KEYNOTE (GREEN)

Deep learning methods for molecular representation and property prediction

Zhen Li [a], Mingjian Jiang [c], Shuang Wang [d], Shugang Zhang [b,*]

[a] College of Computer Science and Technology, Qingdao University, Qingdao 266071, China
[b] College of Computer Science and Technology, Ocean University of China, Qingdao 266100, China
[c] School of Information and Control Engineering, Qingdao University of Technology, Qingdao 266033, China
[d] College of Computer Science and Technology, China University of Petroleum, 266580 Qingdao, China

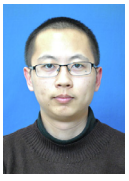
With advances in artificial intelligence (AI) methods, computer-aided drug design (CADD) has developed rapidly in recent years. Effective molecular representation and accurate property prediction are crucial tasks in CADD workflows. In this review, we summarize contemporary applications of deep learning (DL) methods for molecular representation and property prediction. We categorize DL methods according to the format of molecular data (1D, 2D, and 3D). In addition, we discuss some common DL models, such as ensemble learning and transfer learning, and analyze the interpretability methods for these models. We also highlight the challenges and opportunities of DL methods for molecular representation and property prediction.

Keywords: Molecular representation; Deep learning; Self-supervised learning; Drug discovery; Property prediction

Introduction

Molecular properties are important factors in many fields, including chemistry, drug discovery, and healthcare, and are related to quantum mechanics, physical chemistry, biophysics, physiology, and so on.[1] Computer-aided methods are able to predict molecular properties quickly, providing overviews of the molecules of interest before specific experiments begin. Such approaches are referred to as quantitative structure–activity relationship (QSAR) or quantitative structure–property relationship (QSPR) models. Furthermore, with the development of machine learning (ML) methods, the accuracy and speed of molecular property prediction (MPP) have also improved, accelerating other related applications, such

Zhen Li is an associate professor with Qingdao University. His research interests include graph convolution models, machine learning, and bioinformatics. He is currently focusing on deep learning methods for computer-aided drug discovery.



Mingjian Jiang is a lecturer with Qingdao University. His main research interests include virtual screening, molecular design, and drug–target affinity prediction.

Shuang Wang is currently a lecturer with the China University of Petroleum (East China). Her research interests mainly include deep learning-based drug design, such as for drug design, and molecular property and drug–target affinity predictions.

Shugang Zhang is a lecturer with the Ocean University of China. His research interests include computational cardiology and AI-based drug discovery. He is currently focusing on *in silico* drug design and protein function prediction with deep learning approaches.

* *Corresponding author.Zhang, S. (zsg@ouc.edu.cn)*

https://doi.org/10.1016/j.drudis.2022.103373

as drug–target affinity predictions[2,3] and molecular synthesis predictions.[4,5] In particular, DL methods, an important branch of ML, have received significant attention.[6] Such approaches allow the more precise discovery of the relationships between a structure and its properties.

The first decision in MPP using DL models is how to represent a molecule. The molecular formula is a common representation for molecules (e.g., $C_{30}H_{35}N_7O_4S$ represents imatinib mesylate); however, such representation is difficult for DL models to predict the properties of molecules because of the lack of structural information. Therefore, a more advanced sequence-based representation, namely the Simplified Molecular-Input Line Entry System (SMILES),[7] was proposed and has become a popular representation of molecules. In a SMILES string, atoms and chemical bonds are represented by letters and punctuation, respectively, and branches are described using parentheses. In Fig. 1a, imatinib mesylate is converted into a SMILES string. However, because a SMILES string mgiht not correspond to a valid molecule, self-referencing embedded strings (SELFIES)[8] were proposed to solve this problem, whereby each SELFIES string corresponds to a valid molecule.

In addition, fingerprints are another type of sequence-based molecular representation that incorporates molecular structure information, such as extended connectivity fingerprints (ECFP)[9] and molecular access system (MACCS)[10] (Fig. 1). They are normally used as input for traditional ML methods[11] or as the auxiliary input combined with other types of data.[12] In recent years, the mathematical representation of molecules, including topological and geometric models, has achieved enormous success.[13–21] Nguyen *et al.*[22] used algebraic topology-based representations to characterize molecules, and persistent homology was introduced to enrich the topological representation.

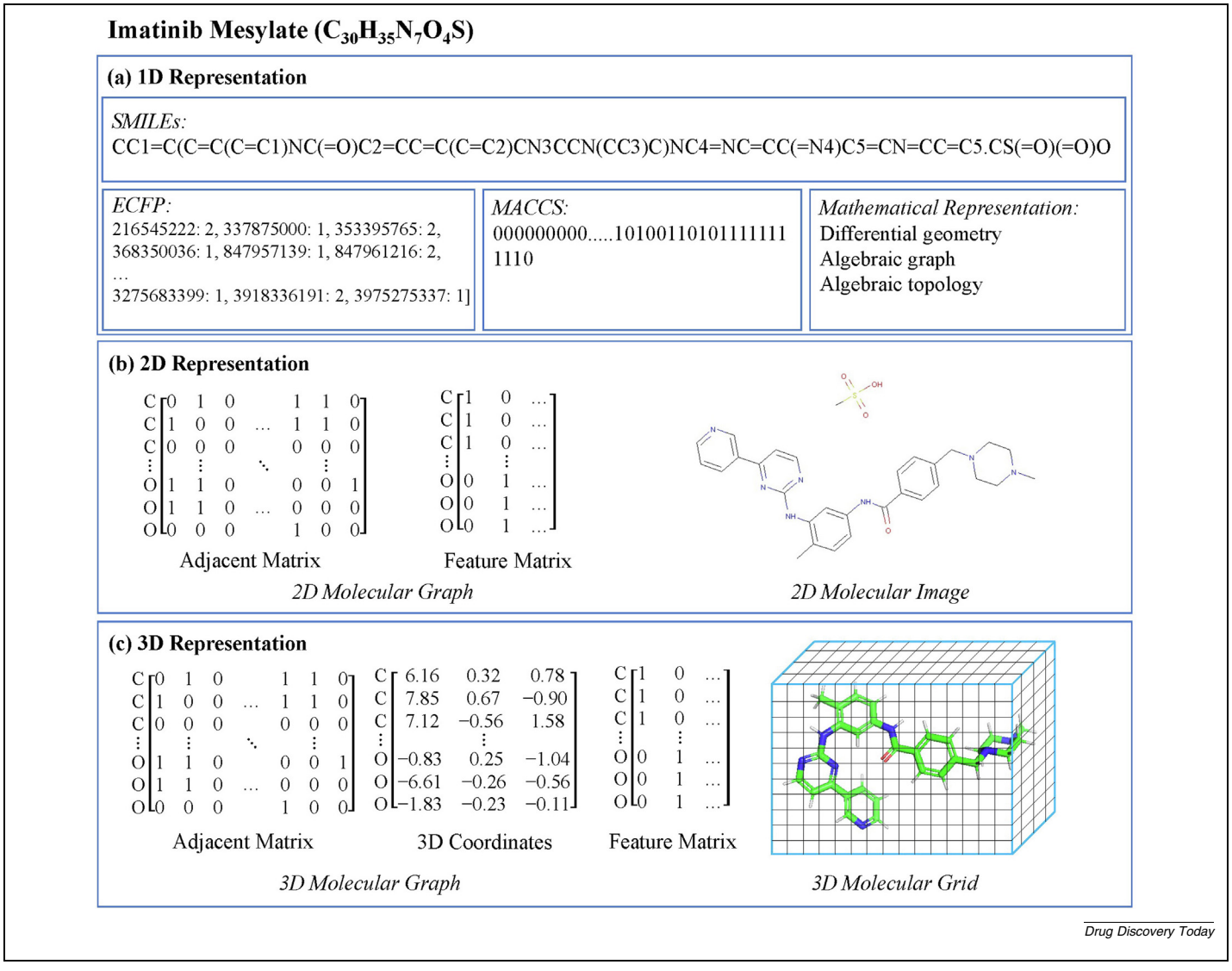


FIGURE 1

1D, 2D, and 3D representations of imatinib mesylate ($C_{30}H_{35}N_7O_4S$). **(a)** For the 1D representation, several representation forms are demonstrated here, including Simplified Molecular-Input Line Entry System (SMILES), the extended connectivity fingerprint (ECFP) and molecular access system (MACCS) fingerprint, and some mathematical representation methods. **(b)** For the 2D representation, the molecular graph can be presented in the form of two matrices (i.e., the adjacent matrix and the feature matrix). The molecular image on the right (generated by RDKit[68]) is another type of 2D representation for imatinib mesylate. **(c)** Two 3D representation methods: a 3D molecular graph and a 3D molecular grid (generated by PyMOL[116]).

MathDL[23] combined differential geometry, algebraic graphs, and algebraic topology to form rotational and translational invariant molecular fingerprints. Based on persistent homology, Liu *et al.*[24] developed a persistent spectral hypergraph model, in which persistent attributes were used as fingerprints.

Although the SMILES string is simple and fast, it is still not able to capture the spatial relationship between atoms comprehensively. As shown in Fig. 1, in the C = C5 benzene ring, five carbon atoms are grouped into C5; however, different carbon atoms have different relationships with other atoms and are located at different places in the molecule. Furthermore, they might correspond to different properties. Thus, using only SMILES is not enough to predict certain properties.

The structural information embedded in the molecular 2D data is helpful in MPP, which is divided into two types (molecular graph data and molecular image data; Fig. 1b). Graph data are an efficient way to learn molecular representation. The atoms of a molecule are regarded as nodes in a molecular graph, whereas the chemical bonds are regarded as edges. With the development of graph convolutional networks (GCNs),[25,26] the information of neighboring nodes can be gathered more directly and efficiently, which is useful in capturing spatial relationships among atoms within a molecule. At the same time, the molecular image, obtained by converting the molecule into a pixel-based rasterization image, is another 2D representation format for molecules, with each pixel in the image representing a bond, atom, or blank background. Given the development of DL methods in the image field, researchers have attempted to transfer these methods to molecular images for MPP.

The 3D structure provides the most detailed information about a molecule. Analogous to 2D molecular data, there are two types of 3D molecular data: 3D molecular graphs and 3D molecular grid (Fig. 1c). The 3D molecular graph records the 3D locations of each atom, and the 3D molecular grid is a special 3D image in which the voxels in the grid indicate different elements or attributes of molecular conformation through different methods.

In this review, we highlight DL models using for molecular representation. We first introduce molecular representation and property prediction methods and highlight newly emerging DL methods, such as ensemble learning and transfer learning, which have been used to solve some common problems in molecular representation. We also present a brief overview of the interpretability of DL models and highlight associated challenges and future research avenues.

Sequence-based methods

SMILES is the most direct and simple way to depict a molecule. It is similar to natural language, in which each atom is a word in the sentence. Given rapid progress in the natural language processing (NLP) area, NLP methods could be applied for the embedding of a SMILES sequence.

Data augmentation methods

Inconsistency in SMILES has to be overcome before they can be processed using DL models. For a molecule, there could be many valid SMILES sequences depending on the SMILES grammar. A starting atom and a traverse order might correspond to a sequence; thus, we can choose any atom as the starting point and any branch as the first one to go through. The canonical SMILES ensures that each molecule has only one SMILES string according to certain rules. However, various SMILES formats for the same molecule could enhance the learning ability of the DL model when non-canonical SMILES are used as inputs. This is because non-canonical SMILES can also benefit DL models by offering latent features associated with the grammar of SMILES and chemical properties. Therefore, data augmentation or enumeration is recommended to enlarge the coverage of strings, thereby ensuring that the model is able to learn multiple strings of a molecule.

Given that each molecule has different lengths, there are fewer possible notations of short strings compared with long strings. In Conv2S[27], the SMILES strings were generated randomly and continuously until $L^N/(L + 1)$ was < 1 %, where L and N are the length and number of generated SMILES, respectively. To overcome the problem of unbalanced data sets, molecules with fewer SMILES strings are complemented by the repetitive SMILES strings to ensure that all molecules have the same number of SMILES strings. Kimber *et al.*[28] conducted a comprehensive analysis of five different SMILES augmentation methods. They found that the augmentation method improved the performance of the DL model, and the results achieved using canonical SMILES were better than using single random SMILES.

Convolutional neural network models

Convolutional neural networks (CNNs) can be used for sequence data processing. For example, the ConvS2S[27] model converts SMILES to an integer list and then adds the position embedding to inform the model of the position of the corresponding letter. Lim *et al.*[29] also performed a character-level embedding of SMILES, in which embedding vectors were generated for each letter. A CNN layer with a multihead self-attention module was introduced to process the input embedding, and two fully connected layers were added to output the prediction. SMILES convolution fingerprints (SCFPs)[30] were combined multiple atom properties, including type, degree, charge, and chirality, to form a feature vector of an atom. The SMILES sequence could be converted into a matrix, the length of which was the maximum length of the SMILES sequence. Two convolutional and pooling layers with a subsequent global pooling layer were constructed to extract the representation, and the large contribution of the corresponding filter indicated the important substructure by backtracking.

Given that CNN-based methods require a fixed length of input samples, the drug SMILES must be padded or truncated before being sent into the network. Typically, the maximum or average length of the SMILES strings in the data set can be chosen as the fixed length of the model input samples. However, both methods will result in data loss and introduction of noise, one of the major problems of CNN-based methods.

Recurrent neural network models

Recurrent neural networks (RNNs) and variants, such as long short-term memory (LSTM) and gated recurrent unit (GRU), are widely used in NLP to process sequence data. To process molecular sequence data, an accurate and robust RNN model for

SMILES sequence is crucial to extract features of molecules. Hou *et al.*[31] proposed a bidirectional-LSTM (Bi-LSTM) with a channel and spatial attention network improved by Bayesian optimization, which could specifically identify the prime factors in the SMILES sequence. Nazarova *et al.*[32] proposed two backpropagation methods of RNN and compared the performance of binary and decimal representation of SMILES in polymer property prediction. They found that the binary representation was more accurate compared with the decimal one. The combination of CNN and RNN can also improve the performance of representation. Li *et al.*[33] converted each character in SMILES to a vector using one-hot encoding, and a hybrid architecture of stacked CNN and RNN layers was introduced for representation extraction. It was reported that CNN was able to improve the performance of prediction compared with plain RNN models.

Although the RNN is suitable for sequence processing, using only sequences and ignoring other information (e.g., chemical context or molecular structures) is not a comprehensive way to learn the molecular representation. In particular, the atomic relation, atomic group, as well as bond types within a molecule might also be of relevance to the molecular property, which could be introduced in certain ways to improve the performance. Moreover, the interpretability of the model using sequence is still defective. Given that the molecular branch is fused with the main sequence, it is difficult for the RNN model to distinguish the motif and branch without other more specific settings; thus, the key atoms from the same functional group might be located far away from each other. Even if the attention mechanism is introduced, it only focuses on single letters or adjacent letters.

Substructure learning methods

Normally, the functional group is the key part of the molecule, and the molecular properties and activities are highly related to functional groups and substructures. However, the SMILES sequence does not contain this type of information directly; thus, methods focusing on the functional groups hiding in SMILES have been developed. SMILES pair encoding[34] learns a vocabulary of high-frequency SMILES substrings and convert SMILES according to the learned vocabulary, which can be fed into DL models. Mol2Context-vec[35] extracts the substructure with the help of ECFP. The substructure comprises multiple atoms, including a central atom and all atoms within a given radius surrounding the central atom. Each substructure has its own identifier. The substructure sequence is the input of a Bi-LSTM, which captures the interactions between atomic groups. Mol-BERT[36] summarizes the vectors of each substructure as a molecular representation for unsupervised learning and downstream tasks. S2DV[37] defines a split character to retain the substructure information, and a sliding window with predefined size to process the sequence. Branch chains and double characters indicating a single atom are replaced by a single character, and each branch is expanded and processed in the same way. Each sliding window generates a vector to depict the relationship between the compound and its substructure.

Sequence-based self-supervised learning methods

In recent years, self-supervised learning (SSL) has developed rapidly. SSL can use a large amount of unlabeled data sets and design the pretext task to learn the intrinsic feature of data, thereby reducing the demand for labeled samples.[38] Generally, sequence-based SSL models can be divided into contrastive and generative learning methods.[39] The contrastive learning method is to construct pseudo-label data to learn the difference between the positive and negative samples, whereas the generative method encodes the input to latent features and decodes it to reconstruct the input, whereby the latent features can then be used as the representation of the input. From the data information perspective, the contrastive learning method tries to find the interdata information, whereas the generative learning method focuses on the intradata information.[40]

In the NLP area, the bidirectional encoder representation from transformers (BERT)[41] is a widely used SSL method to learn the feature of words, in which a transformer comprises an encoder and a decoder. BERT-like methods can be applied in SMILES sequences to extract atom or molecular features (Fig. 2a). MOL-BERT[42] combines three tasks to generate the molecular representation. The first is the masked language model (MLM) used in BERT. The second one is the SMILES equivalence method, which uses two SMILES sequences from the same molecules as one class, and two SMILES sequences of different molecules as the other class for training. The third uses molecular chemical characteristics for prediction. All three models are trained jointly to output the molecular representation. SMILES-BERT[43] is also based on BERT, but only the MLM has been retained and a self-attention layer has been introduced to use the sequential information.

In addition to BERT-based methods, there are other generative methods using the encoder and decoder architecture for molecular representation. Hu *et al.*[44] used a GRU-based encoder-decoder generative model to generate latent features of a fixed size to represent molecules from SMILES, and a CNN model was introduced for downstream prediction tasks. The Molecular Prediction Model Fine-Tuning (MolPMoFiT)[45] used the language model to predict the next word according to a sequence of words, which could also extract features for downstream tasks.

The translation model used in NLP can also be implemented in the SMILES sequence data (Fig. 2b). Sequence-to-sequence (seq2seq) is a popular tool containing an encoder and a decoder for the translation task, the object of which is to convert a sequence to another sequence. Similarly, the output of the encoder could also be the representation for other tasks. When using the seq2seq method for SMILES, the main objective is to find two corresponding sequences for training. Winter *et al.*[46] proposed a method for translating two semantically equivalent representations of molecular structures (i.e., SMILES and IUPAC name). Transformer-CNN[47] trained a transformer model to conduct a SMILES canonicalization task, in which the input was noncanonical SMILES and the output was the corresponding canonical SMILES.

Thus, SSL is a promising method because of its ability to discover the inner features of input data without labels. For sequence-based SSL methods, contrastive learning methods that help to find the inter-relationships of SMILES strings are still lacking, which is a future direction for molecular representation research.

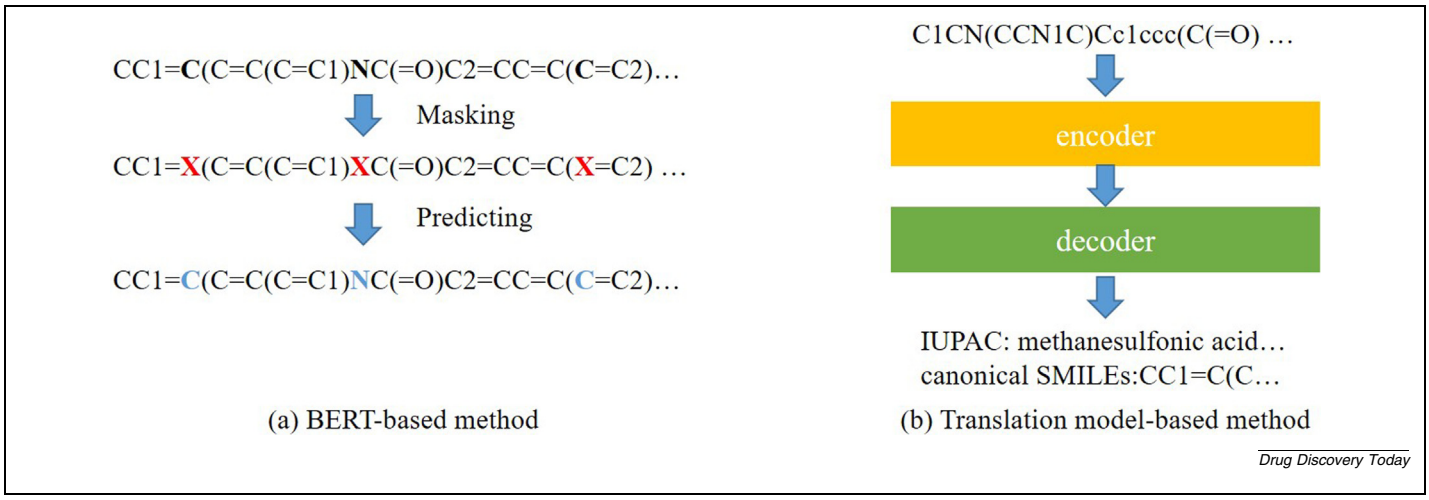


FIGURE 2

Different types of self-supervised learning (SSL) method in Simplified Molecular-Input Line Entry System (SMILES). **(a)** Bidirectional encoder representation from transformers (BERT)-based method. The SMILES sequence is used as input and some atoms are randomly masked. The language model is then trained to predict these masked items for representation learning. **(b)** Translation model-based method. The model is trained to translate the input SMILES sequence to another type of sequence. The latent feature encoded by the encoder is used as the molecular representation.

Graph-based methods

Graphs are a more direct structure that can store and represent most structural information. In the graph model, atoms are set as nodes and bonds are set as edges, and each node has its own feature. With the help of graph data, structural information within molecules can be used by GCNs, which is designed for non-Euclidean graph data. They are capable of capturing information on the relationship between connected nodes. Generally, there are two types of GCN: spatial convolution and spectral convolution. The former updates the feature of each node by gathering information about its neighboring nodes using certain message-passing rules in the spatial domain. The latter converts the graph data into spectral domain by performing eigenvalue decomposition on Laplacian matrices.

Spectral GCN models

We introduce the spectral GCN methods firstly. LanczosNet[48] uses the Lanczos algorithm to build the low-rank approximations of Laplacians for graph spectral convolution, which could be used to exploit multiscale information and design learnable spectral filters. Shang *et al.*[49] proposed a consistent edge-aware multi-view spectral GCN model with a new flexible spectral filter from the Chebyshev approximation; the molecular graph was decomposed into multiple views of the graph according to the type of edges, and a consistent edge-mapping method that learned the attention weights of edges was used to ensure the edge consistency.

Thus, there are fewer spectral methods than spatial methods in molecular representation and property prediction. This is because molecules with different atoms will produce graphs of different sizes, whereas spectral GCN models can only handle graphs of a fixed size. As a result, data alignment operations, such as padding or truncation, are still needed when processing the input graph data samples, which will impair the data integrity and affect the final performance of the models.[25].

Spatial GCN models

Spatial GCN models are more widely used in drug discovery and MPP. Generally, spatial GCN models require two matrices as inputs: an adjacent matrix and a feature matrix. The former indicates the spatial interconnection of atoms within a molecule and can be obtained from molecular graphs, and the latter is normally defined by different methods. DeepAtomicCharge[50] uses the message-passing neural network (MPNN) with skip connection to predict the atomic charge. AttentiveFP[51] is another molecular representation method derived from GCN, which automatically learns nonlocal intramolecular interactions and captures the hidden edges from specified tasks through an attention mechanism. Multiphysical GNN[52] combines the scale-specific graph neural network and the element-specific graph neural network to capture various atomic interactions from different scales for multiphysical representations.

The edge is also regarded as an important element that should be considered in the convolution process. Cross-dependent graph neural networks[53] consider atoms and bonds equally important. Both atom-central and bond-central views are constructed, and a cross-dependent message-passing scheme is proposed between the two views. TrimNet[54] proposes a triplet-attentive edge network to gather information through atom–bond–atom arrangements to improve the extraction of edge information. A pair of atoms along with the bond between them are concatenated into a triplet, and the multi-head attention is used to gather the message of a node from its neighboring nodes and edges.

Directed graphs are a special type of graph that contain directed edges indicating the direction of message passing. Such property is often used to handle the problem of oversmoothing, which frequently occurs in graph-learning models. For example, the Edge Memory Neural Network[55] focuses on passing messages of edges rather than nodes. Each edge owns two states corresponding to two opposite directions; thus, each state is updated

based only on its upstream nodes, thereby avoiding the disappearance of useful information brought by the oversmoothing problem.

Tree-based methods

A graph can be transformed into a tree structure by selecting a starting atom, which breaks the circle of the graph and provides another view of it. Moreover, according to a breadth-first search (BFS) or depth-first search (DFS) method, a tree can be converted to a sequence of atoms using an RNN model for representation learning.

Su *et al.*[56] and Wang *et al.*[57] both developed QSAR modeling methods based on the molecular tree structure. The molecules are encoded to signature descriptors, and the tree-structured long short-term memory (Tree-LSTM), which is good at capturing long-range dependencies, was used to depict molecular tree data structures and correlate them to molecular properties.

Junction Tree[58] decomposes the molecule into substructures first, and then generates a tree-structured graph based on these substructures. Although the Junction Tree was proposed for molecular generation, the encoder part can be isolated for property prediction. Inspired by this, Wang *et al.*[59] proposed a multichannel tree-based molecular prediction method. A molecule is transformed into a substructure-based graph, and a BFS method is applied to traverse the graph to generate a tree structure. A GRU-based neural network with attention mechanisms is then applied to learn the molecular features at multiple levels.

However, because of the different traversing methods, such as BFS or DFS, and different root atom selection methods, the tree structure is not unique. Multiple structures of the generated molecular tree will affect the generalization of the model. The definition of the canonical structures is not robust and cannot ensure that all property-related information, especially the connection information, is comprehensively exhibited in the structures because transferring a graph to a tree will have to break down one or more connections.

Graph-based self-supervised learning methods

Similar to the performance on sequence data, SSL methods also achieved remarkable performance on graph data. Wu *et al.*[40] defined another type of SSL method on graph data, the predictive method, which designs prediction-based pretext tasks based on self-generated labels.

In the contrastive learning method, how to construct positive and negative samples is a crucial step. In this regard, MolCLR[60] augmented the molecules in three ways: atom masking, bond deletion, and subgraph removal. Samples from the same molecule were denoted as positive pairs, and the others were denoted as negative pairs for contrastive loss calculation. However, MOCL[61] argues that augmentation methods, such as node dropping, edge perturbation, and subgraph extraction, will affect the properties of the molecule and are not suitable for contrastive learning. Instead, MOCL adopts a substructure substitution strategy, whereby a substructure is replaced by a bioisostere that shares similar properties. In the NLP domains, many features are extracted at the word and sentence levels. These two levels in NLP are analogous to the node and graph levels in a graph, which represent the local and global features, respectively. Li *et al.*[62] pretrained a model at both the node and graph level. At the graph level, each molecule was decomposed into two parts, and the model needed to predict whether the two parts came from the same molecule. The process of contrastive learning method on molecular graph is shown in Fig. 3a.

The generative method reconstructs the input through a encoder–decoder model. The molecular graph BERT[63] combines local message passing and GNN into the BERT model for pretraining. Koge *et al.*[64] used a molecular hypergraph grammar variational autoencoder (VAE)[65] to extract the embedding of molecules, which embedded the molecular structures and physical properties into the latent feature of VAE. The process of generative learning method based on molecular graphs is shown in Fig. 3b.

The predictive model normally used in MPP is used to predict the type or attribute of a node or an edge. GROVER[66] combines two-level SSLs. The first SSL task, defined at the node/edge level, was to predict the property of a subgraph. The second SSL task, defined at the graph level, was to predict the occurrence of different motifs; combining both two levels could provide structural and semantic information about a molecule. Moreover, the concatenation of SSL and supervised learning provides a new way to understand a molecule. For example, SUGAR[67] combines both supervised and SSL methods based on subgraphs, and the two losses coming from the classification and mutual information maximization are grouped into the final loss function. The evaluation results demonstrated that the model performance was improved by the introduction of SSL. The process of predictive learning method based on molecular graphs is shown in Fig. 3c.

Image-based methods

DL methods have achieved great success in the image processing field, which also sheds light on QSAR/QSPR. More concretely, the molecules can be converted to images, so that traditional DL models can be used for QSAR/QSPR tasks, among which CNN is the most widely adopted for molecular feature extraction.

For image generation-based methods, the simplest way is to use molecular images directly, which can be converted by softwares, such as RDKit[68] and Open Babel.[69] However, this type of image introduces a large blank area without valid information. In addition, scale sensibility is another problem because all molecules are converted into images of the same size. In other words, the size of the same atom/structure is vibrational in different molecules because of the fixed size of the whole molecular image.

Other image generation methods try to avoid this problem. Yoshimori *et al.*[70] generated one map for each atom, and formed a molecular topographic map by adding all atomic maps together as a 28 × 28 heat map. MolMap[71] maps molecular descriptors and fingerprint features into 2D feature maps, resulting in a method that combines hand-crafted features into a 2D space to capture the intrinsic correlations of molecular features.

The frequency domain is another important field in image processing. Tchagang *et al.*[72] converted molecules to images through frequency-domain methods. They first converted the molecule to a 1D Coulomb matrix, and a time–frequency-like (TFL) method was then introduced to generate a TFL image, which could encode the structural, geometric, energetic, elec-

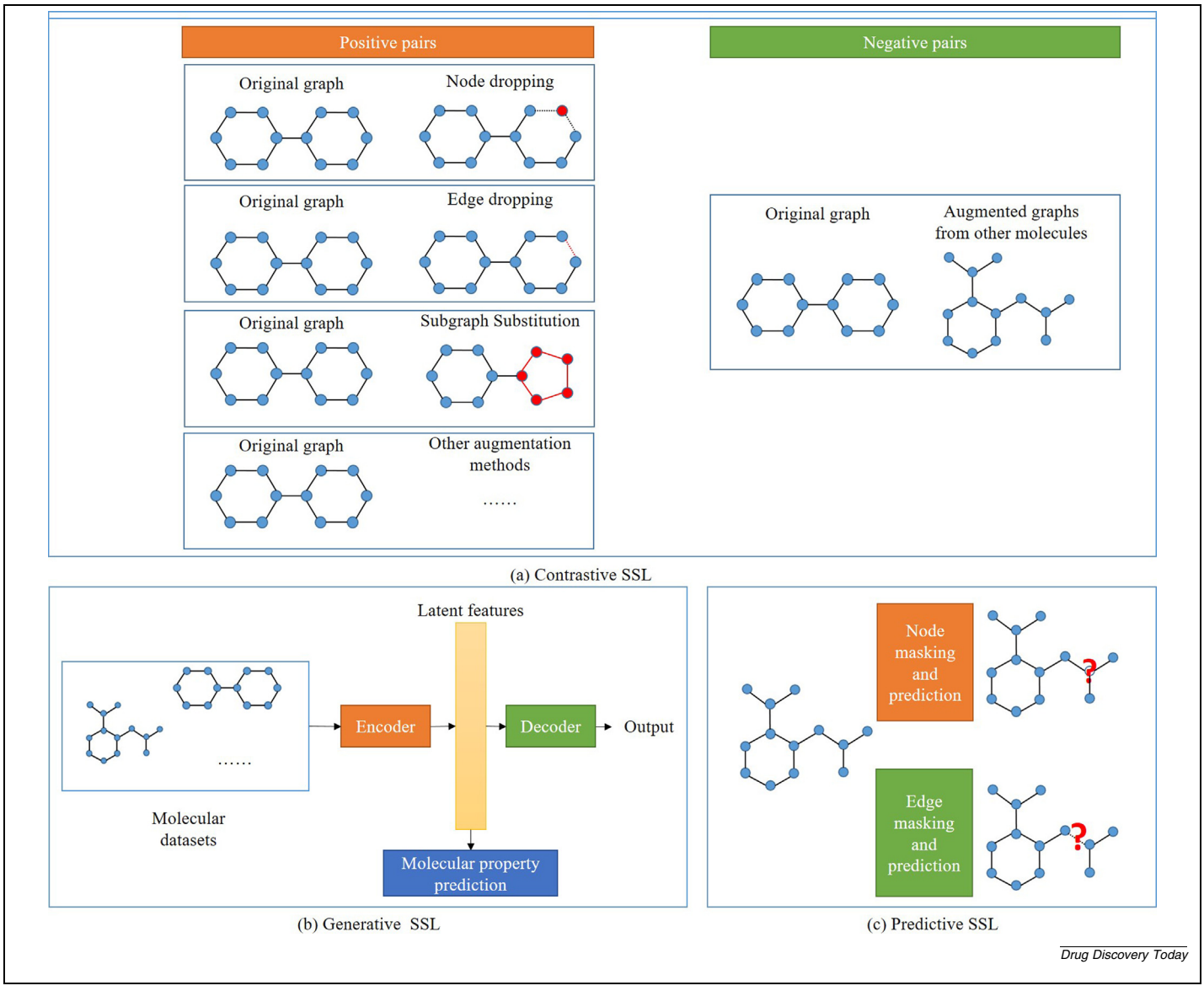


FIGURE 3

Different types of self-supervised learning (SSL) method using graphs. **(a)** Contrastive SSL: uses data augmentation methods, such as node dropping, edge dropping, and subgraph substitution for positive pairs, while selecting other augmented graphs randomly for negative pairs. **(b)** Generative SSL: the input is reconstructed through the encoder–decoder model to obtain the latent features for representation. **(c)** Predictive SSL: randomly masks some nodes or edges and lets the model predict outcomes; by doing so, the model is able to learn latent features and generate effective molecular representations.

tronic, and thermodynamic properties of the molecule. MolPSI[73] is an equal-sized molecular 2D image representation method based on a spectral graph model, in which a filtration process is used to generate a series of multiscale topological representations and geometric information.

Moreover, molecular images have been used to detect functional groups through a pretrained model, which was further fine-tuned for activity cliff prediction[74]. The learned weights were extracted to highlight the key functional groups that distinguished compounds forming activity cliff and nonactivity cliff pairs.

Thus, image-based methods are not mainstream for MPP because they have to convert data samples to the Euclidean space, which is not suitable for molecular property prediction because of the absence of atom and bond attributes. Nevertheless, image-based DL models have been proposed and developed in recent years. To take advantage of image-based DL models, image generation methods that are able to discover the relationship among atoms from certain views warrant further research, which will be helpful for the generation of new effective molecular representations.

3D Graph-based methods

The conformation of a molecule normally contains the atomic 3D coordinates of the molecule, which are also known as its geometric data and can provide additional spatial information for MPP. The first problem to be solved is that the 3D molecular conformation data set is limited. To enlarge the application fields of

geometry-based methods, conformation generation is introduced on 2D data sets. Merck Molecular Force Field (MMFF94),[75] which is embedded in the RDKit, can be used for conformation generation. Moreover, Hamiltonian neural networks[76] were proposed to predict the conformation of molecules and the predicted 3D coordinates are fed into an MPNN-based fingerprint generator for molecular representation.

However, there are still relatively few methods focusing on the geometry-based data compared with 1D- and 2D-based methods. Analogous to the 2D data, geometry-based methods are also divided into two categories: 3D graph-based methods and 3D grid-based methods.

3D graph-based methods mainly adjust or improve the GCN with the introduction of geometric information. Except for the adjacent matrix and feature matrix, 3D-GCN[77] introduces a relative position matrix comprising interatomic 3D positions to ensure translational invariance in the convolutional process. It was shown that the trained model had an intrinsic characteristic of rotational randomness without additional augmentation methods. Lu *et al.*[78] designed a GCN capturing multilevel quantum interactions from the conformation and spatial information of molecules. The different orders of neighboring nodes were involved sequentially in the model to ensure that the node representation could cover higher-order interactions. However, unlike other large graph data, such as social media and citation network graphs, the number of nodes in the molecular graph is limited. In some cases, high-order neighbors might cover a large number of atoms in a molecule and consequently lead to oversmoothing (see above).

To solve the problem of messages passing in 3D graphs, the spherical message passing method[79] was proposed. Specifically, the 3D coordinate of each atom is converted into the spherical coordinate system, and the neighboring node of the origin node is specified by a 3-tuple comprising the edge length, angle between edges, and torsion angle. Compared with the plain 3D coordinates, the 3-tuple is more flexible and able to more accurately depict the structure of molecules. Spherical message passing is invariant to the translation and rotation of input molecules. In addition, the geometric message passing neural network (GemNet)[80] also uses the spherical representations of molecules to ensure that the model is invariant to translation and equivariant to permutation and rotation.

Moreover, the SSL method is an important branch of 3D graph-based methods to discover the distinct features of a graph. Fang *et al.*[81] proposed a self-supervised framework that fully uses the molecular geometry information. They constructed a novel bond-angle graph, in which chemical bonds within a molecule were regarded as nodes rather than edges, whereas the angle formed between two bonds was considered as the edge between them. The items of 1-hop neighborhoods of certain atoms were masked for bond length and bond angle prediction to extract local representation. Liu *et al.*[82] proposed an SSL method containing contrastive learning and generative learning methods between 3D and 2D views of molecules. In contrastive learning, both 3D and 2D graphs from the same molecule are regarded as the positive pairs to train the model. In generative learning, the model is trained to generate the 3D conformers from their 2D topology. The two strategies were combined for molecular representation.

3D Grid-based methods

3D grids is another representation method using molecular geometric data, which places each atom in one or more voxels of the grid. Indeed, macromolecules, such as proteins, can be better represented by a 3D grid,[83] but such data still show good performance in MPP, especially for some quantum mechanics properties. The 3D CNN is the best choice for 3D grid data; thus, a more powerful and informative grid for 3D CNN can improve the performance of MPP.

Libmolgrid[84] provides a library to generate voxel grids of 3D molecular data to represent molecules. The resolution of the grid is an important factor affecting the results. Multi-resolution 3D-DenseNet[85] uses an atom-centered Gaussian density model to create the 3D gird, and multi-channel grids with different scales (4–14 Å) are generated as input data for 3D-DenseNet processing and prediction. Casey *et al.*[86] also generated two 3D spatial point grids (electron charge density and electrostatic potential) by using a Gaussian model. The rotation augmentation method was also applied to the above two methods to enlarge the training data set instead of varying the size, Tran *et al.*[87] selected C, H, O, N, S, and Cl as independent channels of the 3D grid, and a CNN-based autoencoder was used for molecular representation. Kuzminykh *et al.*[88] found that the regular 3D grid of molecules was too sparse and affected the performance of CNN, and instead proposed a wave transform smoothing method to fill the neighboring voxels of each atom.

Thus, there is still a lack of 3D graph-based and 3D grid-based methods for MPP, and the long analysis time required is a serious issue, especially for the 3D grid-based method. Moreover, graph convolution on 3D graph data remains an open question. Extending traditional GCN approaches to 3D scenarios by simply adding the 3D location information does not fully exploit the advantages of geometric data. Although spherical message passing has attempted to fit a specific GCN on a 3D graph, this remains a promising approach to design a 3D graph message-passing mechanism.

Hybrid data-based methods and ensemble learning

All the aforementioned 1D, 2D, and 3D representation methods present the molecule in different ways, and combining them could provide a multiview of a molecule. GraSeq[89] combines molecular graph and SMILES sequences, and uses GCN and biLSTM for encoding. Karim *et al.*[90] combined SMILES, fingerprints, molecular graphs, and 2D and 3D descriptors with multiple DL models for quantitative toxicity prediction. Normally, the fingerprints are regarded as a popular auxiliary factor, and both image-based[91] and graph-based[92] methods have combined fingerprints to improve the prediction performance.

Ensemble learning could also connect multiple classifiers to enhance the performance of the joint model over each individual one. Kosasih *et al.*[93] built an ensemble of three GINs. Busk *et al.*[94] also built an ensemble of multiple MPNNs, which were initialized with random parameters and trained individually on the same data set. Moreover, they used the variance of all classifiers to represent the predicted uncertainty, and the calibrated results improved the performance of the model. Karim *et al.*[95] converted the SMILES string into a one-hot vector indicating

the absence of each character, combing molecular image and 2D numerical features as the input, and an RNN, a CNN, and a fully connected neural network were trained on these three types of data, respectively. The outputs of these networks were combined by an ensemble averaging method or a meta-neural network.

Transfer learning, multi-task learning, and meta-learning

The deficiency of experimental data sets is another problem in MPP. Using large data sets from other domains to help to find the pattern of the target domain with fewer data would be effective in overcoming this problem. Transfer learning, multi-task learning, and meta-learning are all suggested for this purpose.

For transfer learning, the model is first trained on a large data set for certain pretext tasks, thereby learning a general representation of molecules. The learned general representation is then used for the downstream task (usually with limited samples) to transfer *a priori* knowledge. MRlogP[96] uses transfer learning on low-accuracy predicted logP values on the large data set (500 000 molecules), and the parameters are fine-tuned on a small accurate data set of 244 drug-like compounds. In the computer vision area, transfer learning is more frequently used with the help of the ImageNet data set,[97] which provides more than 1 000 000 images. Zhong *et al.*[98] used the pretrained model on the ImageNet and transferred it to the molecular image data for QSAR tasks.

Multi-task learning trains all tasks simultaneously and shares the representations to improve the generalization of the prediction. Liu *et al.*[99] used multi-task learning to predict 12 quantum chemical properties from the QM9 and Alchemy data sets. Moreover, an atom-centered symmetry function was selected as an auxiliary prediction target in the framework to improve the generalizability and transferability.

In recent years, the meta-learning method has emerged to solve the few-shot problem, also called 'learning to learn'. In the training process, the meta-learning divides the training data set into different meta tasks to learn the well-initialized model parameters with high generalization ability. The model is updated by a small number of gradient descents on a new task to enhance the performance of the model. For example, Meta-MGNN[100] combined graph neural network, SSL, and task weight-aware meta-learning in Tox21 and six tasks in SIDER for MPP. Wang *et al.*[101] proposed a property-aware embedding method considering the relationship between different molecular properties and different molecular substructures, and a meta-learning method selectively updated the parameters within tasks to model generic- and property-aware knowledge separately.

Thus, all three types of learning method can find the relationship between different tasks and solve the problem of limited data. Normally, the transfer learning method can use large-scale data sets, such as ChEMBL[102] and ZINC,[103] to learn a generic representation of molecule and fine-tune it to adapt to specific datasets. Multi-task learning improves the generalization of a model by learning several related tasks simultaneously, whereas meta-learning predicts unseen tasks with limited data,[104] which is promising in the MPP field for the few-shot problem and to avoid expensive, time-consuming and laborious experimental data collection.

Interpretability of the DL model on molecular property prediction

The most controversial area of DL is its interpretability.[105] Interpretable DL methods are divided into two classes: passive and active. Passive methods use the parameters in the DL model for explanation, whereas active methods change the training process to improve the interpretability of the model. Jiménez-Luna *et al.*[106] stated that transparency, justification, informativeness, and uncertainty estimation are the main aspects of the interpretability of AI methods in drug design.

Pope *et al.*[107] evaluated three prominent interpretability methods on GCN and reported that salient subgraphs could be explained as functional groups. Amides, tricholoromethyl, sulfonamides, and aromatic structures are all highlighted through interpretability methods. Jiménez-Luna *et al.*[108] proposed a feature attribution approach using the trained MPNN model to produce an importance score for each node, and atoms were colored according to their importance scores. The method could recognize the pharmacophore motif and identify the property cliffs.

For molecular property prediction, the passive method is still the main method used for understanding the relationship between the precise substructure of the molecule and its property. The attention mechanism could learn the weight of different parts of input to ensure that the DL model could focus on the important part. The concept of the attention mechanism could also be used for the interpretability of DL models, which can find the significant atoms or groups and their corresponding contributions to the molecular property. For example, Tang *et al.*[109] visualized self-attention values to detect which parts of molecule contributed to its lipophilicity or water solubility. Wu *et al.*[110] developed a multi-task graph attention framework for toxicity prediction. The framework extracted features of molecular substructures, with each substructure being assigned an attention weight. Highest-weighted substructures, such as acyl chloride, semicarbazone, nitrite, and nitrosamide (known as structural alerts), were detected.

In addition, the uncertainty estimation is an important way to evaluate the reliability of model. Ryu *et al.*[111] used a Bayesian GCN for the prediction of molecular properties. It replaced the standard dropout with a 'Concrete' dropout, and estimated the model uncertainty through a Bayesian approach. The authors found that uncertainty can be used as the confidence indicator of prediction. Hirschfel *et al.* [112] evaluated many uncertainty estimators of molecular property predictions and proposed that joining multiple weak uncertainty estimators could lead to more consistent performance.

Molecular property prediction challenges and future work

Self-supervised learning methods in 3D data

The SSL method is a promising direction to discover the distinct features of molecules. Both 1D data- and 2D data-based methods have been proposed in recent years, and have achieved relatively good performance. For SSL methods, the design of the pretext

task is the most crucial step. However, the abundant information hidden in 3D molecular data remains to be fully exploited. 3D SSL methods have already been applied in many fields; thus, designing a novel molecular 3D SSL method would be very useful for predicting molecular properties. Moreover, virtual screening is used to find the ligand–target pair with high binding affinity, which is a key step in drug discovery and design.[113,114] In virtual screening software, the conformations of both the target and ligand are required. Therefore, an accurate and comprehensive 3D representation of the molecule will also be helpful for virtual screening.

Graph convolution methods with experience

The GCN method has become the mainstream method in molecule-related tasks because of its excellent performance on graph data. However, there is still room for improvement, such as introducing empirical data and expert knowledge. We cannot ignore the effect of human experience on the DL model, and there are some ways of injecting expertise into the model by defining the types of atoms, bonds, and functional groups.[61,115] However, some high-level or complex domain experiences can be more powerful in the MPP, such as the relationship between the motif and properties, which have not yet been fully exploited. Designing this type of high-level experience as the input of the DL model remains an open question.

1D, 2D, and 3D data fusion and selection methods

Generally, high-dimensional data contain more information compared with low-dimensional data. If this is the case, then the performance of models based on 2D data should be better than that based on 1D data. However, as we discussed above, the hybrid method uses 1D, 2D, and 3D data, and ablation experiments have demonstrated their individual roles in MPP, indicating that the information contributed by low-dimension data was not completely covered by high-dimension data. A question raised here is why 1D data still contribute to the model performance when high-dimension data (2D or 3D data) are already being used. This can be answered from the following two perspectives. First, some information mgiht be lost when converting 1D sequences to 2D graphs. Second, the DL model cannot fully exploit the hidden information in high-dimension data; therefore, the 1D data still serve as an auxiliary information source. Whatever the reason, how to determine an appropriate type (or the optimal combination of multiple types) of data remains an open question.

Meta-learning methods

Transfer learning, multi-task learning, and meta-learning are all used to solve the lack of experimental data available for certain properties. We have already reviewed some of the appropriate methods here and, in our opinion, we believe that the meta-learning method is one of the most promising research directions at present. To be more specific, meta-learning is an ideal approach for practical applications, because there might be only a few instances for certain tasks (e.g., predicting some rare molecular properties), when traditional ML or DL models cannot be used because of the limited amount of data samples. Thus, the meta-learning method for MPP warrants further research.

The interpretability of DL models

We have discussed a few methods regarding the interpretability of DL. Unlike traditional tasks in image processing, most molecule-related tasks are highly specialized and need chemical experts to analyze the underlying mechanisms, such as the role of molecular substructures. Such characteristics of molecule-related tasks are somewhat in contradiction to the 'black-box' nature of DL models. As a result, improving the interpretability of DL models is always necessary. More concretely, locating the key functional elements within the model by analyzing both the successfully predicted and the failed data samples will benefit not only the final performance of DL models, but also the discovery of novel QSAR theory. In our opinion, the active method is a more powerful tool to increase the interpretability of DL models by adding specific parameters to them.

Concluding remarks

In this review, we have surveyed DL methods on multiple types of molecular data and the emerging methods including transfer learning, meta learning, and so on. In addition, we have also discussed the interpretability methods of molecules in DL models. Significant progress in drug discovery has been made using the DL method. However, it still faces more challenges to improve the performance, robustness and interpretability of molecular representation and property prediction.

Data availability

No data was used for the research described in the article.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research was funded by the Shandong Key Science and Technology Innovation Project (2021CXGC011003), Shandong Provincial Postdoctoral Program for Innovative Talents (SDBX2020003), Natural Science Foundation of China (62202498), Shandong Provincial Natural Science Foundation (ZR2022QF111, ZR2021QF023), and Fundamental Research Funds for the Central Universities (21CX06018A).

References

1 Z. Wu, B. Ramsundar, E.N. Feinberg, J. Gomes, C. Geniesse, A.S. Pappu, et al., MoleculeNet: a benchmark for molecular machine learning, Chem Sci 9 (2018) 513–530.

2 Z. Yang, W. Zhong, L. Zhao, C.C. Yu-Chian, MGraphDTA: deep multiscale graph neural network for explainable drug-target binding affinity prediction, Chem Sci 13 (2022) 816–833.

3 W. Yuan, G. Chen, C.Y.C. Chen, FusionDTA: attention-based feature polymerizer and knowledge distillation for drug–target binding affinity prediction, Brief Bioinform 23 (2022) bbab506.
4 F. Wang, X. Feng, X. Guo, L. Xu, L. Xie, S. Chang, Improving de novo molecule generation by embedding LSTM and attention mechanism in CycleGAN, Front Genet 12 (2021) 709500.
5 S. Wang, T. Song, S. Zhang, M. Jiang, Z. Wei, Z. Li, Molecular substructure tree generative model for de novo drug design, Brief Bioinform 23 (2022) bbab592.
6 F. Wang, X. Diao, S. Chang, L. Xu, Recent progress of deep learning in drug discovery, Curr Pharm Des 27 (2021) 2088–2096.
7 D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, J Chem Inf Comput Sci 28 (1988) 31–36.
8 M. Krenn, F. Häse, A. Nigam, P. Friederich, A. Aspuru-Guzik, Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation, Mach Learn Sci Technol 1 (2020) 45024.
9 D. Rogers, M. Hahn, Extended-connectivity fingerprints, J Chem Inf Model 50 (2010) 742–754.
10 J.L. Durant, B.A. Leland, D.R. Henry, J.G. Nourse, Reoptimization of MDL keys for use in drug discovery, J Chem Inf Comput Sci 42 (2002) 1273–1280.
11 Y. Ding, M. Chen, C. Guo, P. Zhang, J. Wang, Molecular fingerprint-based machine learning assisted QSAR model development for prediction of ionic liquid properties, J Mol Liq 326 (2021) 115212.
12 L. Xie, L. Xu, R. Kong, S. Chang, X. Xu, Improvement of prediction performance with conjoint molecular fingerprint in deep learning, Front Pharmacol 11 (2020) 606668.
13 M. Wang, Z. Cang, G.W. Wei, A topology-based network tree for the prediction of protein–protein binding affinity changes following mutation, Nat Mach Intell 2 (2020) 116–123.
14 D.D. Nguyen, G.W. Wei, AGL-score: algebraic graph learning score for protein–ligand binding scoring, ranking, docking, and screening, J Chem Inf Model 59 (2019) 3291–3304.
15 Z. Cang, L. Mu, G.W. Wei, Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening, PLoS Comput Biol 14 (2018) e1005929.
16 D.D. Nguyen, Z. Cang, K. Wu, M. Wang, Y. Cao, G.W. Wei, Mathematical deep learning for pose and binding affinity prediction and ranking in D3R Grand Challenges, J Comput Aided Mol Des 33 (2019) 71–82.
17 Z. Cang, G.W. Wei, Integration of element specific persistent homology and machine learning for protein–ligand binding affinity prediction, Int J Numer Method Biomed Eng 34 (2018) e2914.
18 Z. Cang, G.W. Wei, TopologyNet: Topology based deep convolutional and multi–task neural networks for biomolecular property predictions, PLoS Comput Biol 13 (2017) e1005690.
19 Z. Meng, K. Xia, Persistent spectral–based machine learning (PerSpect ML) for protein–ligand binding affinity prediction, Sci Adv 7 (2021) eabc5329.
20 J. Wee, K. Xia, Forman persistent Ricci curvature (FPRC)-based machine learning models for protein–ligand binding affinity prediction, Brief Bioinform 22 (2021) bbab136.
21 X. Liu, H. Feng, J. Wu, K. Xia, Dowker complex based machine learning (DCML) models for protein–ligand binding affinity prediction, PLoS Comput Biol 18 (2022) e1009943.
22 D.D. Nguyen, K. Gao, J. Chen, R. Wang, G.W. Wei, Unveiling the molecular mechanism of SARS-CoV-2 main protease inhibition from 137 crystal structures using algebraic topology and deep learning, Chem Sci 11 (2020) 12036–12046.
23 D.D. Nguyen, K. Gao, M. Wang, G.W. Wei, MathDL: mathematical deep learning for D3R Grand Challenge 4, J Comput Aided Mol Des 34 (2020) 131–147.
24 X. Liu, H. Feng, J. Wu, K. Xia, Persistent spectral hypergraph based machine learning (PSH-ML) for protein–ligand binding affinity prediction, Brief Bioinform 22 (2021) bbab127.
25 M. Sun, S. Zhao, C. Gilvary, O. Elemento, J. Zhou, F. Wang, Graph convolutional networks for computational drug development and discovery, Brief Bioinform 21 (2020) 919–935.
26 J. Xiong, Z. Xiong, K. Chen, H. Jiang, M. Zheng, Graph neural networks for automated de novo drug design, Drug Discov Today 26 (2021) 1382–1393.
27 J.H. Chen, Y.J. Tseng, Different molecular enumeration influences in deep learning: an example using aqueous solubility, Brief Bioinform 22 (2021) bbaa092.
28 T.B. Kimber, M. Gagnebin, A. Volkamer, Maxsmi: maximizing molecular property prediction performance with confidence estimation using SMILES augmentation and deep learning, Artif Intell Life Sci 1 (2021) 100014.
29 Lim S, Lee YO. Predicting chemical properties using self-attention multi-task learning based on SMILES representation. In: *25th International Conference on Pattern Recognition (ICPR)*. Pisctaway; IEEE; 2021: 3146–53.
30 M. Hirohara, Y. Saito, Y. Koda, K. Sato, Y. Sakakibara, Convolutional neural network based on SMILES representation of compounds for detecting chemical motif, BMC Bioinformatics 19 (2018) 526.
31 Y. Hou, S. Wang, B. Bai, H.C.S. Chan, S. Yuan, Accurate physical property predictions via deep learning, Molecules 27 (2022) 1668.
32 A.L. Nazarova, L. Yang, K. Liu, A. Mishra, R.K. Kalia, K. Nomura, et al., Dielectric polymer property prediction using recurrent neural networks with optimizations, J Chem Inf Model 61 (2021) 2175–2186.
33 C. Li, J. Feng, S. Liu, J. Yao, A novel molecular representation learning for molecular property prediction with a multiple SMILES-based augmentation, Comput Intell Neurosci 2022 (2022) 8464452.
34 X. Li, D. Fourches, SMILES pair encoding: a data-driven substructure tokenization algorithm for deep learning, J Chem Inf Model 61 (2021) 1560–1569.
35 Q. Lv, G. Chen, L. Zhao, W. Zhong, C.C. Yu-Chian, Mol2Context-vec: learning molecular representation from context awareness for drug discovery, Brief Bioinform 22 (2021) bbab317.
36 J. Li, X. Jiang, Mol-BERT: an effective molecular representation with BERT for molecular property prediction, Wirel Commun Mob Comput 2021 (2021) 7181815.
37 J. Shao, Q. Gong, Z. Yin, W. Pan, S. Pandiyan, L. Wang, S2DV: converting SMILES to a drug vector for predicting the activity of anti-HBV small molecules, Brief Bioinform 23 (2022) bbab593.
38 Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: *25th International Conference on Machine Learning*. Pisctaway; IEEE; 2021: 1597–607.
39 Liu X, Zhang F, Hou Z, Mian L, Wang Z, Zhang J, *et al*. Self-supervised learning: generative or contrastive. *IEEE Trans Knowl Data Eng*. Published online June 21, 2021. http://dx.doi.org/10.1109/TKDE.2021.3090866.
40 Wu L, Lin H, Tan C, Gao Z, Li SZ. Self-supervised learning on graphs: contrastive, generative, or predictive. *IEEE Trans Knowl Data Eng*. Published online December 1, 2021. http://dx.doi.org/10.1109/TKDE.2021.3131584.
41 Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv*. 2018; 2018: arXiv181004805. 2018.
42 Fabian B, Edlich T, Gaspar H, Segler M, Meyers J, Fiscato M, *et al*. Molecular representation learning with language models and domain-relevant auxiliary tasks. *arXiv*. 2020; 2020: arXiv201113230.
43 S. Wang, Y. Guo, Y. Wang, H. Sun, J. Huang, in: SMILES-bert: large scale unsupervised pre-training for molecular property prediction, Association for Computing Machinery, New York, 2019, pp. 429–436.
44 S. Hu, P. Chen, P. Gu, B. Wang, A deep learning-based chemical system for QSAR prediction, IEEE J Biomed Heal Informatics 24 (2020) 3020–3028.
45 X. Li, D. Fourches, Inductive transfer learning for molecular activity prediction: next-gen QSAR models with MolPMoFiT, J Cheminform 12 (2020) 1–15.
46 R. Winter, F. Montanari, F. Noé, D.A. Clevert, Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations, Chem Sci 10 (2019) 1692–1701.
47 P. Karpov, G. Godin, I.V. Tetko, Transformer-CNN: Swiss knife for QSAR modeling and interpretation, J Cheminform 12 (2020) 1–12.
48 R. Liao, Z. Zhao, R. Urtasun, R.S. Zemel, Lanczosnet: multi-scale deep graph convolutional networks, arXiv 2019 (2019).
49 C. Shang, Q. Liu, Q. Tong, J. Sun, M. Song, J. Bi, Multi-view spectral graph convolution with consistent edge attention for molecular modeling, Neurocomputing 445 (2021) 12–25.
50 J. Wang, D. Cao, C. Tang, L. Xu, Q. He, B. Yang, et al., DeepAtomicCharge: a new graph convolutional network–based architecture for accurate prediction of atomic charges, Brief Bioinform 22 (2021) bbaa183.
51 Z. Xiong, D. Wang, X. Liu, F. Zhong, X. Wan, X. Li, et al., Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism, J Med Chem 63 (2019) 8749–8760.
52 X.S. Li, X. Liu, L. Lu, X.S. Hua, Y. Chi, K. Xia, Multiphysical graph neural network (MP-GNN) for COVID-19 drug design, Brief Bioinform 23 (2022) bbac231.
53 H. Ma, Y. Bian, Y. Rong, W. Huang, T. Xu, W. Xie, et al., Cross-dependent graph neural networks for molecular property prediction, Bioinformatics 38 (2022) 2003–2009.
54 P. Li, Y. Li, C.Y. Hsieh, S. Zhang, X. Liu, H. Liu, et al., TrimNet: learning molecular representation from triplet messages for biomedicine, Brief Bioinform 22 (2021) bbaa266.

55 M. Withnall, E. Lindelöf, O. Engkvist, H. Chen, Building attention and edge message passing neural networks for bioactivity and physical–chemical property prediction, J Cheminform 12 (2020) 1–18.
56 Y. Su, Z. Wang, S. Jin, W. Shen, J. Ren, M.R. Eden, An architecture of deep learning in QSPR modeling for the prediction of critical properties using molecular signatures, AIChE J 65 (2019) e16678.
57 Z. Wang, Y. Su, W. Shen, S. Jin, J.H. Clark, J. Ren, et al., Predictive deep learning models for environmental properties: the direct calculation of octanol–water partition coefficients from molecular graphs, Green Chem 21 (2019) 4555–4565.
58 W. Jin, R. Barzilay, T. Jaakkola, Junction tree variational autoencoder for molecular graph generation, Proc Machine Learn Res 80 (2018) 2323–2332.
59 S. Wang, Z. Li, S. Zhang, M. Jiang, X. Wang, Z. Wei, Molecular property prediction based on a multichannel substructure graph, IEEE Access 8 (2020) 18601–18614.
60 Y. Wang, J. Wang, Z. Cao, A.B. Farimani, MolCLR: molecular contrastive learning of representations via graph neural networks, arXiv 2021 (2021).
61 Sun M, Xing J, Wang H, Chen B, Zhou J. MoCL: data-driven molecular fingerprint via knowledge-aware contrastive learning from molecular graph. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. New York; Association for Computing Machinery: 2021: 3585–94.
62 P. Li, J. Wang, Y. Qiao, H. Chen, Y. Yu, X. Yao, et al., An effective self-supervised framework for learning expressive molecular global representations to drug discovery, Brief Bioinform 22 (2021) bbab109.
63 X.C. Zhang, C.K. Wu, Z.J. Yang, Z.X. Wu, J.C. Yi, C.Y. Hsieh, et al., MG-BERT: leveraging unsupervised atomic representation learning for molecular property prediction, Brief Bioinform 22 (2021) bbab152.
64 D. Koge, N. Ono, M. Huang, M. Altaf-Ul-Amin, S. Kanaya, Embedding of molecular structure using molecular hypergraph variational autoencoder with metric learning, Mol Inform 40 (2021) 2000203.
65 H. Kajino, Molecular hypergraph grammar with its application to molecular optimization, Proc Machine Learn Res 97 (2019) 3183–3191.
66 Y. Rong, Y. Bian, T. Xu, W. Xie, Y. Wei, W. Huang, et al., Self-supervised graph transformer on large-scale molecular data, Adv Neural Inf Process Syst 33 (2020) 12559–12571.
67 Q. Sun, J. Li, H. Peng, J. Wu, Y. Ning, P.S. Yu, et al., SUGAR: Subgraph neural network with reinforcement pooling and self-supervised mutual information mechanism, Association for Computing Machinery, New York, 2021: 2081–91..
68 Landrum G. *RDKit: Open-Source Cheminformatics Software*. http: //www.rdkit.org [Accessed September 20, 2022].
69 N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G.R. Hutchison, Open Babel: an open chemical toolbox, J Cheminform 3 (2011) 1–14.
70 A. Yoshimori, Prediction of molecular properties using molecular topographic map, Molecules 26 (2021) 4475.
71 W.X. Shen, X. Zeng, F. Zhu, C. Qin, Y. Tan, Y.Y. Jiang, et al., Out-of-the-box deep learning prediction of pharmaceutical properties by broadly learned knowledge-based molecular representations, Nat Mach Intell 3 (2021) 334–343.
72 A.B. Tchagang, J.J. Valdés, Time frequency representations and deep convolutional neural networks: a recipe for molecular properties prediction, in: In: *2021 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*. Pisacataway, 2021, pp. 1–5.
73 P. Jiang, Y. Chi, X.S. Li, X. Liu, X.S. Hua, K. Xia, Molecular persistent spectral image (Mol–PSI) representation for machine learning models in drug design, Brief Bioinform 23 (2022) bbab527.
74 J. Iqbal, M. Vogt, J. Bajorath, Learning functional group chemistry from molecular images leads to accurate prediction of activity cliffs, Artif Intell Life Sci 1 (2021) 100022.
75 T.A. Halgren, Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94, J Comput Chem 17 (1996) 490–519.
76 Z. Li, S. Yang, G. Song, C.L. HamNet, onformation-guided molecular representation with Hamiltonian neural networks, arXiv 2021 (2021).
77 H. Cho, I.S. Choi, Enhanced deep-learning prediction of molecular properties via augmentation of bond topology, ChemMedChem 14 (2019) 1604–1609.
78 Lu C, Liu Q, Wang C, Huang Z, Lin P, He L. Molecular property prediction: a multilevel quantum interactions modeling perspective. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Menlo Park; AAAI; 2019: 1052–60.
79 Y. Liu, L. Wang, M. Liu, X. Zhang, B. Oztekin, S. Ji, Spherical message passing for 3D graph networks, arXiv 2021 (2021).
80 J. Klicpera, F. Becker, S. Günnemann, GemNet: Universal directional graph neural networks for molecules, Adv Neural Inf Process Syst 34 (2021) 6790–6802.
81 X. Fang, L. Liu, J. Lei, D. He, S. Zhang, J. Zhou, et al., Geometry-enhanced molecular representation learning for property prediction, Nat Mach Intell 4 (2022) 127–134.
82 S. Liu, H. Wang, W. Liu, J. Lasenby, H. Guo, J. Tang, Pre-training molecular graph representation with 3D geometry, arXiv 2021 (2021).
83 L. Xie, L. Xu, S. Chang, X. Xu, L. Meng, Multitask deep networks with grid featurization achieve improved scoring performance for protein–ligand binding, Chem Biol Drug Des 96 (2020) 973–983.
84 J. Sunseri, D.R. Koes, Libmolgrid: graphics processing unit accelerated molecular gridding for deep learning applications, J Chem Inf Model 60 (2020) 1079–1084.
85 S. Liu, J. Li, K.C. Bennett, B. Ganoe, T. Stauch, M. Head-Gordon, et al., Multiresolution 3D-DenseNet for chemical shift prediction in NMR crystallography, J Phys Chem Lett 10 (2019) 4558–4565.
86 A.D. Casey, S.F. Son, I. Bilionis, B.C. Barnes, Prediction of energetic material properties from electronic structure using 3D convolutional neural networks, J Chem Inf Model 60 (2020) 4457–4473.
87 N. Tran, D. Kepple, S. Shuvaev, A. Koulakov, DeepNose: using artificial neural networks to represent the space of odorants, Proc Machine Learn Res 97 (2019) 6305–6314.
88 D. Kuzminykh, D. Polykovskiy, A. Kadurin, A. Zhebrak, I. Baskov, S. Nikolenko, et al., 3D molecular representations based on the wave transform for convolutional neural networks, Mol Pharm 15 (2018) 4378–4385.
89 Z. Guo, W. Yu, C. Zhang, M. Jiang, N.V. Chawla, in: GraSeq: graph and sequence fusion learning for molecular property prediction, Association for Computing Machinery, New York, 2020, pp. 435–443.
90 A. Karim, V. Riahi, A. Mishra, M.A.H. Newton, A. Dehzangi, T. Balle, et al., Quantitative toxicity prediction via meta ensembling of multitask deep learning models, ACS Omega 6 (2021) 12306–12317.
91 J.G. Meyer, S. Liu, I.J. Miller, J.J. Coon, A. Gitter, Learning drug functions from chemical structures with convolutional neural networks and random forests, J Chem Inf Model 59 (2019) 4438–4449.
92 J.Y. Ryu, M.Y. Lee, J.H. Lee, B.H. Lee, K.S. Oh, DeepHIT: a deep learning framework for prediction of hERG-induced cardiotoxicity, Bioinformatics 36 (2020) 3049–3055.
93 E.E. Kosasih, J. Cabezas, X. Sumba, P. Bielak, K. Tagowski, K. Idanwekhai, et al., On graph neural network ensembles for large-scale molecular property prediction, arXiv 2021 (2021).
94 J. Busk, P.B. Jørgensen, A. Bhowmik, M.N. Schmidt, O. Winther, T. Vegge, Calibrated uncertainty for molecular property prediction using ensembles of message passing neural networks, Mach Learn Sci Technol 3 (2021) 15012.
95 A. Karim, J. Singh, A. Mishra, A. Dehzangi, M.A.H. Newton, A. Sattar, Toxicity prediction by multimodal deep learning, Lect Notes Comp Sci 11669 (2019) 142–152.
96 Y.K. Chen, S. Shave, M. Auer, MRlogP: transfer learning enables accurate logP prediction using small experimental training datasets, Processes 9 (2021) 2029.
97 Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway; IEEE; 2009: 248–55.
98 S. Zhong, J. Hu, X. Yu, H. Zhang, 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 408 (2021) 127998.
99 Z. Liu, L. Lin, Q. Jia, Z. Cheng, Y. Jiang, Y. Guo, et al., Transferable multilevel attention neural network for accurate prediction of quantum chemistry properties via multitask learning, J Chem Inf Model 61 (2021) 1066–1082.
100 Z. Guo, C. Zhang, W. Yu, J. Herr, O. Wiest, M. Jiang, et al., Few-shot graph learning for molecular property prediction, arXiv 2021 (2021).
101 Y. Wang, A. Abuduweili, Q. Yao, D. Dou, Property-aware relation networks for few-shot molecular property prediction, arXiv 2021 (2021).
102 D. Mendez, A. Gaulton, A.P. Bento, J. Chambers, M. De Veij, E. Félix, et al., ChEMBL: towards direct deposition of bioassay data, Nucleic Acids Res 47 (2019) D930–D940.
103 J.J. Irwin, K.G. Tang, J. Young, C. Dandarchuluun, B.R. Wong, M. Khurelbaatar, et al., ZINC20—a free ultralarge-scale chemical database for ligand discovery, J Chem Inf Model 60 (2020) 6065–6073.
104 H. Wang, H. Zhao, B. Li, Bridging multi-task learning and meta-learning: towards efficient training and effective adaptation, Proc Machine Learn Res 139 (2021) 10991–11002.
105 Y. Zhang, P. Tiňo, A. Leonardis, K. Tang, A survey on neural network interpretability, IEEE Trans Emerg Top Comput Intell 5 (2021) 726–742.
106 J. Jiménez-Luna, F. Grisoni, G. Schneider, Drug discovery with explainable artificial intelligence, Nat Mach Intell 2 (2020) 573–584.

107 Pope PE, Kolouri S, Rostami M, Martin CE, Hoffmann H. Explainability methods for graph convolutional neural networks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Piscatawy; IEEE: 2019: 10772–81.

108 J. Jiménez-Luna, M. Skalic, N. Weskamp, G. Schneider, Coloring molecules with explainable artificial intelligence for preclinical relevance assessment, J Chem Inf Model 61 (2021) 1083–1094.

109 B. Tang, S.T. Kramer, M. Fang, Y. Qiu, Z. Wu, D. Xu, A self-attention based message passing neural network for predicting molecular lipophilicity and aqueous solubility, J Cheminform 12 (2020) 1–9.

110 Z. Wu, D. Jiang, J. Wang, C.Y. Hsieh, D. Cao, T. Hou, Mining toxicity information from large amounts of toxicity data, J Med Chem 64 (2021) 6924–6936.

111 S. Ryu, Y. Kwon, W.Y. Kim, A Bayesian graph convolutional network for reliable prediction of molecular properties with uncertainty quantification, Chem Sci 10 (2019) 8438–8446.

112 L. Hirschfeld, K. Swanson, K. Yang, R. Barzilay, C.W. Coley, Uncertainty quantification using neural networks for molecular property prediction, J Chem Inf Model 60 (2020) 3770–3780.

113 S. Wang, M. Jiang, S. Zhang, X. Wang, Q. Yuan, Z. Wei, et al., MCN–CPI: multiscale convolutional network for compound–protein interaction prediction, Biomolecules 11 (2021) 1119.

114 S. Zhang, M. Jiang, S. Wang, X. Wang, Z. Wei, Z. Li, SAG-DTA: Prediction of drug–target affinity using self-attention graph network, Int J Mol Sci 22 (2021) 8993.

115 T. Hasebe, Knowledge-embedded message-passing neural networks: improving molecular property prediction with human knowledge, ACS Omega 6 (2021) 27955–27967.

116 O. Trott, A.J. Olson, AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading, J Comput Chem 31 (2010) 455–461.