embedding function. This $p_{\tau,i}$ further coadapts with embeddings of molecules in S_τ on the relation graph as $h_{\tau,i}$, which is taken as the final molecular embedding and used for class prediction. Reproduced with permission from ref 229. Copyright 2021 NeurIPs.

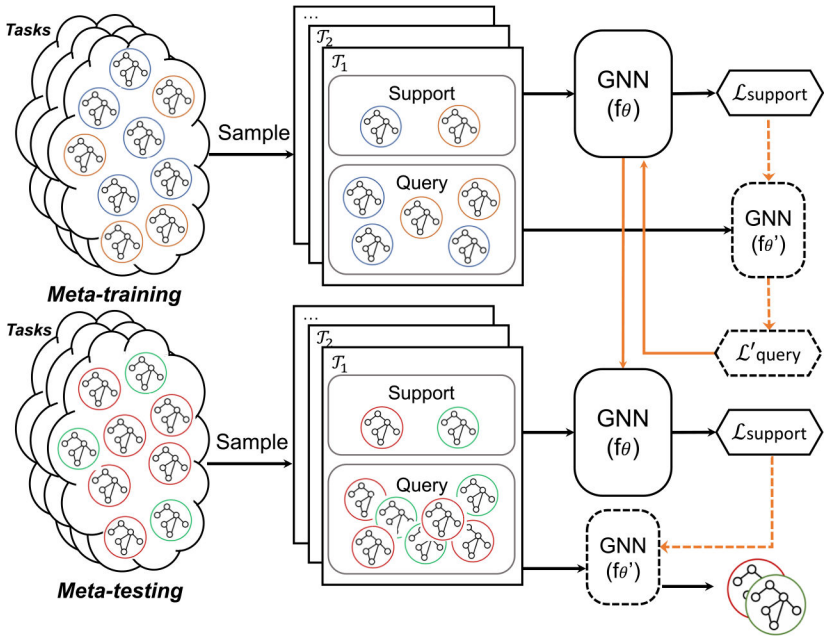


Figure 8.

Overall framework of a Meta-MGNN model for toxicity predictions: It first samples a batch of training tasks. For each task, there are a few data examples in the support set. These examples are fed into a GNN parametrized by θ. Then the support loss $\mathscr{L}_{\text{support}}$ is calculated and utilized to update the GNN parameters to θ'. Next, the examples in the corresponding query set are fed into the GNN parametrized by θ' and calculate the loss $\mathscr{L}'_{\text{query}}$ for this task. The same process is repeated for other training tasks. Later, we compute the summation of $\mathscr{L}'_{\text{query}}$ over all sampled tasks and use it to further update the GNN parameters for testing. Reproduced with permission from ref 232. Copyright 2021 Web of Conferences.

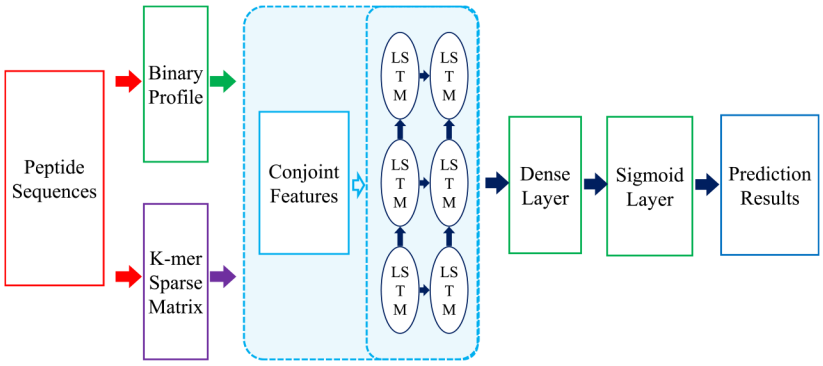


Figure 9.

Illustration of a LSTM architecture utilizing *k*-mer sparse matrices and binary contour features for predicting anticancer peptides.[263]

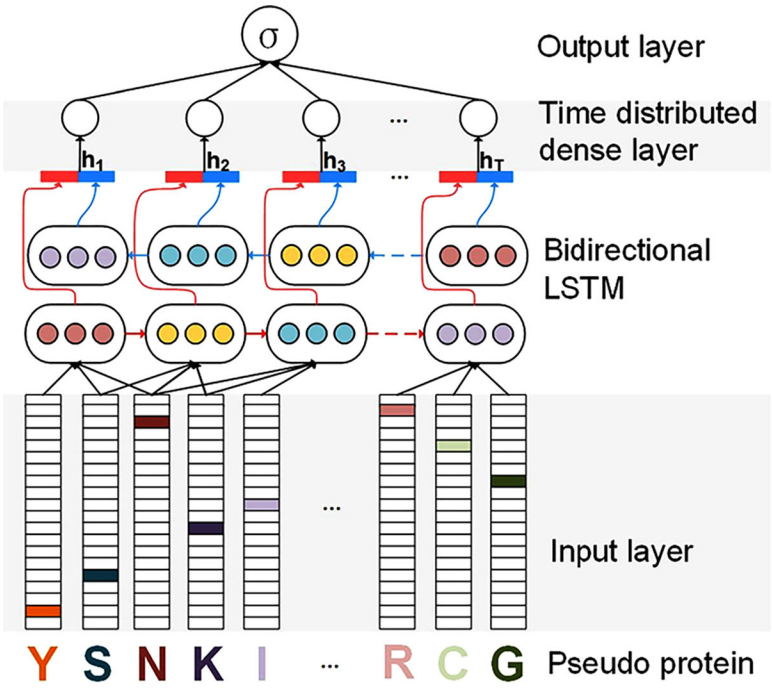


Figure 10.

Structure of ProDec-BLSTM for protein remote homology detection. The input layer converts the pseudo proteins into feature vectors by one-hot encoding. Next, the subsequences within the sliding window are fed into the BLSTM layer for the extraction of the sequence patterns. Then, the time-distributed dense layer weighs the extracted patterns. Finally, the extracted feature vectors are fed into an output layer for prediction. Reproduced with permission from ref 264. Copyright 2017 Springer Nature.

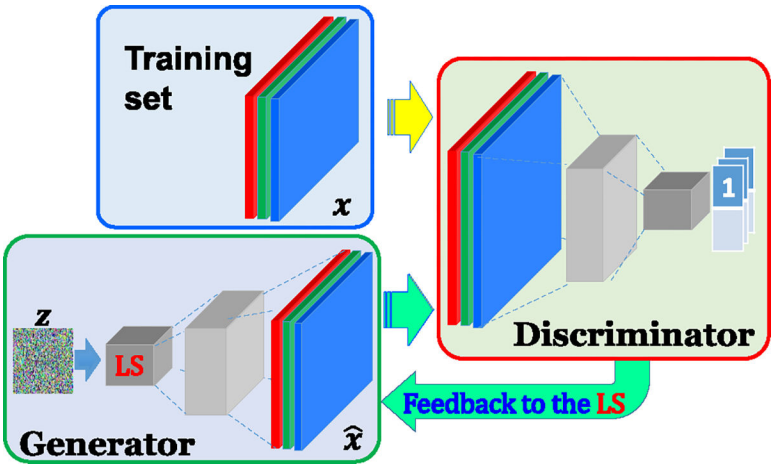


Figure 11.
Framework of a general-purpose GAN model. It consists of a discriminator and a generator.

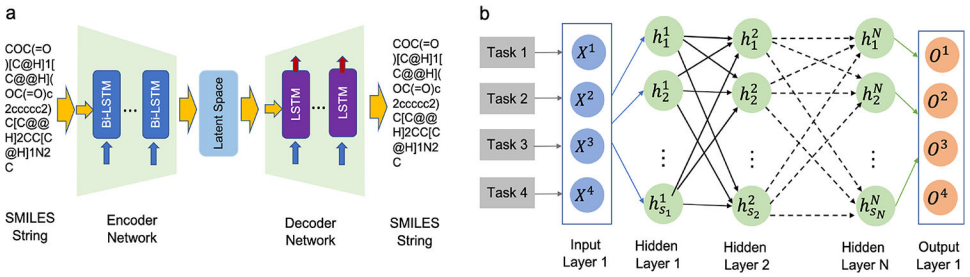


Figure 12.

Workflow of an autoencoder-assisted multitask ANN model for enhancing small data sets inferred by interactomics networks of cocaine addition targets.[315] (a) Sequence-to-sequence autoencoder model is used to create uniform features for different data sets. BLSTM and LSTM are used in encoder and decoder networks, respectively. (b) An MT-DNN model is connected to the autoencoder for regression predictions.

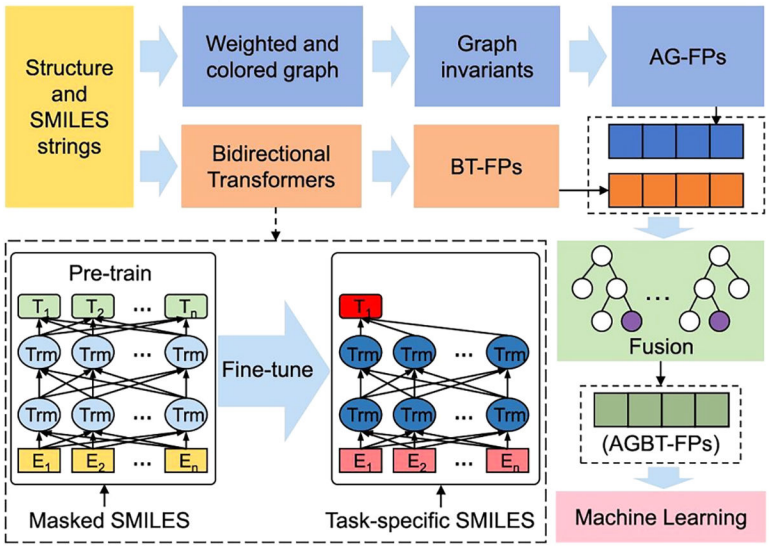


Figure 13.

Illustration of the AGBT model. For a given molecular structure and its SMILES strings, AG-FPs are generated from an element-specific algebraic subgraphs module and BT-FPs are generated from a deep bidirectional transformer module, as shown inside the dashed rectangle, which contains the pretraining and fine-tuning processes, and then finally completes the feature extraction using task-specific SMILES as input. Then the RF algorithm is used to fuse, rank, and select optimal fingerprints (AGBT-FPs) for ML.[334]

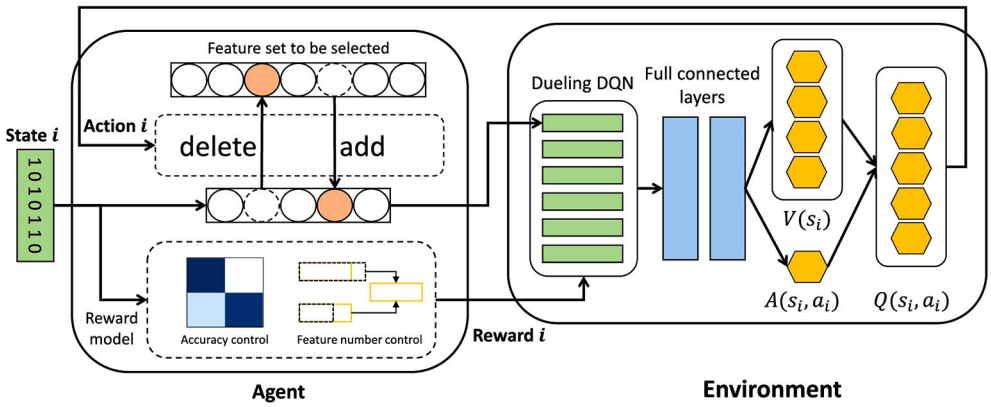


Figure 14.

RL model for feature selection in protein–ligand binding. The DQN based reinforcement learning is used to further select features to train a classifier. We formulate a new reward function to balance classification accuracy and number of features. The action set contains two basic operations, adding and deleting based on the χ^2 test, to search for the optimal state. Reproduced with permission from ref 372. Copyright 2022 Elsevier.

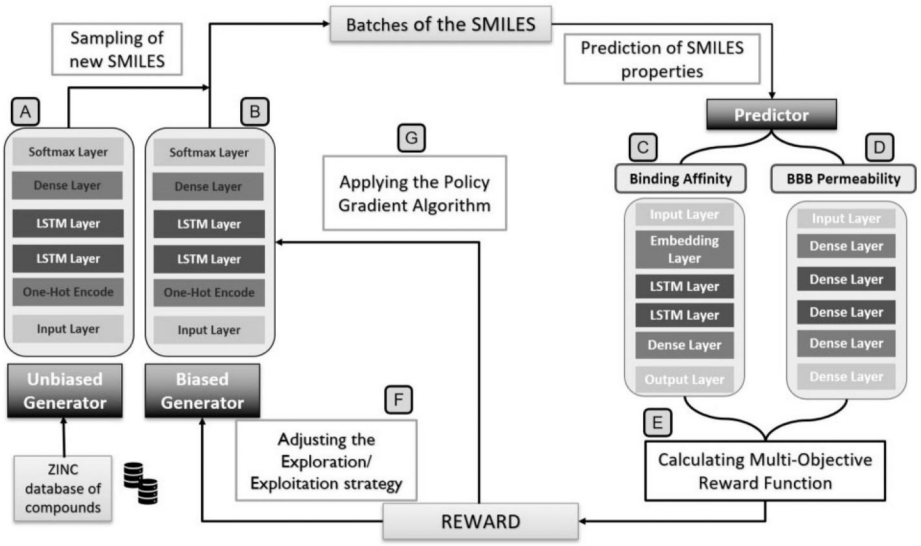


Figure 15.

General framework of molecular generation contains 4 DL modules: an unbiased generator (A), a biased generator (B) which shares the same architecture, and two QSAR models for predicting the binding affinity (C) and BBB permeation (D). The DL modules were interconnected by a policy-based reinforcement Learning approach (G) applied with a particular exploration/exploitation strategy (F) based on a multiobjective reward function (E). Reproduced with permission from ref 378. Copyright 2021 Oxford University Press.

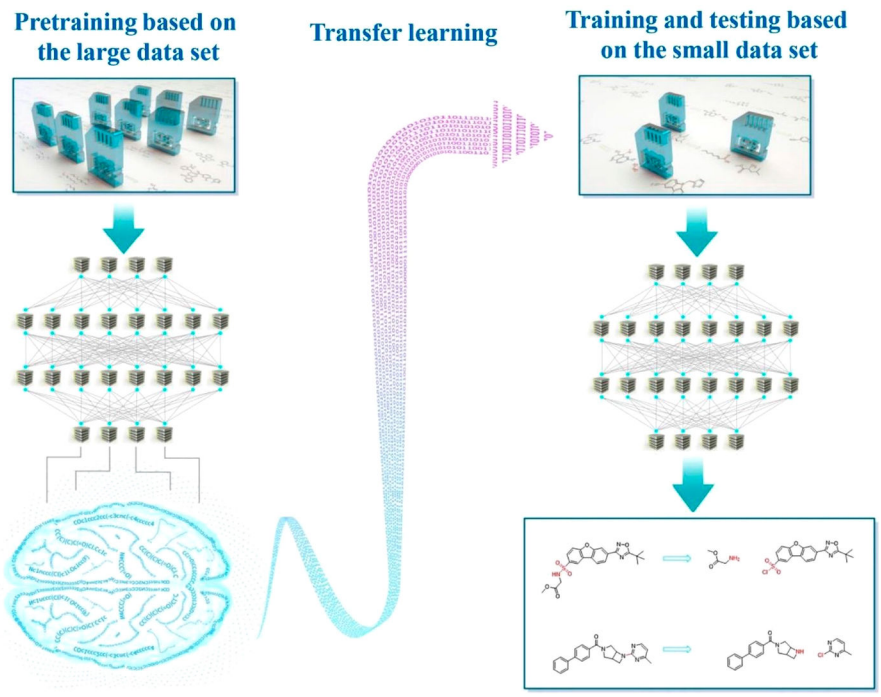


Figure 16.

Illustration of transfer-learning-aided retrosynthetic analysis. To improve the accuracy of the antisynthetic analysis, a migration learning strategy in terms of the seq2seq and transformer models was employed. In this analysis, a large chemical reaction data set was pretrained to acquire specialized knowledge of chemical reactions. Such learned knowledge is then successfully transferred to a smaller data set. With the chemical information attained from the pretraining, the final model yields higher accuracy. Reproduced with permission from ref 416. Copyright 2020 Multidisciplinary Digital Publishing Institute.

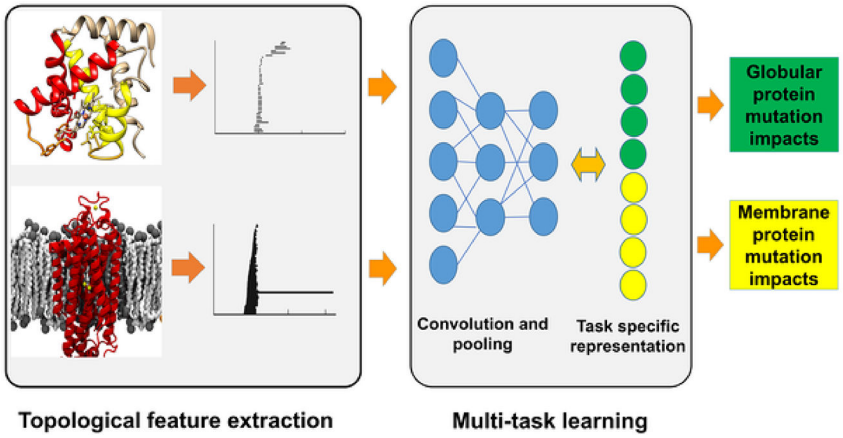


Figure 17.

Schematic illustration of a multitask topological DL model.[419] Topological invariants extracted by element-specific persistent homology are shared among globular proteins and membrane proteins.

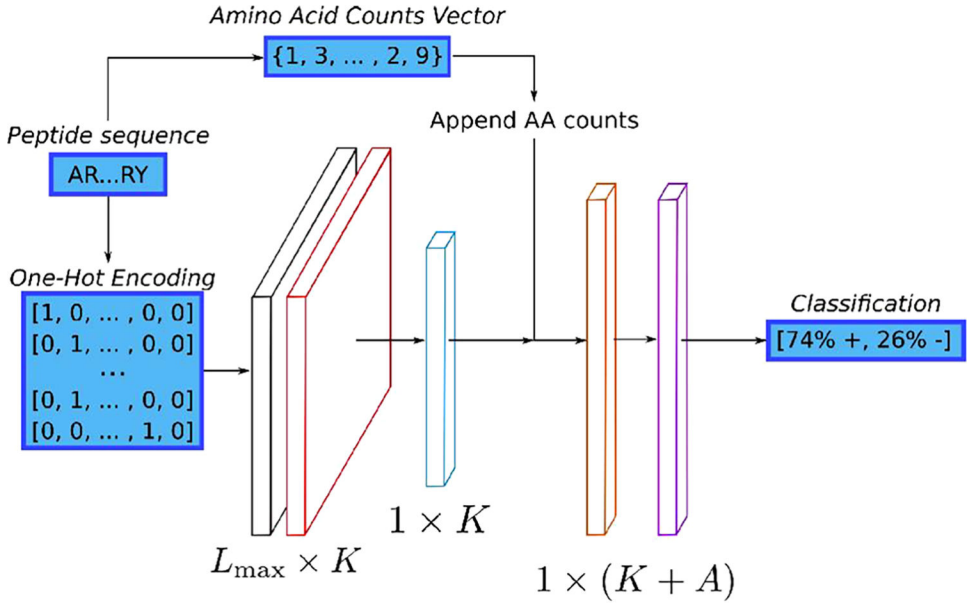


Figure 18.

Neural network structure for active learning. Here, $L_{\max}$ is the maximum width of a peptide in the data set (although a convolution can use any length), K is the number of motif classes, and A is the length of the amino acid alphabet. Peptides are first translated to a one-hot encoding ($L_{\max} \times A$) and a vector of normalized amino acid counts ($1 \times N$). The output of the max pool layer is passed through one fully connected layer with ReLU activation, then, amino acid counts are appended to the output. This is then passed into two more fully connected layers with a final output dimension of 2 for positive and negative class labels. Labels below neural network layers indicate the dimensionality of the data as they pass through the layer. Reproduced with permission from ref 443. Copyright 2020 American Chemical Society.

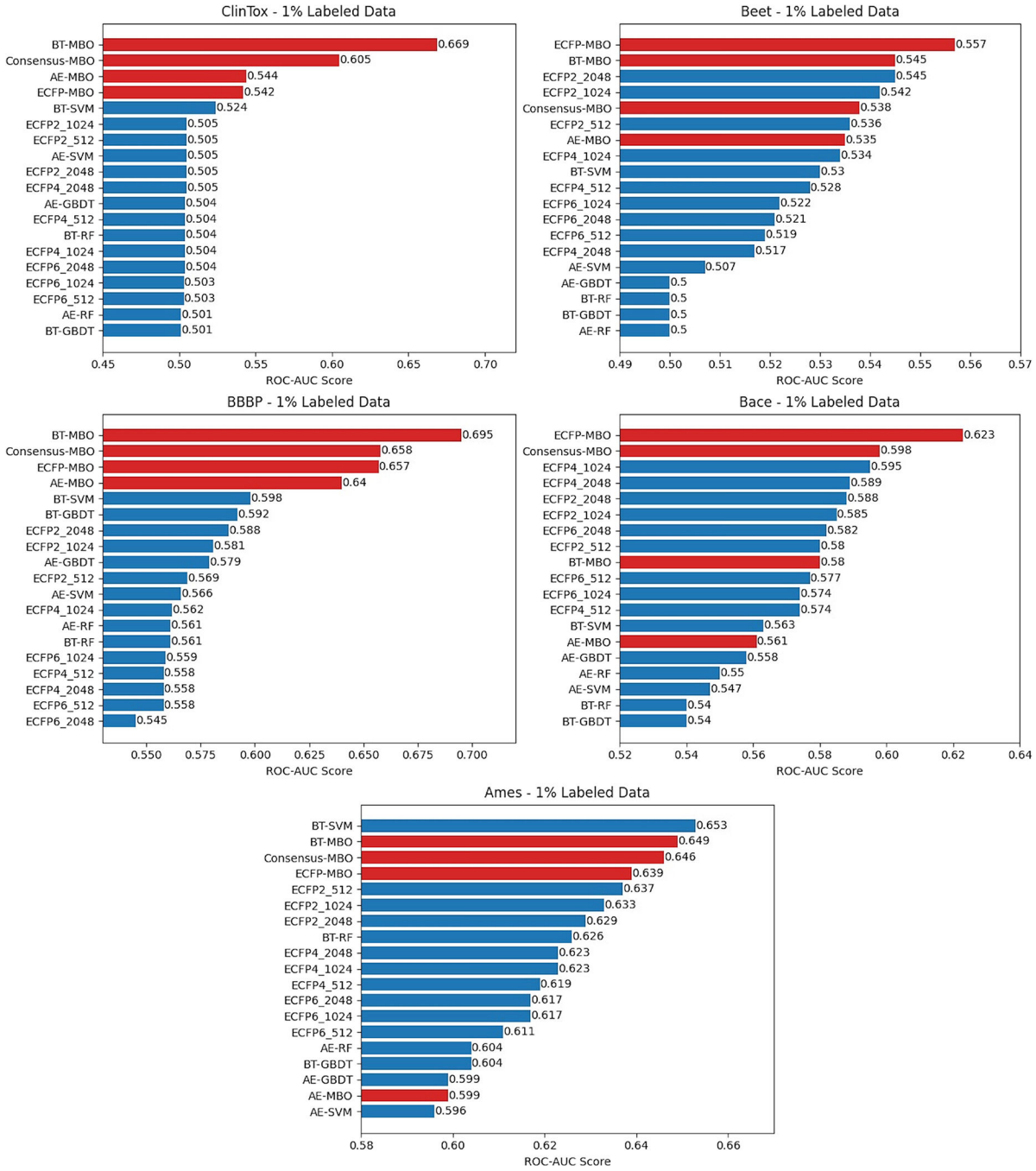


Figure 19.

Comparison of MBO-based proposed methods (shown in red) with other methods (shown in blue) on the five benchmark molecular data sets for 1% labeled data.[456]

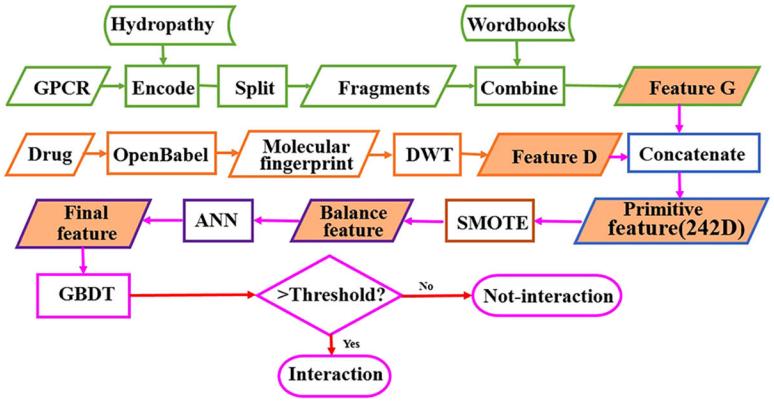


Figure 20.

Illustration of a three-step process that can continuously select features to improve the accuracy of drug interactions during the experiment. In the first step, features are obtained through a GPCR module and merged with molecular fingerprints. Then, SMOTE (synthetic minority oversampling technique) and ANN are employed to generate the final features. Finally, GBDT is used to predict drug interactions. Reproduced with permission from ref 483. Copyright 2021 Frontiers Media SA.

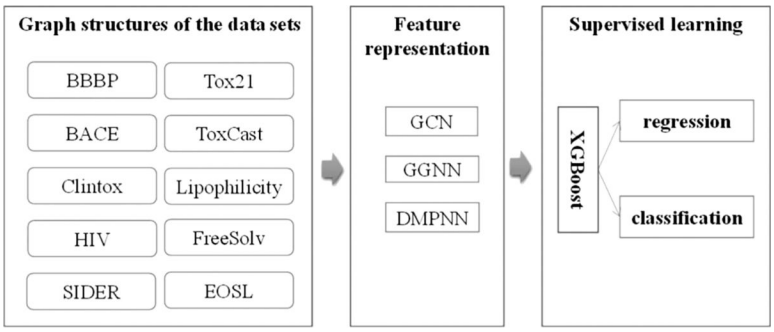


Figure 21.

Workflow of a joint GNN and XGBoost model. Molecular descriptors are extracted by a GNN model, and the prediction is produced by a supervised learner XGBoost for classification or regression. Reproduced with permission from ref 486. Copyright 2021 American Chemical Society.

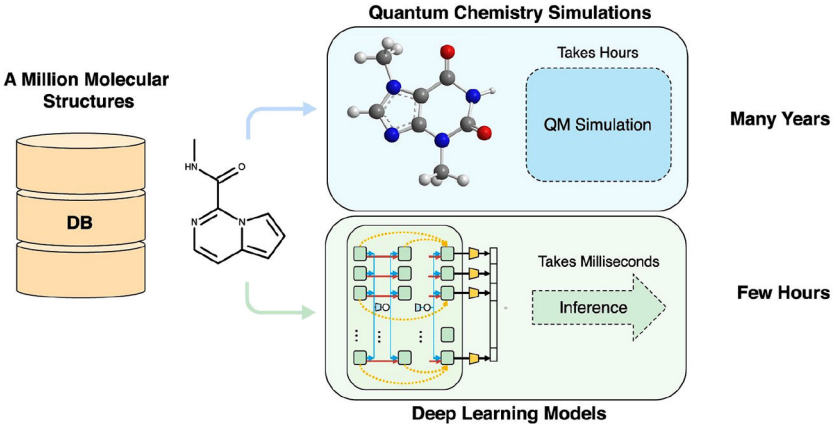


Figure 22.

Illustration of speed differences in computing a given molecular property for a database of 1 million molecules using DFT versus ML. On average, QM simulations require approximately 5 h per molecular structure, leading to a total processing time of 5 × 3600(s) × 10^6 ≈ 500 years. In contrast, a trained DL model needs only 5 ms per molecular structure and just a few hours for 1 million molecules. Reproduced with permission from ref 528. Copyright 2022 American Chemical Society.

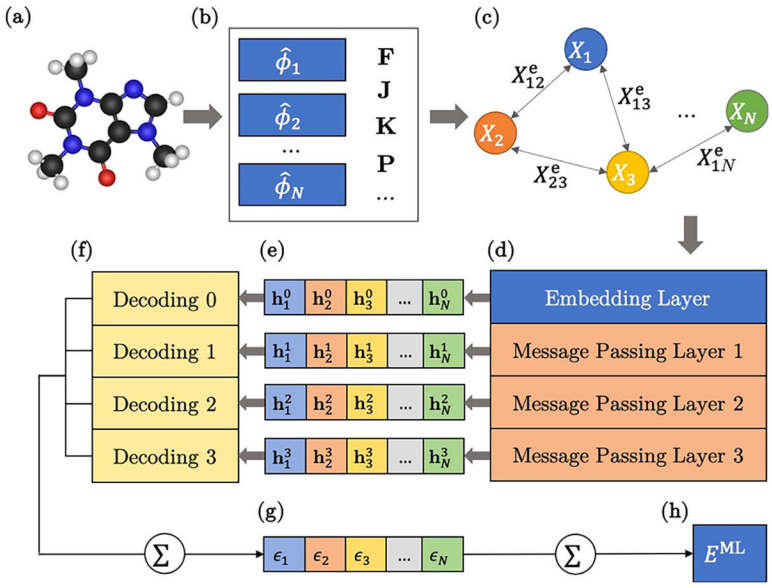


Figure 23.

Illustration of the ORBNET workflow. (a) A low-cost mean-field electronic structure calculation is performed for the molecular system, and (b) the resulting SAAOs and the associated quantum operators are constructed. (c) An attributed graph representation is built with node and edge attributes corresponding to the diagonal and off-diagonal elements of the SAAO tensors. (d) The attributed graph is processed by the embedding layer and message-passing layers to produce transformed node and edge attributes. (e) The transformed node attributes for the encoding layer and each message passing layer are extracted and (f) passed to MPL-specific decoding networks. (g) The node-resolved energy contributions ϵ_u are obtained by summing the decoding networks outputs nodewise, and (h) the final extensive energy prediction is obtained from a one-body summation over the nodes. Reproduced with permission from ref 529. Copyright 2020 AIP Publishing.

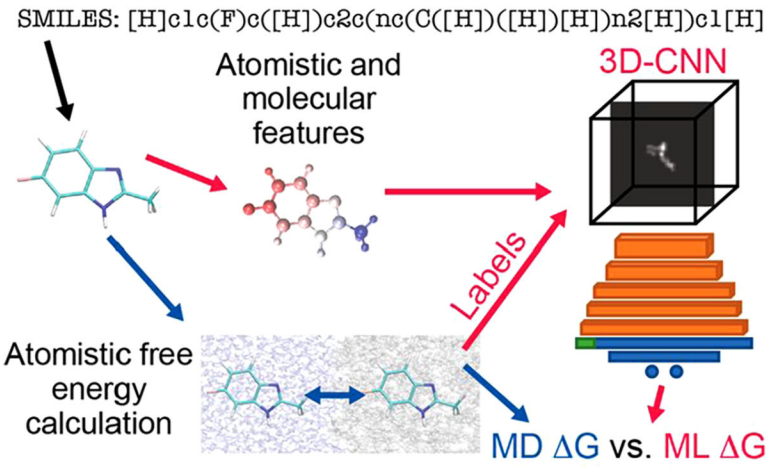


Figure 24.

Illustration of a deep CNN. The relative free energy for moving 15 000 small molecules between water and cyclohexane was computed with atomistic MD simulations. From the simulations, features, such as each atom's partial charge, the average number of water contacts, and molecular features, including the number of hydrogen bonds and size/shape, were extracted. A 3D-CNN and spatial graph CNN were then constructed using the atomic and molecular features to predict the free energies of transfer. Reproduced with permission from ref 530. Copyright 2020 American Chemical Society.

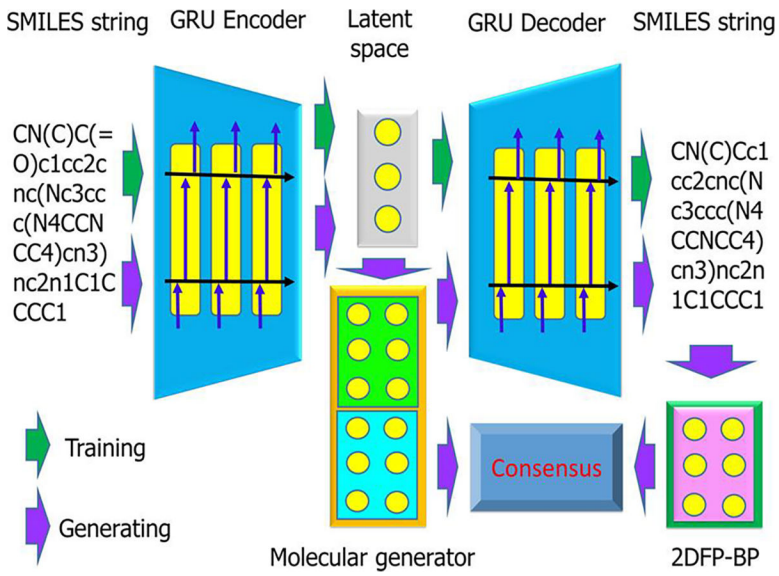


Figure 25.
Schematic illustration of a generative network complex for molecular generation.[568]

Table 1.

Prediction Results Based on Different ML Methods for Identifying Drug–Target Interactions[a]

Method	AUC	Accuracy
Naive Bayes	0.54285	0.445622
neural net	0.55611	0.544142
SVM	0.56119	0.597514
logistic regression	0.62449	0.619996
nearest neighbors	0.71011	0.663864
random torest	0.87473	0.817584
our approach	0.91095	0.871931

[a]Reproduced with permission from ref 40. Copyright 2018 Springer Nature.

Table 2.

Performance in Secondary Structure Prediction by ProteinUnet and SPIDER3-Single on TS1197 and CASP13 According to the Mean Accuracy and SD at the Protein Sequence Level[a]

		TS1197			CASP13		
		mean %	**SD %**	***p*-value**	**mean %**	**SD %**	***p*-value**
Q3	ProteinUnet	73.53	8.70	0.0152	74.39	8.13	0.0128
	SPIDER3-Single	73.18	9.04		75.12	7.65	
Q8	ProteinUnet	61.82	10.86	<0.0001	60.81	12.17	0.8961
	SPIDER3-Single	61.34	11.15		60.81	12.79	

[a]Reproduced with permission from ref 201. Copyright 2021 Wiley.

Table 3.

Performance of Different Classification Strategies for Cancer-Staging Data[284]

	RF	NB	RF with SMOTE	NB with SMOTE	DNN with SMOTE	RF with WGAN	NB with WGAN	DNN with WGAN
accuracy	0.5484	0.3226	0.5806	0.5484	0.6452	0.2903	0.6129	0.7097
F-measure	0.5673	0.1220	0.6254	0.5600	0.6605	0.2931	0.6260	0.7007
g-mean	0.4550	0	0.4811	0.5233	0.6332	0.2627	0.6043	0.6839

Table 4.

Classification Results for Different Classifiers for Cancer Data[a]

classifiers	accuracy (mean)
Decision tree	0.608
KNN ($k = 3$)	0.864
SVM	0.84
VGG	0.781
ResNet	0.849
Gene-GAN (nonamplified)	0.85
Gene-GAN (mixed)	0.892

[a]Reproduced with permission from ref 74. Copyright 2022 Springer Nature.

Table 5.

Comparison of the Results of Three Kinase Data Sets on Different Methods[a]

	JAK1		JAK2		JAK3	
data set	**ROC (%)**	**AP (%)**	**ROC (%)**	**AP (%)**	**ROC (%)**	**AP (%)**
Non-MoLGNN	91.8 ± 0.6	85.4 ± 0.7	87.2 ± 0.3	85.5 ± 0.8	81.2 ± 2.5	80.2 ± 2.0
GINVAE Only	94.2 ± 0.1	88.6 ± 0.2	88.0 ± 0.9	87.3 ± 0.4	84.6 ± 0.2	84.7 ± 0.8
Motif Only	93.4 ± 0.4	87.8 ± 1.1	85.8 ± 0.9	87.5 ± 0.2	86.1 ± 0.1	86.7 ± 0.4
MoLGNN	94.5 ± 0.3	89.2 ± 0.5	89.6 ± 0.2	89.9 ± 0.4	89.2 ± 0.2	89.5 ± 0.2

[a]Reproduced with permission from ref 335. Copyright 2020 NeurIPS.

Table 6.

Comparison of Prediction Accuracy on Different Data Sets[a]

data sets algorithms	breast cancer (%)	glioblastoma (%)	lung cancer (%)
proposed DRL model	98.3	99.2	97.34
SVM	91.32	92.34	93.42
RF	78.9	81.23	82.34
ANN	94.5	93.47	94.5

[a]Reproduced with permission from ref 379. Copyright 2022 Springer Nature.

Table 7.

Performance of the Three Models on Six Data Sets[a]

data set	graph convolution with dropout	semi-supervised with dropout	semi-supervised with SVGD
FreeSolv	0.531 ± 0.061	0.439 ± 0.093	0.688 ± 0.053
ESOL	0.112 ± 0.035	0.306 ± 0.079	0.553 ± 0.026
CatS	0.049 ± 0.036	0.066 ± 0.044	0.310 ± 0.019
MeltingPoint	0.192 ± 0.016	0.284 ± 0.035	0.337 ± 0.013
p450	0.167 ± 0.015	0.185 ± 0.049	0.213 ± 0.010
malaria	0.315 ± 0.028	0.317 ± 0.031	0.378 ± 0.019

[a]Reproduced with permission from ref 441. Copyright 2019 Royal Society of Chemistry.

Table 8.

Results of the R^2 Comparison of Accuracy between the Model in the Text and Other Methods for Molecular Property Predictions[341]

method	IGC_{50}	LC_{50}	LC_{50}-DM	average
BTAMDL 2	0.793	0.778	0.741	0.771
BTAMDL 1	0.795	0.776	0.733	0.768
MDL consensus	0.792	0.772	0.721	0.762
GBDT consensus	0.777	0.692	0.472	0.647
hierarchical[482]	0.719	0.710	0.695	0.708
single-model[482]	NA	0.704	0.697	0.701
FDA[482]	0.747	0.626	0.565	0.646
group contribution[482]	0.682	0.686	0.671	0.680
nearest neighbor[482]	0.600	0.667	0.733	0.667
test consensus[482]	0.764	0.728	0.739	0.744
3D MDL consensus[338]	0.802	0.789	0.678	0.765

Table 9.

An Overview of Major Machine Learning and Deep Learning Approaches in Different Fields with a Variety of Algorithms for Small Data Challenge

applied field	algorithm	ref
Basic Machine Learning Algorithm Approach		
drug–target interaction	gradient boosted decision trees (GBDT)	40
drug-induced ototoxicity	support vector machine (SVM), message-passing neural networks (MPNNs)	111
compound activity	random forest (RF), *k*-nearest neighbor (KNN)	114
miRNA expression	random forest (RF)	115
Artificial Neural Networks Approach		
molecule lipophilicity	MRlogP, a neural network-based predictor of log *P*	131
molecule lipophilicity	multiple linear regression (MLR) and artificial neural network (ANN)	132
molecule lipophilicity	GA-MLR and GA-ANN	133
pharmacokinetics	quantitative structure–pharmacokinetic relationship model	134
dislocation nucleation	artificial neural networks (ANNs), random forest (RF), support vector machine (SVM)	135
molecular dynamics simulations	*k*-nearest neighbor (k-NN) and artificial neural network (ANN)	136
Convolutional Neural Networks Approach		
drug-induced liver injury	natural language processing (NLP) inspired convolutional neural networks (CNNs)	168
environmental applications	molecular image-convolutional neural networks (CNNs) with transfer learning	169
molecular dynamics simulations	deep learning encoder–decoder convolutional neural networks (CNNs)	172
U-Net Approach		
binding sites prediction	Voxel-based 3D U-Net	200
protein structure prediction	single-sequence-based U-Net convolutional network	201
medical image segmentation	an automatic liver parenchyma segmentation network based on the U-Net architecture	202
Graph Neural Networks Approach		
molecular property prediction	property-aware relation networks with graph neural networks-based classifier	229
machine learning algorithm	model agnostic meta-learning (MAML) and first-order MAML (FO-MAML)	231
molecular property prediction	meta-MGNN	232
Long Short-Term Memory Approach		
protein structure prediction	deep asymmetric convolutional LSTM neural network (DeepACLSTM)	258
medicinal science	deep learning long short-term memory (DL-LSTM)	263
protein remote homology detection	ProDec-BLSTM	264
anticancer peptide prediction	bidirectional long short-term memory (BLSTM)	272
short-term load forecasting	bidirectional LSTM	273
Generative Adversarial Networks Approach		
protein solubility prediction	protein log *S* generative adversarial nets (Pro-GAN)	281
multiclassification for cancer staging	Generative Adversarial Network (GAN) combined with a deep neural network (DNN)	284
Generative Adversarial Networks Approach		
cancer classification	Gene-GAN Wasserstein generative adversarial	74
cancer prognosis prediction	network-based deep adversarial data augmentation (wDADA)	289
brain network	BrainNetGAN	290
antiviral drug	sequence-based binary classifier	291

applied field	algorithm	ref
Autoencoders Approach		
materials science	variational autoencoder (single task learning) (VAE), VAE-L with linear regression (multitask learning), VAE-NL with nonlinear regression (multitask learning)	64
biopolymerization	VAE+ANN, VAE+GAN, VAE+RF, GAN+RF, ANN, RF	41
drug discovery	autoencoder-assisted multitask ANN	315
Transformers Approach		
drug discovery	self-supervised Motif Learning Graph Neural Network (MoLGNN)	335
molecular property prediction	algebraic graph-assisted bidirectional transformer	334
drug repositioning	multitask self-supervised learning	344
Reinforcement Learning Approach		
molecular property prediction	deep reinforcement learning with new reward function	372
human microRNA-disease association	RFLMDA (combining the Q-learning algorithm with reinforcement learning)	373
molecule design	deep reinforcement learning	378
cancer type classification	deep reinforcement learning	379
Transfer Learning Approach		
pharmacokinetic parameter prediction	integrated transfer learning and multitask learning approach	412
pharmacogenomics	adversarial inductive transfer learning (AITL)	413
retrosynthetic analysis	sequence-to-sequence (seq2seq) transfer learning	416
chemical reaction prediction	transformer-transfer learning	572
molecular property prediction	MRlogP (neural network-based predictor of log *P*)	418
protein structure prediction	algebraic topology-based multitask and multichannel CNN	419
Active Learning Approach		
domain applicability	semi-supervised learning and Bayesian deep learning	441
molecular property prediction	active learning -based semi-supervised GNN	573
Graph-Based Semi-supervised Learning Approach		
graph-based semi-supervised learning	an autoencoder coupled with Merriman Bence Oshe scheme (AE-MBO) and a bidirectional encoder transformer coupled with Merriman Bence Oshe scheme (BT-MBO)	456
machine learning algorithm	Multikernel manifold learning (MML) and multiscale MBO (MMBO)	458
machine learning algorithm	Poisson Merriman Bence Oshe (MBO)	460