SCIENCE CHINA
Information Sciences

• REVIEW •

July 2025, Vol. 68, Iss. 7, 170103:1–170103:16
https://doi.org/10.1007/s11432-024-4453-4

Special Topic: AI for Biology

Self-supervised learning in drug discovery

Yangyang CHEN[1], Zixu WANG[1], Jianmin WANG[2], Yanyi CHU[3], Qingpeng ZHANG[4], Zhong Alan LI[5] & Xiangxiang ZENG[6*]

[1]*Department of Computer Science, University of Tsukuba, Tsukuba 305-8577, Japan*
[2]*Department of Integrative Biotechnology, Yonsei University, Incheon 21983, Republic of Korea*
[3]*Department of Pathology, School of Medicine, Stanford University, Stanford CA 94305, USA*
[4]*Institute of Data Science & Department of Pharmacology and Pharmacy, The University of Hong Kong, Hong Kong 999077, China*
[5]*Department of Biomedical Engineering, Faculty of Engineering, The Chinese University of Hong Kong, Hong Kong 999077, China*
[6]*College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China*

Received 26 December 2024/Revised 12 March 2025/Accepted 29 March 2025/Published online 23 June 2025

Abstract Recent advances in deep learning have proven highly effective in medical applications, notably in drug discovery. Among various deep learning techniques, self-supervised learning (SSL) has shown considerable advantages over traditional supervised learning by utilizing vast amounts of unlabeled data for model training. This review discusses both classic and state-of-the-art SSL-based methods in the drug discovery field, detailing their applications from small molecule and peptide drug discovery to antibody design and vaccine development, which provides a current and accessible guide to drug discovery. Furthermore, this review suggests the challenges faced by SSL in drug discovery, such as data quality, model interpretability, and computational resource constraints, and outlines its potential future directions. As deep learning technology advances, we anticipate that SSL-based models will increasingly promote drug research and development, potentially revolutionizing the pharmaceutical industry.

Keywords drug discovery, self-supervised learning, drug representation

Citation Chen Y Y, Wang Z X, Wang J M, et al. Self-supervised learning in drug discovery. Sci China Inf Sci, 2025, 68(7): 170103, https://doi.org/10.1007/s11432-024-4453-4

1 Introduction

The primary goal of drug research and development [1] is to pinpoint specific target molecules with optimal properties within the extensive and continually growing chemical space. However, rapid advancements in chemistry have rendered traditional experimental screening methods for these targets increasingly impractical and inefficient. The advent of deep learning technologies [2–7] has markedly sped up the target identification process, yet these models depend heavily on extensive known datasets for effective training. Labeled data [8,9] acquisition often consumes a significant amount of time and resources during drug research and development, especially for emerging biological targets and complex disease states, where the validity and accessibility of such data are often very limited. In traditional machine learning applications, this scarcity of data often leads to inefficient model training. This dependence creates a significant bottleneck, especially when investigating novel compounds with scant existing data. Despite the difficulty of obtaining high-quality labeled data, a large amount of relevant unlabeled data exists. This data contains a wealth of information such as molecular structures, gene expression levels, protein interactions, and functional annotations. These insights can help in understanding biological processes, identifying potential drug targets, and advancing research. Training deep learning models directly using this unlabeled data is a potential solution [9].

Self-supervised learning (SSL) [10–13] is an advanced machine learning framework that provides a way for deep learning models to effectively learn data representations using unlabeled data. In the context of SSL, the model uses the information in the dataset itself as the learning objective (e.g., predicting

* Corresponding author (email: xzeng@xmu.edu.cn)

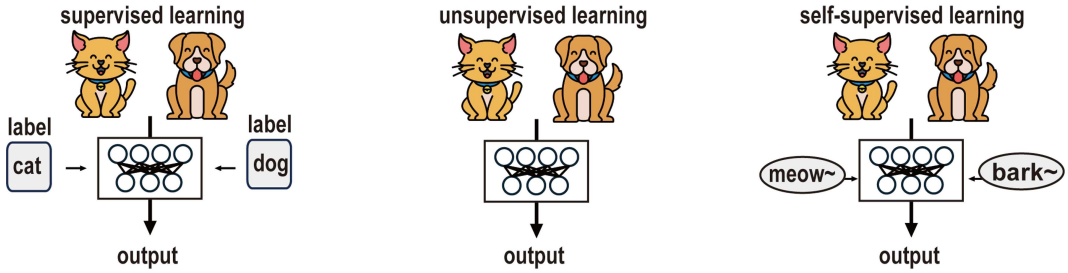

Figure 1 (Color online) Comparative illustration of three machine learning methodologies: supervised learning, unsupervised learning, and self-supervised learning. Supervised learning uses explicitly provided labels to guide the learning process. Unsupervised learning operates without labels, identifying data patterns and structures autonomously. Self-supervised learning generates its own labels from inherent features of the data, such as audio cues, to train the model.

missing fragments in a molecular sequence), thereby learning a comprehensive representation of the entire dataset. The introduction of SSL into the field of drug discovery, i.e., through pre-training on large-scale unlabeled chemical databases, not only reduces the dependence on expensive experiments, but also enables the rapid identification and optimization of potential candidate molecules in the early stages of drug research and development. Specifically, powerful feature representation enables the model to capture the complex relationship between molecular structure and its properties, and achieve efficient similarity comparison and pattern recognition in high-dimensional feature spaces. This ability enables researchers to quickly screen candidate molecules with potential biological activity and excellent pharmacokinetic properties. In addition, powerful feature representation also supports the fine optimization of candidate molecules, further improving their drug properties and safety by adjusting and improving their structural characteristics. This not only speeds up the screening and optimization process of candidate molecules, but also improves the success rate of discovering highly effective drug candidates, thereby significantly shortening the drug development cycle and reducing development costs.

In this review, we begin by defining SSL and outlining its principles, and then comparing it with supervised and unsupervised learning. Subsequently, we explore in detail the utilization of SSL in drug discovery and its specific applications. Lastly, we review the current status and limitations of SSL in drug research and development and discuss the prospective advancements of SSL in this field.

2 Overview of self-supervised learning

The SSL algorithm [10–17] is designed to extract discriminative features from numerous unlabeled instances without depending on manual annotations. Figure 1 concisely illustrates the differences among the three principal machine learning paradigms: supervised learning [18–21], unsupervised learning [22] and self-supervised learning. Supervised learning utilizes labeled training data to train models, where each sample is paired with a corresponding output label. However, it requires a substantial volume of high-quality labeled data, which can be expensive and time-consuming to acquire, and may lead to model overfitting [23, 24] and reduced generalization capacity [25]. Unsupervised learning seeks to identify inherent structures and associations in data without using labels, but it lacks explicit feedback or indicators, complicating the verification of model outputs. SSL functions as a variant of unsupervised learning by generating pseudo-labels [26] from the data itself, mimicking the supervised learning process. This approach not only capitalizes on unlabeled data but also enhances supervised learning tasks, for instance, by serving as a pre-training step [27–29] to boost the performance of supervised models.

The method of SSL mainly covers three types of tasks: prediction tasks [8, 30–32], generation tasks [10, 33–35], and contrastive learning [10, 12, 36–40]. In predictive tasks, the model must infer missing information from existing data features to internally establish the inherent structure and logical relationships of the data. For instance, in text processing, the model may need to predict missing words or the subsequent part of a sentence [27]. Generation tasks enable the model to learn the data distribution and generate new data samples that align with the statistical characteristics of the original data [39]. This not only enhances the model's understanding of data structures but also supports data augmentation in scenarios where data is limited. Through this approach, SSL allows models to self-train without external labels by analyzing and mimicking the input characteristics of data. While diffusion models [41] also generate high-quality outputs without labeled data, they primarily focus on iterative refinement for sample

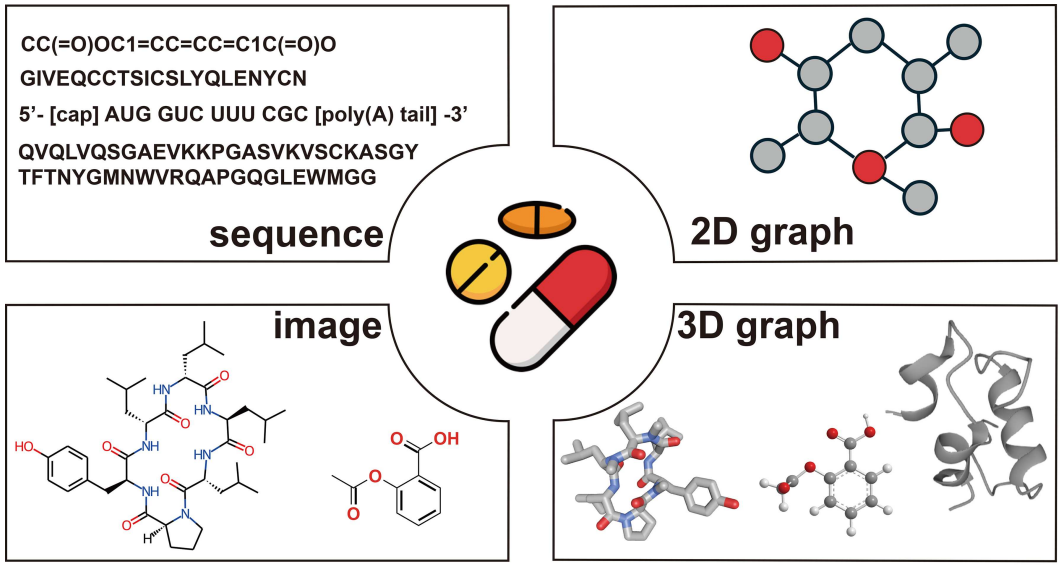

Figure 2 (Color online) Four representations of a drug molecule. The "sequence" displays the underlying code of the molecule in a linear format, such as SMILES and peptide sequence. The "2D graph" focuses on representing the molecule as a network of nodes (atoms) and edges (bonds), providing a simplified visual abstraction that emphasizes connectivity and relationships between elements. The "image" represents the molecule in a traditional 2D chemical structure diagram, showing connections and arrangements of atoms. The "3D graph" visualizes the molecule in three-dimensional space, highlighting its spatial configuration essential for understanding molecular interactions. Each method of representation is essential for facilitating various types of molecular and pharmacological studies.

generation. In contrast, SSL enhances not only generation quality but also downstream predictive tasks by capturing rich structural and semantic information. In contrastive learning tasks, models learn by comparing different data instances, aiming to cluster similar data points closer and separate dissimilar ones further in the feature space. Collectively, these methods enhance the scalability and robustness of model training by utilizing large volumes of unlabeled data.

3 How can self-supervised learning be applied to drug discovery?

Designing a novel drug entails navigating a complex landscape that necessitates meeting multiple benchmarks, including on-target potency, desirable physicochemical properties, and adherence to pharmacological safety standards [42,43]. Traditional drug discovery relies heavily on chemists manually screening extensive chemical libraries to identify candidate molecules and conduct experimental validation—a process that is both time-consuming and labor-intensive. Although deep learning has hastened this process, most models still depend on labeled data (supervised learning). In contrast, SSL leverages vast amounts of unlabeled molecular data to train models, effectively teaching the model to recognize and generate candidate molecules that satisfy predefined criteria without the need for manual labeling. Self-supervised learning bridges the gap between the abundance of large-scale biological datasets and the precision required for targeted drug design, offering a cost-effective strategy to accelerate the discovery and optimization of new therapeutics.

3.1 Representations of drug molecules

The representation of molecules [44] is fundamental to drug discovery, as it dictates the interpretation and processing of molecular data by computational models. Figure 2 illustrates four types of molecular representations: (1) sequence-based [45,46], (2) 2D graph-based [47], (3) image-based [48], and (4) 3D graph-based [49].

Sequence-based representations treat drug molecules as biological sequences, encoding them in a linear format such as SMILES [45], amino acid sequences [50], or InChI [51]. This approach is advantageous for employing techniques developed for genomics and proteomics, like sequence alignment and pattern recognition, which are effective for molecule design and similarity assessment. Two-dimensional (2D) graph-based representations visualize molecules as graphs with atoms as nodes and bonds as edges,

Table 1 Comparison of SSL across different molecular representations.

Representation	Strengths	Limitations
Sequence	1. Simple representation 2. Mature NLP-based models available 3. Efficient pretraining on large datasets	1. Structural information lost 2. Sensitive to minor input perturbations
2D Graph	1. Captures local and global molecular structures effectively	1. Computationally expensive compared to sequences 2. Limited scalability to large molecules
3D Graph	1. Encodes geometric and spatial molecular properties 2. Crucial for protein–ligand interaction modeling	1. Dependent on accurate 3D conformations 2. High computational cost
Image	1. Compatible with established vision-based SSL methods 2. Enables transfer learning from pretrained vision models	1. Structural connectivity not explicitly represented 2. Requires careful image preprocessing

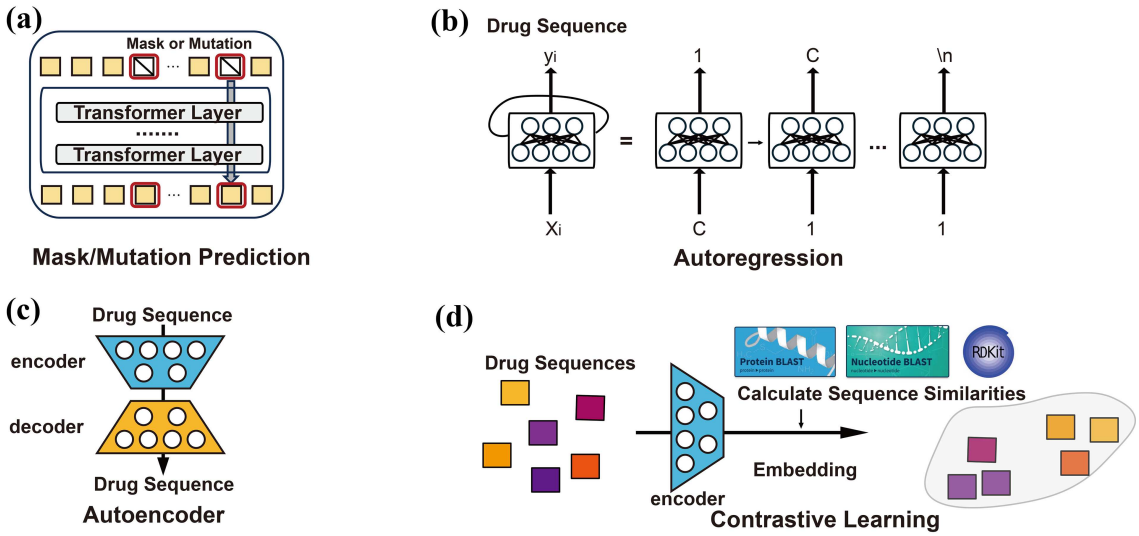

Figure 3 (Color online) Self-supervised learning in drug sequences is generally divided into four methods. (a) Mask/mutation prediction: it involves masking or mutating specific elements in a drug sequence. These modified sequences are then processed by transformer layers, which predict the altered elements to facilitate learning of sequence dependencies. (b) Autoregression: each element of the drug sequence is predicted sequentially based on its predecessors. This helps the model in learning temporal sequence patterns. (c) Autoencoder: this framework compresses a drug sequence into a lower-dimensional representation using an encoder and then reconstructs it through a decoder to capture essential sequence features. (d) Contrastive learning for sequence similarity: this method employs techniques where similarities between drug sequences are calculated and transformed into embeddings by an encoder, enabling effective differentiation between sequences.

facilitating the use of graph theory and network analysis to explore molecular structure and predict properties such as solubility or reactivity. Image-based representations transform molecular structures into 2D images, making them amenable to computer vision techniques. This method capitalizes on advances in deep learning for image classification and object recognition to identify and classify molecular patterns. Three-dimensional (3D) graph-based representations create detailed three-dimensional models of molecules, capturing spatial relationships and conformations vital for understanding drug interactions with targets like enzymes or receptors. Techniques such as molecular docking and dynamics simulations use these representations to predict the binding affinity and stability of drug-target complexes, essential for identifying lead compounds in drug design. Table 1 lists the advantages and disadvantages of the four different representations. By selecting the appropriate molecular representation, researchers can enhance the efficacy and accuracy of their computational models, thereby accelerating the discovery process and increasing the success rate of new drug development.

3.2 Self-supervised learning in drug sequences

In drug research and development, sequence-based representation methods are highly valued for their simplicity and have become the foundational approach for representing drug molecules. As depicted in Figure 3, SSL strategies are particularly promising for analyzing sequence data. They enhance the understanding of characteristics of drug molecule sequences and bolster the capability to discover new drug molecules through the following core technologies.

BERT [29], a classical model leveraging masked prediction for pre-training, has been widely used in natural language processing and has proven effective in understanding complex semantic structures [52].

As depicted in Figure 3(a), the method involves randomly masking parts of drug molecule sequences or introducing mutations (i.e., replacing them with new characters), and trains the model to predict these altered sections. This technique uses the sequences themselves as training targets, incorporating contextual semantics to deepen the comprehension of the model about the critical functional domains of drug molecules. Furthermore, by pre-training with a large corpus of unlabeled drug molecule data, the model develops rich features that enhance its performance in subsequent tasks, enabling more precise predictions of activity, toxicity assessments, or pharmacological analyses. For example, this model can be utilized for the rapid screening of new molecules to identify promising drug candidates with high affinity or specific targets [53, 54], thereby not only speeding up the drug development process but also helping to minimize the high costs associated with advancing unsuitable candidates into clinical trials.

In addition to predicting individual positions within sequences, generating complete sequences is an effective method of self-supervised training. As depicted in Figure 3(b), autoregressive models [55] such as recurrent neural networks (RNNs) based [56] process sequence data through a forward-generating process that anticipates subsequent elements within a sequence. This method proves invaluable for understanding and generating biological sequences, including small molecules [57], DNA [58], RNA [59], and protein sequences. In the field of drug discovery, these models not only serve as pre-training parameters for models predicting properties but also aid scientists in designing novel drug molecules from scratch and estimating the likelihood of new sequence occurrences. The training process of autoencoders (AE) [60], including variational autoencoders (VAE) [61], represents a distinct form of self-supervised learning. As shown in Figure 3(c), autoencoders map sequences to a feature space using an encoder and subsequently reconstruct the original sequence using a decoder. Similarly, in drug discovery, a well-trained encoder can provide pre-training parameters for property prediction models [62], facilitating not only the learning of data feature distributions but also enabling the trained decoder to generate new sequences.

Contrastive learning [37] is an effective strategy for sequence self-supervised learning, focusing on developing meaningful feature representations by comparing different data samples. This method trains models to reduce the distance between sequences considered positive samples in the feature space and to increase the distance from negative samples. The selection of positive and negative samples varies greatly, especially in drug discovery, where molecular sequences that share similar structures and functional properties are typically viewed as positive samples. As depicted in Figure 3(d), tools such as RDKit [63] and FASTA [64] facilitate molecular sequence similarity analysis, enabling the classification of highly similar sequences as positive samples and those that are less similar or dissimilar as negative samples. Based on this approach to sample construction, the training goal of the model is to draw positive samples closer together in the feature space while distancing negative samples. This strategy significantly improves the ability of the model to learn molecular features [38, 65], thereby enhancing its capability to identify and interpret more complex patterns.

3.3 Self-supervised learning in drug 2D graph

A graph-based model represents the molecular structure of a drug molecule as a graph, where atoms are depicted as nodes and chemical bonds as edges [66, 67]. This configuration enables the direct representation of molecular structural information and topological relationships, crucial for elucidating and predicting the chemical properties and biological activities of molecules. Figure 4 outlines several prevalent graph-based SSL methods employed in the field of drug discovery.

In Figure 4(a), drug molecules are represented as graphs, with nodes symbolizing atoms and edges depicting chemical bonds between them [68]. An encoder-decoder architecture is used to learn representations of these molecular graphs. Initially, the encoder converts the molecular graph into a latent space representation, capturing critical structural and chemical information. Subsequently, the decoder employs this latent representation to predict various properties of the molecule, such as atomic types or other potential chemical properties. Pre-training on these properties helps the model capture basic chemical characteristics, facilitating downstream drug development tasks.

Beyond property prediction, subgraph sampling and molecular graph reconstruction are key self-supervised techniques (Figure 4(b)) [69]. Subgraphs are typically generated through methods such as random sampling or based on node importance, and the decoder then tries to reconstruct subgraphs from the original graph. This process helps the model learn essential substructural features while omitting extraneous nodes or edges, thereby reducing the size of the graphs processed and consequently lowering computational complexity and memory demands. Such reductions make managing large-scale chemical

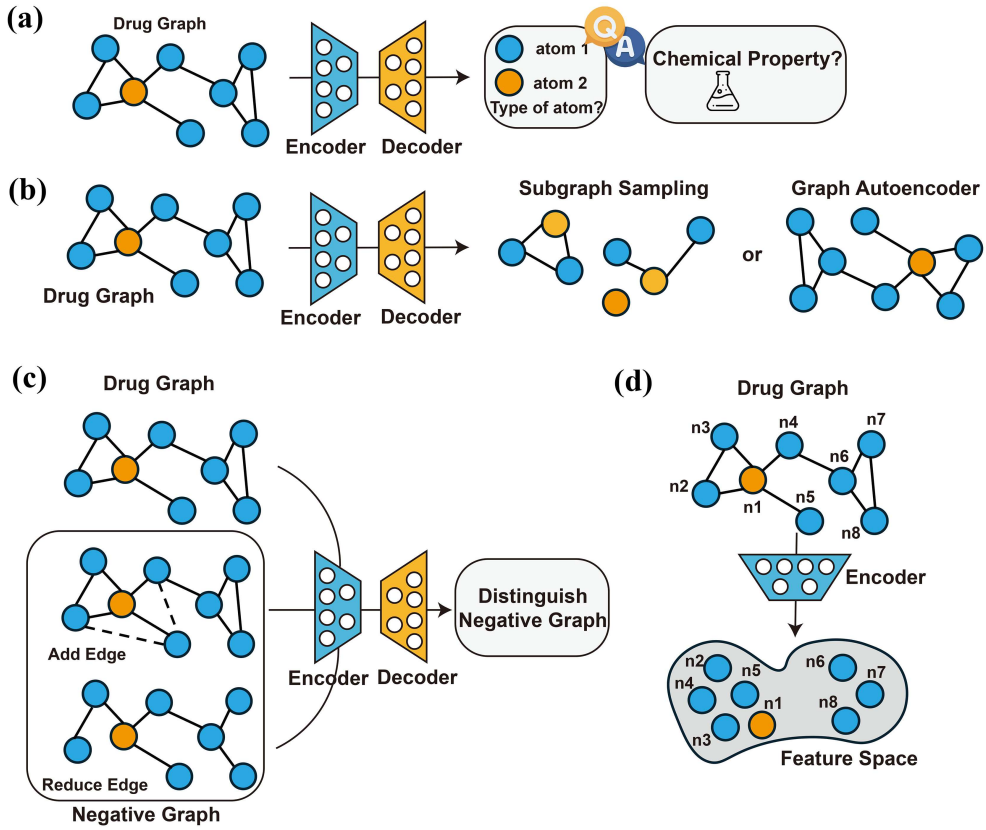

Figure 4 (Color online) Self-supervised learning for drug 2D graphs is generally divided into four methods. (a) Attribute prediction: the encoded features of a drug graph are utilized to predict the node type, molecular properties. (b) Sub-graph generation or graph autoencoder: the features of the drug graph are encoded by an encoder and reconstructed into either its sub-graph or the full graph. (c) Contrastive learning for samples: this involves randomly adding or removing edges from the drug graph and identifying these altered graphs as negative samples through contrastive learning. (d) Contrastive learning for nodes: a positive sample is defined as one where the path length is shorter than that of a reference node, and vice versa.

libraries more manageable, particularly during the early phases of drug discovery, which demand quick screening and assessment of numerous candidate compounds. Molecular graph reconstruction uses a graph autoencoder to fully reconstruct the entire input graph, requiring the model to learn enough information to accurately replicate the input. This ensures that the learned representations are both highly informative and functional. By reconstructing the entire graph structure, the graph autoencoder compels the model to capture all vital information within the molecule, thereby enhancing the effectiveness of the model in classification, regression, or other predictive tasks related to drug molecules.

Figures 4(c) and (d) introduce two graph-based contrastive learning methods [70]. The first method involves randomly modifying edges (either adding or removing) on the original drug graph to generate new graph variants that are considered negative samples. During this training process, the model must learn to distinguish between the features of the modified graph (negative samples) and the original graph. This approach deepens the model's understanding of how molecules respond to specific modifications by simulating potential chemical changes. Unlike constructing positive and negative samples from the entire graph, the second method constructs them based on individual nodes. For example, nodes are classified as positive or negative samples based on their path distance from a key node (in this case, the yellow node n1). The threshold for classification depends on the specific learning task and objectives, emphasizing the local structural information of drug molecules.

3.4 Self-supervised learning in drug 3D graph

3D molecular structures are primarily composed of the 3D coordinates of atoms, the types of atoms, and the chemical bonds between them, which together define the spatial configuration and properties. This accurate depiction of molecular geometry and stereochemistry is essential for understanding interactions with biological targets. In contrast, 2D molecular graphs provide a simplified planar view focusing

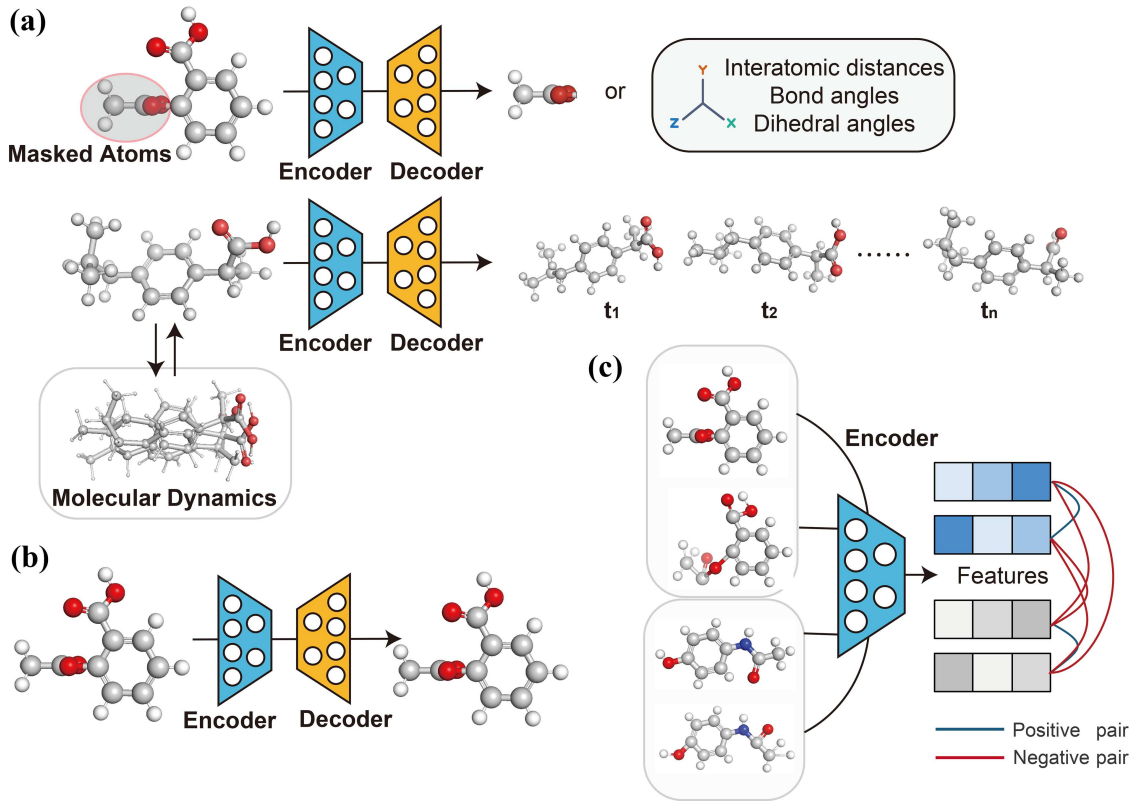

Figure 5 (Color online) Self-supervised learning for drug 3D graph structures is typically divided into three main approaches. (a) Attribute and molecular dynamics prediction: the encoded features of a 3D drug are utilized to predict the masked atoms, interatomic distance, bond angles, dihedral angles, or conformations of molecular dynamics. (b) 3D structural reconstruction: the features of a 3D drug are encoded by an encoder and reconstructed through a decoder. (c) Contrastive learning for conformations: randomly generate different conformations of the drug molecule. Conformations belonging to the same molecule are considered positive samples, and vice versa.

only on atom connectivity and bond arrangements, suitable for basic structural analysis. However, for detailed spatial insights needed in docking studies and molecular dynamics, 3D models are crucial. While various 3D representations exist, such as point clouds, voxel grids, and surface models, this review primarily focuses on 3D molecular graphs (3D-graph). Given the success of 3D graph neural networks (3D-GNNs) in SSL, 3D-graph has become a dominant representation in drug discovery. Taking a small molecule structure as an example, Figure 5 illustrates common 3D graph-based SSL methods used in drug discovery.

In Figure 5(a), the model is trained to predict certain masked atoms within 3D drug molecules or to predict geometric attributes such as interatomic distances, bond angles, and dihedral angles between atom pairs. This approach enables the model to learn and understand the spatial geometric structure of molecules, thereby enhancing the accuracy of predictions for molecular properties, activity, and interactions. In addition, dynamic conformational changes, as opposed to static conformations, better reflect the natural states of molecules. By utilizing trajectory data generated from molecular dynamics simulations [71], we can capture these conformational changes and dynamic behaviors over time. The model can predict structural changes within these time series, deeply learning the dynamic characteristics of molecules. Through this pre-training method, models grasp molecular motion patterns and conformational transition pathways, providing more accurate support for downstream tasks such as drug-target affinity or interaction prediction [72].

Generative models, such as VAEs, are also employed to learn latent representations of 3D drug molecules. In Figure 5(b), an encoder network maps high-dimensional molecular structures into a lower-dimensional latent space, effectively compressing essential information. Simultaneously, a decoder network reconstructs the molecular structures from these latent codes, aiming to minimize the reconstruction error between the original and regenerated molecules. This self-supervised approach allows the model to uncover complex patterns and associations within 3D molecular data, enhancing its capability to perform

downstream tasks such as generating new drug with desired conformations interacted with the target protein.

Figure 5(c) illustrates a classical contrastive learning framework based on three-dimensional molecular structures, wherein positive and negative sample pairs are constructed using different molecular conformations. Specifically, the method employs different conformations of the same molecule as positive sample pairs and conformations of distinct molecules as negative sample pairs. This strategy enables the model to learn feature representations that are invariant or robust to conformational changes, thereby effectively capturing the intrinsic three-dimensional structural characteristics.

3.5 Self-supervised learning in drug image

In recent years, image-based representations of drug molecules have gained popularity [48]. Unlike former representations, images offer a more intuitive and accessible format. This representation method visually presents the molecular structure, effectively utilizing advanced image processing technologies like convolutional neural networks (CNNs) [73] and Swin Transformers [74] to analyze complex intermolecular relationships. Additionally, it simplifies the understanding of the morphological and functional characteristics of molecules for non-specialists. Moreover, image-based representations fully capitalize on the latest advancements in computer vision, including pattern recognition and feature extraction, thereby providing new perspectives and tools for drug discovery.

Similar to SSL methods applied to sequences and graphs, images can also be processed to capture holistic features using encoders (Figure 6(a)) such as Swin Transformers [74]. This technique enables the encoder to analyze and capture the overall structural and textural patterns in an image. Subsequently, the representations learned are employed to predict fundamental properties of the subjects depicted in the images, such as the presence of specific chemical groups or structures. For instance, in the context of chemical imaging, as illustrated in Figure 6(a), the model may predict the presence of phenyl rings or hydroxyl groups based on the encoded features. This predictive capability is crucial for applications in fields like materials science and pharmacology, where an understanding of molecular compositions and configurations plays a pivotal role in influencing the synthesis and functionality of compounds.

Image reconstruction is a widely adopted self-supervised pre-training method in computer vision. This approach enables models to extract intrinsic features from images by learning to restore missing or distorted parts without external annotation. By attempting to reconstruct the original image, the model deepens its understanding of the data, allowing it to learn and articulate the fundamental structure and content of image data, thus improving its generalization abilities in the absence of extensive labeled datasets. As depicted in Figure 6(b), image reconstruction strategies are categorized into masking and disturbance strategies [8, 75]. The masking strategy is implemented at various molecular levels (Figure 6(b)), specifically: (1) atom level: the model learns to identify and infer the properties of various atoms within a molecule by predicting the types of masked atoms; (2) bond level: some bonds are masked, and the model is tasked with predicting the presence and types of these obscured bonds, enhancing its understanding of the connectivity within the molecular structure; (3) motif level: higher-level masking involves specific motifs, such as ring structures or larger molecular components. The model must reconstruct the entire molecular image in the absence of these key components, requiring a comprehensive understanding of the molecule's overall structure and functional areas. Through this masking strategy, the model not only improves its ability to handle complex data but also deepens its understanding of molecular structures, facilitating the recognition of functional features. This strategy provides an effective computational tool for fields such as drug discovery and molecular design, aiding in the development of new pharmaceuticals and the prediction of molecular functions.

The disturbance strategy uses the disarray or perturbation of image segments, employing multiple decoders for their reconstruction to train the model. This strategy is divided into three distinct substrategies (Figure 6(b)) [8, 75]. (1) Rationality prediction: the decoders assess whether each segment is correctly positioned by evaluating the rationality of the disturbed segments in relation to the entire image. This includes determining if these segments are logically placed within the image context. (2) Image reconstruction: even though some segments of the image are disturbed, this strategy requires the decoders to reconstruct the entire image. The decoders work to regenerate the full content from partial information, thus enhancing the model's understanding of the image. (3) Sequential prediction of blocks: this sub-strategy involves numbering the scrambled segments and predicting the correct sequence of these disturbed blocks. Through this approach, the model gains a deeper understanding of the

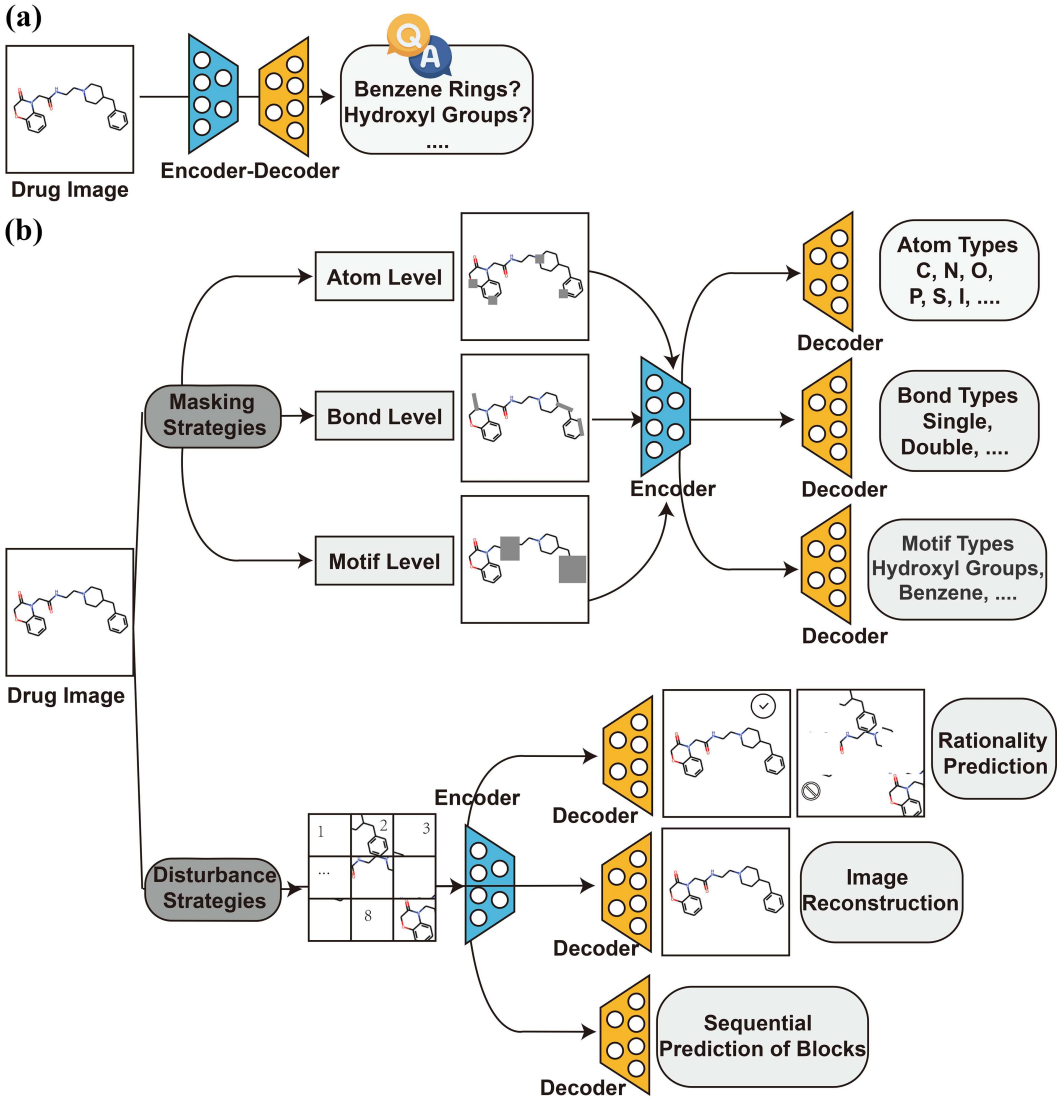

Figure 6 (Color online) Self-supervised learning for drug images is typically divided into two main approaches. (a) Attribute prediction: this approach uses the encoded features of a drug image to predict specific molecular properties, such as the presence of a benzene ring. (b) Reconstruction or generation strategies: there are two distinct strategies within this approach: masking and disturbance. Masking strategies involve reconstructing certain regions (e.g., atoms, bonds, or motifs) of a drug molecule image by masking them. Disturbance strategies involve predicting the reconstructed state or the order of picture blocks after dividing the image into several blocks.

overall structure of the image and the relationships between its segments. These sub-strategies collectively enhance the model's capability to handle complex data by deepening its understanding of spatial relationships and structural integrity within images.

4 Applications of self-supervised learning in drug discovery

SSL has emerged as a powerful approach in drug discovery, particularly for leveraging large-scale unlabeled molecular data. Unlike supervised learning, which requires extensive annotations, SSL can extract meaningful representations from unlabeled data, enhancing generalizability. Compared to physics-based simulations, SSL offers a data-driven approach that reduces computational costs while complementing mechanistic modeling. By integrating SSL with existing AI methods, pharmaceutical applications can achieve both improved efficiency and predictive accuracy, accelerating the drug discovery process. As illustrated in Figure 7, SSL has been extensively employed across five major areas in the field of drug discovery: drug repurposing [76], drug-target interaction (DTI or DTA) [77, 78], drug-drug interaction

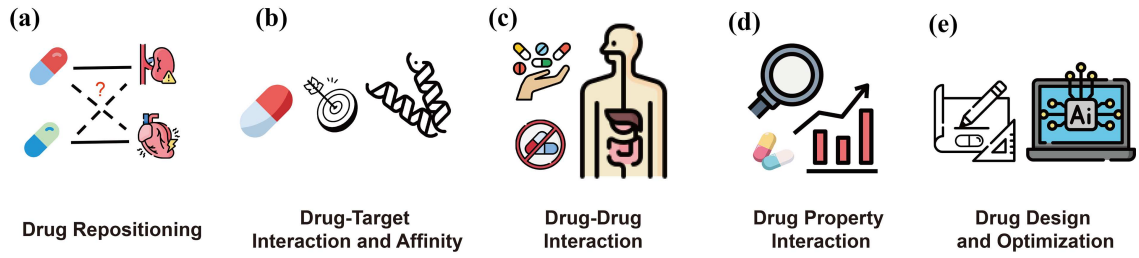

Figure 7 (Color online) Diverse applications of self-supervised learning in drug discovery. (a) Drug repositioning explores alternative uses for existing drugs through predictive modeling of biological processes. (b) Drug-target interaction and affinity evaluate potential drug-target binding affinities, enhancing the prediction of therapeutic effectiveness. (c) Drug-drug interaction examines the interactions between multiple drugs, aiding in the assessment of combination therapies. (d) Drug property interaction focuses on the relationship between drug properties and their biological effects, facilitating optimized drug profiles. (e) Drug design and optimization utilizes computational tools to design new drugs and optimize existing ones, leveraging AI-driven models to predict efficacy and safety profiles.

Table 2 Benchmark datasets commonly used for evaluating SSL models in drug discovery tasks.

Dataset	Description and application	Task(s) evaluated
DAVIS [82]	Experimentally validated binding affinities between small molecules and kinase targets, widely adopted for benchmarking DTI prediction methods.	Drug-target interaction prediction
BindingDB [83]	Comprehensive database of experimentally measured protein-ligand binding affinities, extensively used for drug design and virtual screening.	Drug-target interaction prediction
KIBA [84]	Integration of diverse drug-kinase binding data into a unified scoring system, enabling standardized comparison across prediction models.	Drug-target interaction prediction
MOSES [85]	Standardized benchmarking platform providing datasets and metrics (novelty, validity, diversity) to evaluate molecular generative models.	Molecular generation
ZINC [86]	Open chemical database containing millions of commercially available molecules, commonly utilized in molecular generation and virtual screening benchmarks.	Molecular generation, virtual screening
SAbDab [87]	Specialized database containing experimentally determined antibody sequences and structural data, serving as a benchmark for antibody modeling tasks.	Antibody sequence modeling

(DDI) [79], drug property prediction [27], and drug molecule design and optimization [80, 81]. In Table 2 [82–87], there are some benchmarks commonly used for evaluating SSL models in drug discovery application tasks. This section will comprehensively examine the specific applications of self-supervised learning in the research of various drug types, including small molecule drugs [88], peptide [89], vaccines [90], and antibody drugs [91], from their respective perspectives.

4.1 Small molecule drug discovery

In recent years, SSL techniques have gained prominence in small molecule drug discovery, propelled by the distinctive advantages of small molecules in data representation and processing. Small molecule drugs can be represented flexibly as sequences (SMILES), graphs, or images. This versatility in data representation not only enhances their utility in self-supervised learning frameworks but also allows for the elucidation of biological activity and pharmacological properties from multiple perspectives. Consequently, these approaches have substantially enhanced the model's accuracy and applicability.

For example, K-BERT proposed by Hou et al. [27] used three self-supervised pre-training tasks: prediction based on atomic features, prediction of molecular features, and contrastive learning, which effectively extracts chemical features from SMILES representations. This method endows the model with the analytical ability of a chemist and has demonstrated great potential in predicting molecular properties. TamGen [92] used a GPT-like chemical language model to pretrain a compound decoder on a random sample of 10 million SMILES from PubChem. The compounds generated by TamGen were shown to have better molecular quality and activity. In addition, TamGen further identified 14 compounds with significant inhibitory activity against the tuberculosis ClpP protease, emphasizing the practical potential and real-world applicability of the generative drug design approach. SADR [93] utilized data augmentation and contrastive learning strategies to learn the feature representation of nodes by randomly deleting nodes and their corresponding edges to construct negative samples. It has been extensively tested across three datasets, demonstrating its effectiveness in addressing the challenges posed by sparse datasets. GraphCL-DTA [70] applied graph contrastive learning to drug protein affinity prediction task. It added different random noises to the drug representation to generate two contrastive views and optimize the drug encoder directly, and not need a complex augmentation strategy. ImageMol [8] proposed a self-supervised image representation learning framework to predict of molecular properties and drug targets, which pretrained on 10 million unlabeled drug-like, bioactive molecules, to predict molecular targets of candidate compounds. It used three SSL methods based on molecule image to pretrain the model, in-

cluding predicting chemical structural information, distinguishing rational and irrational molecules and predicting rational permutations. ImageMol provided a general framework for other downstream tasks and has demonstrated its accuracy in identifying molecules that target SARS-CoV-2. Different from ImageMol, MaskMol [75] was an image pre-training framework for activity cliffs. By using pixel masking tasks, it extracted fine-grained information from molecular images, overcoming the limitations of existing deep learning models in identifying subtle structural changes. In recent years, many pharmaceutical companies have achieved success in small molecule drug design using SSL-based pre-trained models. Utilizing de novo generation technology, Insilico Medicine successfully nominated ISM1745 as a preclinical candidate drug targeting PRMT5 in PandaOmics [94]. MindRank has developed ProtMD [72], the first self-supervised pre-training model based on protein dynamics, which significantly improves the accuracy of drug-protein affinity prediction by introducing spatial and temporal dynamics of proteins, assisting medicinal chemistry experts to more accurately screen highly active small molecules and accelerating preclinical research and development.

4.2 Peptide drugs

Peptide drugs are characterized by their high specificity and potency, which reduce side effects and increase therapeutic efficacy. Their biocompatibility ensures low toxicity, and advances in synthetic techniques have enhanced their stability and functional versatility. Since the discovery of insulin nearly a century ago, over 80 peptide drugs have been introduced to the market [89].

Das et al. [95] pre-trained a generative autoencoder on the UniProt protein/peptide sequence database and used the trained classifier to screen whether the points in the generated sample space conformed to the target attributes (such as antibacterial activity and toxicity). Molecular dynamics simulations and synthetic experiments were used to verify two highly effective antibacterial agents against a variety of gram-positive and gram-negative pathogens. PeptideBERT [5] was a specialized protein language model. It can be used to accurately predict key peptide features including hemolysis, solubility and non-contamination by pre-training and then fine-tuning using BERT on amino acids. TPpred-SC [96] combined a pre-trained protein language model with multi-label supervised contrastive learning for predicting multifunctional therapeutic peptides. In TPpred-SC, each sample is selected as an anchor sample in turn. Samples with similar functional characteristics are classified as positive samples, while other samples are defined as negative samples. Experimental results show that TPrEd-SC outperforms existing related methods. DeepAMP [97], a peptide language-based deep generative framework, it screened 321 antimicrobial peptide sequences with relatively high activity (MIC lower than 2.5 against E. coli) from the GRAMPA 19 dataset, randomly masked up to 30% of the site length for each sequence, and finally generated a large number of optimized antimicrobial peptide data pairs containing low activity and high activity. Using the deepAMP, researchers designed and tested 29 AMP candidates in a two-round process, achieving over 90% efficacy against both Gram-positive and Gram-negative bacteria. Beyond the traditional amino acid sequence representation, some peptides, such as cyclic peptide drugs, can additionally be represented using SMILES sequences. The structural and functional properties of cyclopeptides lead to the possibility that they may contain non-standard amino acids in their natural state or during the synthesis process. CyclePermea [98] and MultiCycPermea [99] not only incorporated a peptide encoder based on a pre-trained BERT architecture, but also constructed positive and negative samples using SMILES sequence similarity, and further optimized the cyclopeptide representation using contrastive learning.

4.3 Vaccine development

Since the outbreak of the COVID-19 pandemic, deep learning has played a pivotal role in vaccine development [100, 101]. The capabilities of SSL have enabled researchers to more accurately predict antigenic epitopes of pathogens without explicit labels, facilitating the rapid identification and optimization of viable vaccine candidates. This advancement has not only accelerated vaccine development but also improved the ability to respond to emerging viral variants, which is crucial in addressing global health crises. Recently, mRNA-based vaccines and therapies are becoming increasingly widespread in the treatment of various diseases. CodonBERT [102] was a large language model for mRNA that takes codons as input. Through splicing and masking operations, it is trained with more than 10 million mRNA sequences from different organisms, enabling the model to capture important biological concepts. CodonBERT can also be extended to predict various mRNA properties and performs well on the new crown vaccine dataset. As

a pretrained protein language model, ESM [103] can accurately predict protein structures and functions, providing essential insights into identifying targets for immunotherapy and optimizing vaccine antigen design. Aziz et al. [104] employed immunoinformatics and reverse vaccinology methods combined with deep learning to predict the protein structures of NeoCoV, designing multi-epitope vaccines against this virus and simulating their interactions with immune receptors (TLRs and MHC), thereby significantly reducing the cost and duration of vaccine development. pMTnet-omni [105] utilized ESM to precisely predict the interactions between T-cell receptors (TCR) and peptide-MHC complexes (pMHC), revealing the molecular mechanisms underlying antigen recognition by T cells. This approach facilitates the identification of immunotherapy targets and the selection of high-affinity epitopes, advancing personalized immunotherapies and novel vaccine development.

4.4 Antibody design

Antibodies play a crucial role in drug discovery, primarily by specifically recognizing and binding to disease-related targets such as proteins or pathogen surface molecules, which block or regulate biological processes and ameliorate disease symptoms. Their high specificity and customizability render antibodies ideal therapeutic agents for a wide range of diseases, including cancer, autoimmune diseases, and inflammatory conditions. Furthermore, antibody drugs typically exhibit high efficacy and relatively low side effects, positioning them as central elements in contemporary drug discovery and treatment strategies.

CDRH3 [106] is a key structural domain in antibody molecules, located in the variable region of the antibody and is critical for determining the specificity and affinity of the antibody. PALM-H3 [107] has been proposed for the redesign of the CDRH3 region to endow it with the desired antigen-binding specificity. Initially, a masked language model based on ESM2 [108] was applied for pre-training on unpaired antibody heavy and light chain sequences. Subsequently, the pre-trained model RoFormer was used together with the paired affinity data to construct and fine-tune the A2Binder. Ultimately, these models were employed to generate and train PALM-H3. Through computational simulations and in vitro experimentation, antibodies produced by PALM-H3 have demonstrated significant binding activity with the SARS-CoV-2 antigen, including the newly emerged XBB variant. IgLM [109] designed antibodies based on text-filling in language, enabling it to redesign CDR in antibody sequences using bidirectional context. It was pretrained by autoregressive prediction rather than MLM and generated a sequence based on conditional tokens. It has been shown that IgLM is capable of producing complete antibody sequences from diverse species. Additionally, its infilling approach facilitates the creation of complementarity-determining region (CDR) loop libraries, which are enhanced by better in silico developability profiles.

5 Challenges and prospects

In the field of drug discovery, SSL has demonstrated significant potential, yet it continues to face several challenges [110]. (1) Data quality and consistency [111]: in biomedicine, data often originates from diverse sources and varies in quality. For example, obtaining high-quality 3D molecular structure data is challenging and costly, and computationally generated conformations may lack sufficient accuracy. (2) Complexity and interpretability of models [112]: drug discovery involves intricate biological mechanisms and constantly evolving biomarkers. While SSL models are adept at identifying patterns within the data, their internal decision-making processes remain opaque. This lack of clarity is particularly problematic in drug development, where explicit explanations of model predictions are essential to comply with regulatory standards. (3) Model efficiency [113]: drug discovery necessitates the analysis of extensive datasets encompassing a wide array of molecular structures and biomarkers. For instance, processing image data-particularly high-resolution images-typically requires substantial computational resources. This demand becomes especially pronounced when handling large molecules, thereby constraining the speed and efficiency of model processing. SSL models demand significant computational power and sophisticated algorithms to efficiently process large-scale data, ensuring that training time and resource utilization are optimized. (4) The application of self-supervised learning in drug discovery presents several challenges, including data privacy protection and fairness, necessitating the anonymization of patient information and the mitigation of biases that may affect drug applicability. The predicted drug candidates must undergo rigorous experimental validation to minimize the risks of false discoveries and potential toxicity. Additionally, issues related to intellectual property rights and legal responsibilities

must be clearly defined to ensure compliance and safety, thereby facilitating the reliable integration of self-supervised learning in drug discovery.

We can anticipate several promising directions for expanding the application of SSL in drug discovery. Firstly, integrating SSL with reinforcement learning could significantly enhance the efficiency of new drug research and development. SSL enables the extraction of rich feature representations from drug molecules without requiring labeled data. These features can serve as inputs for reinforcement learning models or be used to refine pre-trained models through reinforcement learning, thus accelerating and improving decision-making and optimization processes [114, 115]. Secondly, merging SSL with multi-view learning is another avenue that merits attention [116, 117]. SSL with multi-view representations of drug molecules offers a promising new avenue for drug development. Multi-view representations can capture various aspects of molecular data, including two-dimensional structural formulas and three-dimensional conformations. This approach enables the extraction of comprehensive molecular embeddings that integrate information from multiple perspectives, thereby enhancing the performance of downstream tasks. Finally, future research should aim to develop scalable SSL frameworks that can be standardized across various stages and goals of drug discovery [13]. Moreover, establishing standardized data processing and model training protocols will enhance the reproducibility of research and the comparability of results, ultimately advancing the entire field of drug discovery.

In summary, the future application of SSL in drug discovery is multifaceted, and the integration of additional learning methods is anticipated to further drive innovation and development in this field. These advancements are expected to enhance efficiency and reduce R&D costs in the discovery of new drugs.

6 Conclusion

In this review, we emphasize the transformative impact of SSL in drug discovery. We began by examining drug molecules represented in various formats, such as sequences, graphs, images, and 3D models, demonstrating how SSL methods (e.g., predictive modeling, generative tasks, and contrastive learning) can utilize large amounts of unlabeled data to enhance model training. This enhancement significantly improves the models' effectiveness and applicability in addressing complex pharmacological challenges. Additionally, we detailed specific applications of SSL in fields such as small molecule drug design, peptide drug development, vaccine optimization, and antibody engineering.

Despite these advancements, SSL in drug discovery encounters challenges with data quality, model interpretability, and computational demands. Addressing these issues through the integration of reinforcement and multi-task learning could further enhance its efficacy, scalability, and cost-effectiveness. By evolving SSL frameworks and establishing standardized protocols, the field of drug discovery is well-positioned to reach unprecedented levels of efficiency and innovation, thereby accelerating the creation of safer and more effective therapeutic solutions.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. U22A2037, 62425204, 62122025, 62450002, 62432011) and Beijing Natural Science Foundation (Grant No. L248013).

References

1 Drews J. Drug discovery: a historical perspective. Science, 2000, 287: 1960–1964
2 LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521: 436–444
3 Chen H, Engkvist O, Wang Y, et al. The rise of deep learning in drug discovery. Drug Discov Today, 2018, 23: 1241–1250
4 Zhang L, Tan J, Han D, et al. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. Drug Discov Today, 2017, 22: 1680–1685
5 Zhang H, Saravanan K M, Wei Y, et al. Deep learning-based bioactive therapeutic peptide generation and screening. J Chem Inf Model, 2023, 63: 835–845
6 Lavecchia A. Deep learning in drug discovery: opportunities, challenges and future prospects. Drug Discov Today, 2019, 24: 2017–2032
7 Jing Y, Bian Y, Hu Z, et al. Deep learning for drug design: an artificial intelligence paradigm for drug discovery in the big data era. AAPS J, 2018, 20: 58
8 Zeng X, Xiang H, Yu L, et al. Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. Nat Mach Intell, 2022, 4: 1004–1016
9 Seeger M. Learning with labeled and unlabeled data. 2000. https://infoscience.epfl.ch/entities/publication/5571b600-619d-4ede-93e5-3ae1e036443a
10 Liu X, Zhang F, Hou Z, et al. Self-supervised learning: generative or contrastive. IEEE Trans Knowl Data Eng, 2023, 35: 857–876
11 Krishnan R, Rajpurkar P, Topol E J. Self-supervised learning in medicine and healthcare. Nat Biomed Eng, 2022, 6: 1346–1352
12 Jaiswal A, Babu A R, Zadeh M Z, et al. A survey on contrastive self-supervised learning. Technologies, 2020, 9: 2

13 Hendrycks D, Mazeika M, Kadavath S, et al. Using self-supervised learning can improve model robustness and uncertainty. In: Proceedings of Advances in Neural Information Processing Systems, 2019
14 Zhai X, Oliver A, Kolesnikov A, et al. S4L: self-supervised semi-supervised learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019. 1476–1485
15 Tsai Y H H, Wu Y, Salakhutdinov R, et al. Self-supervised learning from a multi-view perspective. 2020. ArXiv:2006.05576
16 Schiappa M C, Rawat Y S, Shah M. Self-supervised learning for videos: a survey. ACM Comput Surv, 2023, 55: 1–37
17 Misra I, van der Maaten L. Self-supervised learning of pretext-invariant representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 6707–6717
18 Nasteski V. An overview of the supervised machine learning methods. Horizons, 2017, 4: 56
19 Hastie T, Tibshirani R, Friedman J, et al. Overview of supervised learning. In: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Berlin: Springer, 2009. 9–41
20 Cunningham P, Cord M, Delany S J. Supervised learning. In: Machine Learning Techniques for Multimedia. Berlin: Springer, 2008
21 Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd International Conference on Machine Learning, 2006. 161–168
22 Barlow H B. Unsupervised Learning. Neural Comput, 1989, 1: 295–311
23 Dietterich T. Overfitting and undercomputing in machine learning. ACM Comput Surv, 1995, 27: 326–327
24 Rice L, Wong E, Kolter Z. Overfitting in adversarially robust deep learning. In: Proceedings of the 37th International Conference on Machine Learning, 2020. 8093–8104
25 Bousquet O, Elisseeff A. Stability and generalization. J Mach Learn Res, 2002, 2: 499–526
26 Wu H, Prasad S. Semi-supervised deep learning using pseudo labels for hyperspectral image classification. IEEE Trans Image Process, 2017, 27: 1259–1270
27 Wu Z, Jiang D, Wang J, et al. Knowledge-based BERT: a method to extract molecular features like computational chemists. Brief BioInf, 2022, 23: bbac131
28 Erhan D, Courville A, Bengio Y, et al. Why does unsupervised pre-training help deep learning? In: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, 2010. 201–208
29 Devlin J. BERT: pre-training of deep bidirectional transformers for language understanding. 2018. ArXiv:1810.04805
30 Sun Y, Wang X, Tang X. Deep learning face representation from predicting 10,000 classes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014. 1891–1898
31 Emmert-Streib F, Yang Z, Feng H, et al. An introductory review of deep learning for prediction models with big data. Front Artif Intell, 2020, 3: 4
32 Cova T F G G, Pais A A C C. Deep learning for deep chemistry: optimizing the prediction of chemical patterns. Front Chem, 2019, 7: 809
33 Samanta B, De A, Jana G, et al. NEVAE: a deep generative model for molecular graphs. J Mach Learn Res, 2020, 21: 1–33
34 Salakhutdinov R. Learning deep generative models. Annu Rev Stat Appl, 2015, 2: 361–385
35 Ruthotto L, Haber E. An introduction to deep generative modeling. GAMM-Mitt, 2021, 44: e202100008
36 You Y, Chen T, Sui Y, et al. Graph contrastive learning with augmentations. In: Proceedings of Advances in Neural Information Processing Systems, 2020. 5812–5823
37 Tian Y, Sun C, Poole B, et al. What makes for good views for contrastive learning? In: Proceedings of Advances in Neural Information Processing Systems, 2020. 6827–6839
38 Lee B, Shin D. Contrastive learning for enhancing feature extraction in anticancer peptides. Brief BioInf, 2024, 25: bbae220
39 Le-Khac P H, Healy G, Smeaton A F. Contrastive representation learning: a framework and review. IEEE Access, 2020, 8: 193907
40 Chuang C Y, Robinson J, Lin Y C, et al. Debiased contrastive learning. In: Proceedings of Advances in Neural Information Processing Systems, 2020. 8765–8775
41 Yang L, Zhang Z, Song Y, et al. Diffusion models: a comprehensive survey of methods and applications. ACM Comput Surv, 2024, 56: 1–39
42 Patani G A, LaVoie E J. Bioisosterism: a rational approach in drug design. Chem Rev, 1996, 96: 3147–3176
43 Mandal S, Moudgil M, Mandal S K. Rational drug design. Eur J Pharmacol, 2009, 625: 90–100
44 David L, Thakkar A, Mercado R, et al. Molecular representations in AI-driven drug discovery: a review and practical guide. J Cheminform, 2020, 12: 56
45 Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci, 1988, 28: 31–36
46 Pearson W R. Using the FASTA program to search protein and DNA sequence databases. In: Computer Analysis of Sequence Data: Part I. Berlin: Springer, 1994. 307–331
47 Li J, Cai D, He X. Learning graph-level representation for drug discovery. 2017. ArXiv:1709.03741
48 Li Y, Liu B, Deng J, et al. Image-based molecular representation learning for drug development: a survey. Brief BioInf, 2024, 25: bbae294
49 Martin Y C. 3D database searching in drug design. J Med Chem, 1992, 35: 2145–2154
50 Biemann K, Papayannopoulos I A. Amino acid sequencing of proteins. Acc Chem Res, 1994, 27: 370–378
51 Heller S R, McNaught A, Pletnev I, et al. InChI, the IUPAC international chemical identifier. J Cheminform, 2015, 7: 1–34
52 Jawahar G, Sagot B, Seddah D. What does BERT learn about the structure of language? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019
53 Zheng J, Xiao X, Qiu W R. DTI-BERT: identifying drug-target interactions in cellular networking based on BERT and deep learning method. Front Genet, 2022, 13: 859188
54 Simon E, Bankapur S S. Prediction of drug-target interactions using BERT for protein sequences and drug compound. In: Proceedings of the 16th International Conference on Communication Systems & NETworkS (COMSNETS), 2024. 436–438
55 Gregor K, Danihelka I, Mnih A, et al. Deep autoregressive networks. In: Proceedings of International Conference on Machine Learning, 2014. 1242–1250
56 Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Phys D-Nonlinear Phenomena, 2020, 404: 132306
57 Bedard P L, Hyman D M, Davids M S, et al. Small molecules, big impact: 20 years of targeted therapy in oncology. Lancet, 2020, 395: 1078–1088
58 Cluzel P, Lebrun A, Heller C, et al. DNA: an extensible molecule. Science, 1996, 271: 792–794
59 Corley M, Burns M C, Yeo G W. How RNA-binding proteins interact with RNA: molecules and mechanisms. Mol Cell, 2020, 78: 9–29
60 Michelucci U. An introduction to autoencoders. 2022. ArXiv:2201.03898
61 Kingma D P, Welling M. An introduction to variational autoencoders. FNT Machine Learn, 2019, 12: 307–392
62 Hu Q, Feng M, Lai L, et al. Prediction of drug-likeness using deep autoencoder neural networks. Front Genet, 2018, 9: 585
63 Landrum G. RDKit: a software suite for cheminformatics, computational chemistry, and predictive modeling. Greg Landrum, 2013, 8: 5281

64 Donkor E S, Dayie N, Adiku T K. Bioinformatics with basic local alignment search tool (BLAST) and fast alignment (FASTA). J Bioinformat Sequence Anal, 2014, 6: 1–6
65 Zhang R, Wang X, Wang P, et al. HTCL-DDI: a hierarchical triple-view contrastive learning framework for drug-drug interaction prediction. Brief BioInf, 2023, 24: bbad324
66 Kearnes S, McCloskey K, Berndl M, et al. Molecular graph convolutions: moving beyond fingerprints. J Comput Aided Mol Des, 2016, 30: 595–608
67 García-Domenech R, Gálvez J, de Julián-Ortiz J V, et al. Some new trends in chemical graph theory. Chem Rev, 2008, 108: 1127–1169
68 Ren S, Yu L, Gao L, et al. Multidrug representation learning based on pretraining model and molecular graph for drug interaction and combination prediction. Bioinformatics, 2022, 38: 4387–4394
69 Sun M, Zhao S, Gilvary C, et al. Graph convolutional networks for computational drug development and discovery. Brief BioInf, 2020, 21: 919–935
70 Yang X, Yang G, Chu J. GraphCL-DTA: a graph contrastive learning with molecular semantics for drug-target binding affinity prediction. IEEE J Biomed Health Inform, 2024, 28: 4544–4552
71 Karplus M, Petsko G A. Molecular dynamics simulations in biology. Nature, 1990, 347: 631–639
72 Wu F, Jin S, Jiang Y, et al. Pre-training of equivariant graph matching networks with conformation flexibility for drug binding. Adv Sci, 2022, 9: 2203796
73 Li Z, Liu F, Yang W, et al. A survey of convolutional neural networks: analysis, applications, and prospects. IEEE Trans Neural Netw Learn Syst, 2021, 33: 6999–7019
74 Liu Z, Lin Y, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 10012–10022
75 Cheng Z, Xiang H, Ma P, et al. MaskMol: knowledge-guided molecular image pre-training framework for activity cliffs. 2024. ArXiv:2409.12926
76 Pushpakom S, Iorio F, Eyers P A, et al. Drug repurposing: progress, challenges and recommendations. Nat Rev Drug Discov, 2019, 18: 41–58
77 Chen X, Yan C C, Zhang X, et al. Drug-target interaction prediction: databases, web servers and computational models. Brief Bioinform, 2016, 17: 696–712
78 Zhang Y, Hu Y, Han N, et al. A survey of drug-target interaction and affinity prediction methods via graph neural networks. Comput Biol Med, 2023, 163: 107136
79 Han K, Cao P, Wang Y, et al. A review of approaches for predicting drug-drug interactions based on machine learning. Front Pharmacol, 2022, 12: 814858
80 Song C M, Lim S J, Tong J C. Recent advances in computer-aided drug design. Brief BioInf, 2009, 10: 579–591
81 Prada-Gracia D, Huerta-Yépez S, Moreno-Vargas L M. Application of computational methods for anticancer drug discovery, design, and optimization. Boletín Médico Del Hospital Infantil de México (Engl Ed), 2016, 73: 411–423
82 Davis M I, Hunt J P, Herrgard S, et al. Comprehensive analysis of kinase inhibitor selectivity. Nat Biotechnol, 2011, 29: 1046–1051
83 Liu T, Lin Y, Wen X, et al. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic Acids Res, 2007, 35: D198–D201
84 Tang J, Szwajda A, Shakyawar S, et al. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. J Chem Inf Model, 2014, 54: 735–743
85 Polykovskiy D, Zhebrak A, Sanchez-Lengeling B, et al. Molecular sets (MOSES): a benchmarking platform for molecular generation models. Front Pharmacol, 2020, 11: 565644
86 Sterling T, Irwin J J. ZINC 15—ligand discovery for everyone. J Chem Inf Model, 2015, 55: 2324–2337
87 Dunbar J, Krawczyk K, Leem J, et al. SAbDab: the structural antibody database. Nucleic Acids Res, 2014, 42: D1140–D1146
88 Jayatunga M K P, Xie W, Ruder L, et al. AI in small-molecule drug discovery: a coming wave? Nat Rev Drug Discov, 2022, 21: 175–176
89 Muttenthaler M, King G F, Adams D J, et al. Trends in peptide drug discovery. Nat Rev Drug Discov, 2021, 20: 309–325
90 Plotkin S A. Vaccines: past, present and future. Nat Med, 2005, 11: S5–S11
91 Schrama D, Reisfeld R A, Becker J C. Antibody targeted drugs as cancer therapeutics. Nat Rev Drug Discov, 2006, 5: 147–159
92 Wu K, Xia Y, Deng P, et al. TamGen: drug design with target-aware molecule generation through a chemical language model. Nat Commun, 2024, 15: 9360
93 Jin S, Zhang Y, Yu H, et al. SADR: self-supervised graph learning with adaptive denoising for drug repositioning. IEEE ACM Trans Comput Biol Bioinf, 2024, 21: 265–277
94 Ren F, Aliper A, Chen J, et al. A small-molecule TNIK inhibitor targets fibrosis in preclinical and clinical models. Nat Biotechnol, 2025, 43: 63–75
95 Das P, Sercu T, Wadhawan K, et al. Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. Nat Biomed Eng, 2021, 5: 613–623
96 Yan K, Lv H W, Shao J Y, et al. TPpred-SC: multi-functional therapeutic peptide prediction based on multi-label supervised contrastive learning. Sci China Inf Sci, 2024, 67: 212105
97 Li T, Ren X, Luo X, et al. A foundation model identifies broad-spectrum antimicrobial peptides against drug-resistant bacterial infection. Nat Commun, 2024, 15: 7538
98 Wang Z, Chen Y, Ye X, et al. CyclePermea: membrane permeability prediction of cyclic peptides with a multi-loss fusion network. In: Proceedings of International Joint Conference on Neural Networks (IJCNN), 2024. 1–8
99 Wang Z, Chen Y, Shang Y, et al. MultiCycPermea: accurate and interpretable prediction of cyclic peptide permeability using a multimodal image-sequence model. BMC Biol, 2025, 23: 63
100 Hederman A P, Ackerman M E. Leveraging deep learning to improve vaccine design. Trends Immunol, 2023, 44: 333–344
101 Sarmadi A, Hassanzadeganroudsari M, Soltani M. Artificial intelligence and machine learning applications in vaccine development. In: Bioinformatics Tools for Pharmaceutical Drug Product Development. Hoboken: Wiley, 2023. 233–253
102 Li S, Moayedpour S, Li R, et al. CodonBERT large language model for mRNA vaccines. Genome Res, 2024, 34: 1027–1035
103 Lin Z, Akin H, Rao R, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. bioRxiv, 2022. doi: 10.1101/2022.07.20.500902
104 Aziz S, Waqas M, Halim S A, et al. Exploring whole proteome to contrive multi-epitope-based vaccine for NeoCoV: an immunoinformtics and in-silico approach. Front Immunol, 2022, 13: 956776
105 Han Y, Yang Y, Tian Y, et al. Pan-MHC and cross-species prediction of T cell receptor-antigen binding. bioRxiv, 2023. doi: 10.1101/2023.12.01.569599
106 Jones P T, Dear P H, Foote J, et al. Replacing the complementarity-determining regions in a human antibody with those from a mouse. Nature, 1986, 321: 522–525
107 He H, He B, Guan L, et al. De novo generation of SARS-CoV-2 antibody CDRH3 with a pre-trained generative large language model. Nat Commun, 2024, 15: 6867
108 Rives A, Meier J, Sercu T, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million

protein sequences. Proc Natl Acad Sci USA, 2021, 118: e2016239118
109 Shuai R W, Ruffolo J A, Gray J J. IgLM: Infilling language modeling for antibody sequence design. Cell Syst, 2023, 14: 979–989.e4
110 Pink R, Hudson A, Mouriés M A, et al. Opportunities and challenges in antiparasitic drug discovery. Nat Rev Drug Discov, 2005, 4: 727–740
111 Cong G, Fan W, Geerts F, et al. Improving data quality: consistency and accuracy. In: Proceedings of the 33rd International Conference on Very Large Data Bases, 2007. 315–326
112 Stiglic G, Kocbek P, Fijacko N, et al. Interpretability of machine learning-based prediction models in healthcare. WIREs Data Min Knowl, 2020, 10: e1379
113 Al-Jarrah O Y, Yoo P D, Muhaidat S, et al. Efficient machine learning for big data: a review. Big Data Res, 2015, 2: 87–93
114 Li X, Shang J, Das S, et al. Does self-supervised learning really improve reinforcement learning from pixels? In: Proceedings of Advances in Neural Information Processing Systems, 2022. 30865–30881
115 Buchler U, Brattoli B, Ommer B. Improving spatiotemporal self-supervision by deep reinforcement learning. In: Proceedings of the European Conference on Computer Vision (ECCV), 2018. 770–786
116 Yan X, Hu S, Mao Y, et al. Deep multi-view learning methods: a review. Neurocomputing, 2021, 448: 106–129
117 Geng C, Tan Z, Chen S. A multi-view perspective of self-supervised learning. 2020. ArXiv:2003.00877