Computational and Structural Biotechnology Journal 20 (2022) 4288–4304

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Mini review

On modeling and utilizing chemical compound information with deep learning technologies: A task-oriented approach

Sangsoo Lim [a,1], Sangseon Lee [b,1], Yinhua Piao [c], MinGyu Choi [d,g], Dongmin Bang [e], Jeonghyeon Gu [f], Sun Kim [c,f,g,h,*]

[a] Bioinformatics Institute, Seoul National University, Gwanak-ro 1, Gwanak-gu, Seoul 08826, South Korea
[b] Institute of Computer Technology, Seoul National University, Gwanak-ro 1, Gwanak-gu, Seoul 08826, South Korea
[c] Department of Computer Science and Engineering, Seoul National University, Gwanak-ro 1, Gwanak-gu, Seoul 08826, South Korea
[d] Department of Chemistry, Seoul National University, Gwanak-ro 1, Gwanak-gu, Seoul 08826, South Korea
[e] Interdisciplinary Program in Bioinformatics, Seoul National University, Gwanak-ro 1, Gwanak-gu, Seoul 08826, South Korea
[f] Interdisciplinary Program in Artificial Intelligence, Seoul National University, Gwanak-ro 1, Gwanak-gu, Seoul 08826, South Korea
[g] AIGENDRUG Co., Ltd., Gwanak-ro 1, Gwanak-gu, Seoul 08826, South Korea
[h] MOGAM Institute for Biomedical Research, Yong-in 16924, South Korea

ARTICLE INFO

Article history:
Received 2 April 2022
Received in revised form 29 July 2022
Accepted 29 July 2022
Available online 5 August 2022

Keywords:
Chemical space
Deep learning
Computer-aided drug discovery
Data augmentation
Chemical information modeling

ABSTRACT

A large number of chemical compounds are available in databases such as PubChem and ZINC. However, currently known compounds, though large, represent only a fraction of possible compounds, which is known as chemical space. Many of these compounds in the databases are annotated with properties and assay data that can be used for drug discovery efforts. For this goal, a number of machine learning algorithms have been developed and recent deep learning technologies can be effectively used to navigate chemical space, especially for unknown chemical compounds, in terms of drug-related tasks. In this article, we survey how deep learning technologies can model and utilize chemical compound information in a task-oriented way by exploiting annotated properties and assay data in the chemical compounds databases. We first compile what kind of tasks are trying to be accomplished by machine learning methods. Then, we survey deep learning technologies to show their modeling power and current applications for accomplishing drug related tasks. Next, we survey deep learning techniques to address the insufficiency issue of annotated data for more effective navigation of chemical space. Chemical compound information alone may not be powerful enough for drug related tasks, thus we survey what kind of information, such as assay and gene expression data, can be used to improve the prediction power of deep learning models. Finally, we conclude this survey with four important newly developed technologies that are yet to be fully incorporated into computational analysis of chemical information.

Contents

* Corresponding author.
[1] Equal contribution.

https://doi.org/10.1016/j.csbj.2022.07.049

1. Introduction

Chemical space in drug discovery refers to a collection of chemical compounds satisfying a certain set of properties, and definitions of chemical space vary widely depending on the criteria [1–4]. Out of theoretically possible drug-like chemical compound space, as large as 10^{60} compounds [5], the number of known/identified chemical compounds ranges from thousands in DrugBank to tens of millions in PubChem or ZINC, depending on definitions of drug-like chemical space [6,7]. Before the extensive use of Computer-aided drug discovery (CADD), Lipinski's rule of five (RO5) has long been accepted as the general rule for drug-likeness of a compound [8]. Recent studies reported that there are drugs like Atazanavir, Erythromycin, or Sirolimus that disobey the RO5 in extended/beyond the RO5 space [9,10]. These examples show the difficulty of defining chemical space in terms of specific criteria for chemical compounds. Our current knowledge on properties or characteristics of chemical compounds is still not enough to define chemical space. *One promising alternative is to learn chemical space directly from data.*

There have recently been remarkable advances in artificial intelligence technologies, deep learning in particular, and these technologies have been successfully used to model properties or characteristics of chemical compounds in the context of "tasks" such as absorption, distribution, metabolism, excretion, or toxicity (ADMET) predictions. Defining chemical space in terms of tasks is a supervised learning where tasks are defined quantitatively using the assay data or clinical data. Since chemical compounds were difficult to handle as graph representations, common approaches to represent chemical compounds are to use linear representation, e.g., SMILES, or fingerprint-based representation, e.g., MACCS. Traditional feature-based methods, such as random forest or support vector machine, take these representations as input and identify important features that are effective in performing given tasks for chemical compounds. Deep learning models also use linear or fingerprint representations of compounds as input but learn alternative representations of chemical compounds as *embedding vectors* (See Section 3) when accomplishing given tasks. One major advantage of embedding vectors is that they can be used to compare compounds more effectively by computing similarity among embedding vectors. This is one of the reasons why recent deep learning models outperform traditional feature-based methods. Another important use of embedding vector of chemical compounds is to build chemical space in more general settings before addressing downstream tasks. This general representation of chemical space, known as "pre-training" strategies, can be then specialized for specific tasks. Even with powerful representations of chemical compounds, it is also important to incorporate valuable traditional knowledge such as chemical properties measured by assays, thus many deep learning-based methods use parallel architecture of learning chemical space and utilizing chemical properties to achieve prediction power for specific tasks.

Our survey paper is to summarize this new development in a single article so that drug research community can reach a better understanding in these technologies better and utilize recent computational methods more effectively. We organize the survey as summarized in Fig. 1. Tasks that can be accomplished with chemical information are summarized in Section 2. Recently developed deep learning methods are summarized in terms of technical methods and also tasks in Section 3. Computational methods for creating general representation of chemical space, known as pre-training strategies, are summarized in Section 4. Section 5 discuss how much improvement can be made when traditional knowledge such as chemical properties measured by assays is incorporated into deep learning models. Finally, we discuss four topics important for developing more powerful methods that can be used to accomplish drug discovery tasks more accurately.

2. Tasks: What Can We Do with Chemical Compound Information?

Chemical compound information can be used by computational methods to accomplish tasks such as ADMET prediction. Evaluation on the performance of computational methods requires accu-

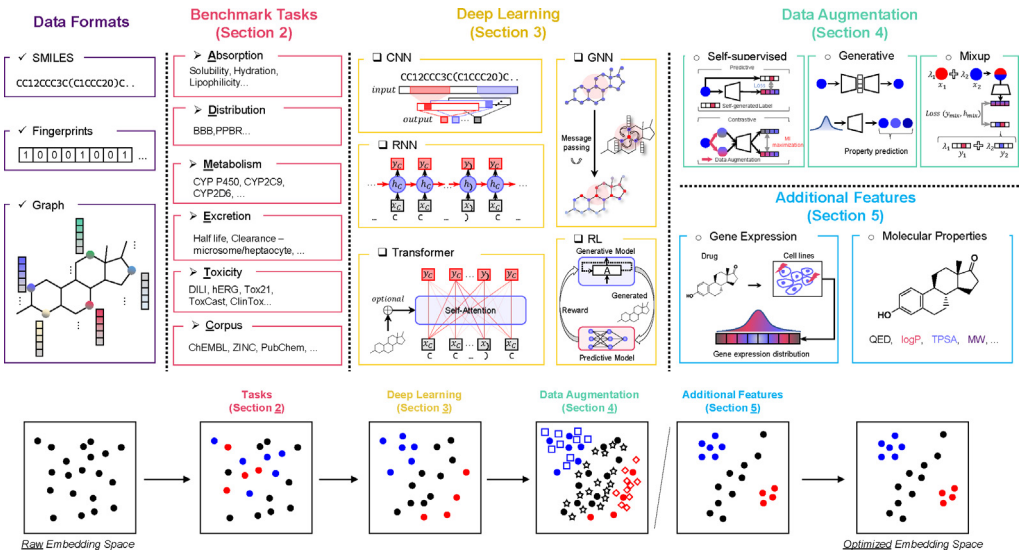

Fig. 1. Overview of the present review on building a chemical space using deep learning methods. In Section 2, benchmark tasks on drug discovery are introduced. In Section 3, several deep learning methods are introduced with selected state-of-the-art approaches. In Section 4, discussions on how to build an improved representation space are made in terms of self-supervised, generative, and mixup methods. Finally, in Section 5, features are introduced that can provide additional information other than structural formats such as gene expression or physico-chemical properties.

Table 1
Chemical tasks in drug discovery. The data imported from [12]. (Binary: Binary classification, Reg: Regression)

Task	Dataset	Size	ML Type	Reference
Absorption	Caco-2 (Cell Effective Permeability)	910	Reg	[17]
	HIA (Human Intestinal Absorption)	578	Binary	[18]
	Pgp (P-glycoprotein) inhibition	1,218	Binary	[19]
	Bioavailability	640	Binary	[20]
	Lipophilicity	4,200	Reg	[11]
	Solubility	9,982	Reg	[21]
	Hydration Free Energy	642	Reg	[11,22]
	Subtotal	16,558		
Distribution	BBBP (Blood–Brain Barrier Permeability)	1,975	Binary	[11,23]
	PPBR (Plasma Protein Binding Rate)	1,797	Reg	[24]
	VDss (Volume of Distribution at steady state)	1,130	Reg	[25]
	Subtotal	4,678		
Metabolism	CYP P450 - 2C19 Inhibition)	12,665	Binary	[26]
	CYP P450 - (2D6 Inhibition)	13,130	Binary	[26]
	CYP P450 - (3A4 Inhibition)	12,328	Binary	[26]
	CYP P450 - (1A2 Inhibition)	12,579	Binary	[26]
	CYP P450 - (2C9 Inhibition)	12,092	Binary	[26]
	CYP2C9 Substrate	666	Binary	[27,28]
	CYP2D6 Substrate	664	Binary	[27,28]
	CYP3A4 Substrate	667	Binary	[27,28]
	Subtotal	16,877		
Excretion	Half Life	667	Reg	[29]
	Clearance (microsome)	1,102	Reg	[24,30]
	Clearance (hepatocyte)	1,020	Reg	[24,30]
	Subtotal	1,592		
Toxicity	LD50	7,385	Reg	[31]
	hERG blockers	648	Binary	[32]
	hERG Central	306,893	Binary/Reg	[33]
	Ames Mutagenicity	7,255	Binary	[34]
	DILI (Drug-Induced Liver Injury)	475	Binary	[35]
	Skin reaction	404	Binary	[36]
	Carcinogens	278	Binary	[28,37]
	Tox21	7,831	Binary	[38]
	ToxCast	8,576	Binary	[39]
	ClinTox	1,484	Binary	[40]
	Subtotal	327,133		
	Total	349,036		

rately labeled databases with qualitative and quantitative information for a spectrum of chemical properties. There are several benchmark databases that have been developed for these purposes. MoleculeNet [11] is a representative database used to develop machine learning models on chemical data. MoleculeNet is a database organized in four different categories such as physiology, biophysics, physical chemistry and quantum mechanics. As the database is constructed to help develop molecular machine learning methods, it has been widely used by state-of-the-art deep learning methods as standard evaluation criteria (See Table S1). Therapeutics Data Commons (TDC) is recently released to focus more on drug discovery tasks for small molecules, peptides and other biological entities [12]. TDC re-organizes small molecule benchmark tasks mainly into ADMET categories for more task-oriented computational method development.

In this section, we focus on ADMET prediction tasks, a set of important criteria to be coordinately optimized for determining the efficacy and selectivity of drugs because they are still considered as major hurdles in drug discovery [13–16]. Table 1 summarizes ADMET benchmark datasets widely used in computer-aided drug discovery. Both absorption and toxicity datasets have been primarily used as benchmark data. There are several reasons for the popular use of these specific datasets. First, as data labels are experimentally determined, the maturation of benchmark dataset is in a close line with that of the experimental techniques. Second, the scope and the amount of data points highly depend on the availability to the public. Third, tasks such as solubility are chemical properties that can be determined straightforward by first-principles knowledge.

2.1. Absorption

Drug absorption tasks are about how effectively a drug engages into the human biological system. Among the datasets in Table 1, lipophilicity [11], hydration free energy [22] and solubility [21] are widely-used. In general, it is recommended to decrease lipophilicity and increase solubility because higher lipophilicity often leads to higher rate of metabolism, poor solubility, higher turn-over, and absorption [41]. However, poor water solubility could lead to slow drug absorption, inadequate bio-availability and induce toxicity [42,43]. Although other datasets like HIA [18] or Pgp [19] are also well established to investigate gastrointestinal or intestinal absorption [44,45], they are not as commonly used as lipophilicity or solubility datasets.

2.2. Distribution

Drug distribution tasks deal with how effectively an absorbed drug can be delivered to desired targets. Blood–brain barrier permeability (BBBP) dataset contains binary labels whether a drug penetrates the brain barrier [23]. Because the brain barrier blocks most foreign molecules, drugs targeting the central nervous system should be permeable to this barrier [46]. Plasma protein binding rate (PPBR) [47] is a regression task of predicting binding rates of drugs to plasma proteins like Albumin. In general, more weakly bound drugs more effectively traverse to the site of actions [48]. However, the two tasks are barely used in CADD because mechanistically, they are not determined by chemical structure only. Both blood barrier penetration and plasma protein binding are related to secondary biological mechanisms - adsorptive-mediated transcytosis, and binding with Albumin, respectively. For BBBP dataset as listed in Table S1, most of recent studies utilized graph neural network (GNN) or Transformer architectures.

2.3. Metabolism

Drug metabolism tasks assess whether a drug is efficiently metabolized to show desired efficacy without adverse side-effects. Predicting whether a drug inhibits or reacts with proteins in CYP 450 systems is a representative task [26–28]. Because the CYP 450 enzymes play crucial roles in the breakdown of xenobiotics, a drug that inhibits these enzymes would cause decreased metabolistic potential, which ultimately leads to drug-drug interactions and adverse effects [49–51]. Drug metabolism datasets have gained minor attention because an interaction between CYP proteins and chemical requires CYP protein structures. Moreover, even if a drug candidate inhibits a specific CYP protein, further biological network analysis is crucial to determine whether the drug causes adverse effects, which is still an open problem [52]. Nevertheless, recent studies addressed CYP datasets by developing a deep featurization strategy to overcome the drawbacks of molecular fingerprints [53,54]. The key to improvement in such methods was using multi-task learning framework to leverage structural diversity from other tasks in prediction of each of the tasks.

2.4. Excretion

Drug excretion tasks are about the rate at which an active drug is removed from the body. Half life is the dataset of measured duration for the concentration of the drug in the body to be reduced by half [29,55]. Drug clearance is defined as the volume of plasma cleared of a drug over a specified time period [24,30,56]. Although pharmacokinetics of drugs is crucial for determining the dosage of a drug, excretion dataset is not widely used in CADD because *in vivo measurement of half life or drug clearance is time-consuming and expensive [57].*

2.5. Toxicity

Toxicity tasks are to predict potential toxicities of drugs in humans. Toxicity is one of the primary causes of compound attrition, early and accurate prediction of toxicity can significantly accelerate the drug discovery and boost the likelihood of being marketed [58]. As toxicity covers extensive area of biological toxicity from heart toxicity (hERG), liver toxicity (DILI), to carcinogenesis, consortium-level efforts to characterize human toxicity experimentally are launched: Tox21 [38], ToxCast [39], and ClinTox [40]. They contain an extensive amount of data compared to others - 7,831, 8,576, and 1,484 compounds, respectively. For Tox21 dataset as listed in Table S1, many of recent studies utilized GNN or Transformer architecture.

2.6. Tasks of Generating Novel Compounds

The goal of generative models is to derive a previously unknown, synthesizable compound with desired chemical properties by utilizing the prior knowledge from a large-scale chemical database such as ZINC or ChEMBL (Section 3.5). Tasks in generative models are to generate list of chemical compounds suitable for experimental validation [59].

There are benchmark platforms for molecule generation tasks, such as MOSES [60] or GuacaMol [61]. These platforms suggest quantitative metrics to assess the performance. For example, basic metrics include validity, uniqueness, and diversity to compare statistics of the chemical distribution between generated and existing compounds. Molecular property statistics, such as partition coefficient, drug-likeness, and synthetic accessibility, are also used to evaluate performance. Some models report pharmacochemical filter scores (Glaxo [62], SureChEMBL [63], or PAINS [64], for example) which are the ratios of valid molecules without

Table 2
Chemical tasks in Toxicity. We report ROC-AUC scores for Tox21 dataset. For acronyms used in "Data" column, 'S' refers to smiles string, 'G' refers to molecular graph, 'F' refers to molecular fingerprint and 'P' refers to molecular properties. In "Model", the results of the methods from 'M Model' are reported from MoleculeNet; 'T' refers to transformer-based methods; 'G' refers to graph-based methods; 'R' refers to RNN-based methods; 'C' refers to CNN-based methods; 'S' refers to shallow embedding methods.

Type	Name	Performance	Data	Model	Year	Ref
Machine Learning in [11]	Logistic Regression	0.781	S	M	2018	[11]
	IRV[a]	0.796	S			
	XGBoost	0.815	S			
Deep Learning in [11]	Weave	0.807	G	M	2018	[11]
	TextCNN	0.838	S			
	GraphConv	0.850	G			
Deep Learning	ChemBERTa	0.728	S	T	2020	[95]
	MICRO-GRAPH	0.770	G	T	2020	[96]
	MoCL	0.780	G	G	2021	[97]
	MolCLR	0.789	G	T	2021	[98]
	SMILES2Vec	0.810	S	RC	2018	[99]
	KCL	0.813	G	T	2021	[100]
	Transformer-CNN	0.820	S	CT	2020	[101]
	DMPNN	0.850	G	G	2019	[102]
	PotentialNet	0.857	G	G	2018	[103]
	FraGAT	0.860	G	G	2021	[104]
	TrimNet	0.860	G	G	2021	[105]
	Mol2Context-Vec	0.860	FP	R	2020	[106]
	CMPNN	0.860	G	G	2020	[107]
	MPAD	0.860	SG	SG	2020	[108]
	GAP	0.880	G	G	2019	[109]
	FP2Vec	0.880	F	CT	2019	[110]
	SA-MTL	0.900	SP	CT	2021	[111]
	TOP	0.950	SP	R	2020	[112]

[a] (Influence Relevance Vector)

toxic or reactive functional groups to total generated molecules. Recently, from a multi-modality point of view, a three-dimensional (3D) molecular design task that takes 3D inter-atomic distance information into account is also being tackled. Models for this task use a quantum mechanics dataset, such as QM9 [65,66], that contains geometries minimal in energy, harmonic frequencies, energies, and so on. The performance for this task is usually assessed by aforementioned basic metrics and chemical stability.

In Zhavoronkov et al. [67], deriving a potent lead compound for DDR1 kinase inhibition was completed within 46 days by developing and utilizing a generative deep learning framework that creates a chemical space with the ZINC clean leads dataset [68] of 4,591,276 molecules and then models the properties of both known DDR1 and common kinase inhibitors (References for datasets in Table S1 of [67]). In Merk et al. [69], two million compounds from ChEMBL22 [70] were used for pre-training by LSTM to generate lead compounds optimized for RXR and PPAR agonists using RXR [71] PPAR [72] datasets.

2.7. Bioactivity and Other Benchmark Tasks

Using diverse datasets from different perspectives, we can evaluate the generalizability of ML models. As such, there are benchmarks other than ADMET closely related to drug discovery. Examples of such bioactivity datasets are BACE, SIDER, MUV and HIV. MUV dataset [11,73] is a subset of PubChem BioAssay [74] that consists of 17 target tasks over 90 thousand chemicals. The aim of this dataset is to validate virtual screening methods. HIV dataset [11,75] was created by the Drug Therapeutics Program (DTP) AIDS Antiviral Screen, to test over 40 k molecules potential to inhibit HIV replication. This dataset is widely used as for recently developed GNN and Transformer models. BACE dataset [11] is a collection of 1,522 compounds with binding results of human beta-secretase 1.

Another major stream of benchmark task is quantum mechanical property prediction. Since molecular properties and chemical reactions are determined by electron configurations and their changes, predicting quantum mechanical properties to describe electronic states is important in drug discovery tasks. *Ab initio* quantum calculation tools such as AMBER [76], Avogadro [77] or CHARMM [78] can provide reliable approximations on modeling molecules, but they require a large amount of time to compute molecule models, which is not suitable for screening many drug candidates. QM7/7b [79], QM8 and QM9 [65] predict quantum mechanical properties of molecules from its three-dimensional structures. The most representative task, QM9, utilizes 134 k stable small organic molecules which are made up of CHONF while possessing up to nine heavy atoms (CONF). QM9 produces the most stable geometry of molecules and 12 quantum mechanical properties corresponding to the conformer, such as harmonic frequencies, dipole moments and polarizabilities. There also exist other types of quantum mechanical tasks including ANI-1 (Accurate NeurAl networK engINe for Molecular Energies - 1) task for potential surface prediction [80] and MD17 (Molecular Dynamics 17) for predicting energy-conserving forces [81,82]. These tools can be helpful for investigating stable compound conformers that are important to dock with target proteins [83].

Many computational tools have developed for lead optimization [84], hit discovery [85], and docking simulation [86]. However, these tools do not use deep learning technologies and we just refer to major survey papers here.

For selected ADMET benchmark datasets, the performance of the selected deep learning methods is summarized and displayed in Table 2 and Supplementary Tables, collecting results from the literature. The details of the selected methods will be discussed in Section 3).

3. Deep Learning Technologies: How Well Can We Accomplish the Tasks with Chemical Information?

Deep learning models take chemical information in various formats. First, string formatted representations include the SMILES [87], SMARTS [88] and SELFIES [89]. Composition of substructures like functional groups is represented as chemical fingerprints, a form of binary vector based on the existence of specific chemical structural features (e.g. number of aromatic rings). According to

the recent review [90], molecular fingerprints can be divided into three different categories: substructure keys-based (MACCS, PubChem, BCI, and TGD), topological or path-based fingerprints (Daylight or OpenEye's Tree), and circular fingerprints (Molprint2D, ECFP, FCFP). Among many fingerprints, three fingerprints are mainly used in machine learning: PubChem [79], Morgan [91], and MACCS keys [92]. Since compounds consist of atoms and covalent bonds; recently chemical graph representations have been utilized as input to graph deep learning methods (See Section 3.4).

Once information of a compound is provided, machine learning models identify important features of the compound in the context of tasks to be accomplished. Often multiple features need to be combined to model compound information in task specific ways. Traditional feature-based machine learning models have limited success in capturing complex relationship of multiple features. Deep learning methods have ability to learn complex relationship directly from data, although many deep learning models are criticized for being blackbox models. Thus, there have recently been many successful examples of deep learning models to accomplish tasks in drug discovery more accurately. Attention-based models, recent developments in deep learning technologies, can be used to overcome the blackbox nature of deep learning models. For example, in toxicity prediction tasks, the presence or absence of a toxicophore in a chemical compound is important for its toxicity label [93,94]. Thus, Convolutional Neural Network (CNN)-based models focus on learning local patterns for toxicity tasks and other models that are based on GNNs or transformers exploit spatial information to learn features related to toxicophore. On the other hand, in the case of solubility prediction, the dipole moment given the degree of non-uniform distributions of positive and negative charges of a molecule is one of the important factors. Thus, multiple deep learning methods are being tried in various ways to learn the entire structural representation of a molecule.

We summarize the performance of various ML and DL methods on selected benchmark tasks (Table 2) and Supplementary Tables which deep learning methods are used for which tasks in Supplementary Table S1.

Different architectures can capture different views about chemical compounds: (1) a sequence view and (2) a graphical view. When a chemical is considered as a sequence, CNN and Recurrent Neural Network (RNN) can capture the local sequence patterns of chemical strings and Transformer considers all pairwise interactions between chemical string elements to embed valid chemical representations. On the other hand, GNN is a well-suited architecture for learning molecules with a graph view, using the a priori topology of the molecular graph to transfer information between nodes and summarize the graph-level representation of molecules. Moreover, to explore more unknown chemical substances with effective representations, Reinforcement Learning (RL) navigates the huge chemical space and generates new representations by learning effective search strategies, which can reflect the properties of the chemical substances in a specific task. In this section, we review each of deep learning technologies in the context of tasks in drug discovery.

3.1. Convolutional Neural Networks

Convolutional Neural Networks (CNN) is a method to capture local information within a specific window of data by either average pooling or maximum pooling to produce reduced representation of feature map. This strategy naturally provides regularization on conventional artificial neural networks (ANNs) and, additionally, the ability to learn hierarchical pattern of given data. CNNs are widely used in image recognition [113–117] due to their excellent ability to learn important features from image, which increases learning efficiency over ANNs.

For molecule representation learning using CNNs, the input molecule is considered as SMILES-like string. A molecule can be represented by a matrix $X \in \mathrm{R}^{n \times d}$, where n denotes the number of elements on the SMILES, and d denotes dimension of the elements. To learn the sequential pattern on SMILES, most CNN-based approaches use 1d CNNs, which is different from 2d CNNs that are used for learning 2d patterns from image data. Specifically, 1d CNNs slide convolutional filter on the X to learn local patterns of SMILES by convolutional operation and extract the effective representations by pooling operation. CNNs have been widely used in drug discovery [118–122], and used to discover patterns related to their properties [110,118].

As CNN kernels are designed to capture localized patterns in SMILES input strings and aggregate the patterns into the final prediction, CNN is favored by tasks for applications where substructures contribute to the molecule-level properties such as solubility, hydration free energy, and lipophilicity. CNN-based methods take input compounds as SMILES, some in combination with RNN or Transformers, to tackle absorption and toxicity benchmark tasks (Table S1). A recently developed method, SA-MTL [123], achieved high performance in BBBP and Tox21 datasets at Area Under the Receiver Operating Characteristic Curve (AUC-ROC) of 0.950 and 0.900, respectively, while outperformed other methods such as SMILES Transformer [124] (AUC-ROC: 0.802), Transformer-CNN [101] (AUC-ROC: 0.82), or BiLSTM-SA [125] (AUC-ROC: 0.842). The authors demonstrated in the ablation study that the performance margin of 0.102 was contributed the most by using self-attention layer. They also suggested for highly imbalanced dataset using a max-pooling layer rather than a discrete output layer in a simple CNN model. FP2VEC [110] achieved AUC-ROC of 0.880 in Tox21 dataset by employing a multi-task learning framework. ConvS2S [126] improved the performances in various datasets including solubility, BBBP, and HIV datasets by suggesting SMILES augmentation scheme.

3.2. Recurrent Neural Networks

Recurrent Neural Network (RNN) is useful for learning relational data or capturing sequential/temporal information because output from the previous state is fed into the current state. Similar to CNNs, RNNs have also been widely used to investigate sequentially formalized molecular representations (e.g., SMILES) [106,127]. Compared to CNN (See Section 3.1), RNN and its variants (e.g., LSTMs or GRUs) can capture long-range relationships among chemical elements in SMILES due to their innate recurrent memory mechanism. However, RNNs are not suitable to capture localized patterns such as functional motifs. Therefore, RNNs and CNNs are often used together to complement each other [99,128,129] in chemical domains.

Depending on how to handle SMILES-like representation, existing works can be grouped in three categories such as atom-level sentence, substructure-level sentence, and SMILES augmentation. A SMILES string can be naturally considered as sentences of atoms with auxiliary symbols (e.g., double bond, branch, or ring). *Atom-level sentence* [99,127] considers each symbol as input features of the model. Though this approach can directly use RNN or CNN architectures, it overlooks the fact that multiple atoms and bonds form one functional group. To overcome this limitation, *substructure-level sentence* [106,110] considers a compound as a sentence of the substructures of the compound. The substructures can be obtained by using any chemical fingerprints [91,130] or fragmentation algorithm [131]. A SMILES string is just one of many possible views on a chemical compound and it is possible to use multiple SMILES representations of a compound, called *SMILES augmentation* [35,128,129,132].

RNN based methods have been widely used for tasks in absorption category and Tox21 datasets (Table S1). TOP [112] achieved an excellent performance in toxicity prediction using Tox21 dataset (mean AUC-ROC: 0.950) by integrating shallow representation on SMILES into biGRU in combination with some molecular descriptors (logP, MW and TPSA). By incorporating the physiochemical properties, TOP resulted in 0.195 performance gain in terms of AUC-ROC. Li et al. [128] achieved comparable performance to existing methods [99,129] in the solubility task. Meanwhile, Mol2Context-vec [106] outperformed other RNN based methods in most benchmark tasks (solubility, lipophilicity, BBBP, and BACE). The authors suggest that learning molecular descriptors such as logP or TPSA solely from SMILES is difficult. Thus, additionally providing such features contributed to the performance improvement.

3.3. Transformer

Transformer performs sequence-to-sequence translation tasks in an encoder-decoder framework. Self-attention mechanism [133] is a core component of the transformer architecture that uses all token pairs to encode contextual information to learn global representations of sequences. Because of this modeling power, Transfomer has been successfully used in the field of natural language processing [134,135]. In computer vision domain, a vision Transformer [136,137] achieved outstanding performance for various machine vision problems.

Following the great success of transformer in computer vision and natural language processing domains, several transformer based models have been proposed for efficient chemical representations. Leveraging the capability of transformer as an encoder, it is usually pre-trained on massive unlabeled chemical compounds either in the form of SMILES or molecular graph, which leads to outstanding performances in downstream tasks such as absorption, distribution, and toxicity prediction [101,123,138–140]. The crucial point of the chemical transformer is fully exploit atom interactions and chemical structure information through self-attention mechanisms.

SMILES-BERT [139] and ChemBERTa [95] embedded chemical SMILES based on transformer or BERT (Bidirectional Encoder Representations from Transformers) [141] to pre-train the semi-supervised learning model on unlabeled large scaled data, where long range atom-level interactions can be learned. Task-specific models by fine tuning the pre-trained model using additional data for downstream tasks improved prediction performance for a number of tasks. For example, SMILES-BERT improved LogP prediction performance by about 2% compared to existing SOTA Seq2Seq-based methods which indicates better utilize the unsupervised information with the Masked SMILES Recovery task, and gets more than 5% and 8% improvement on PM2 and PCBA tasks, respectively.

Transformer-CNN [101] and SA-MTL [123] incorporated CNN and transformer to capture localized chemical substructures and learn interactions between substructures. SA-MTL achieved AUC-ROC of 0.900 on toxicity tasks achieving around $2 \sim 10\%$ higher than existing deep learning models in toxicity prediction and also achieved the highest performance on distribution task (e.g., BBBP dataset) and other tasks (e.g., HIV, SIDER). Transformer can learn chemical information in the form of molecular graphs as well as SMILES. MolAT [138] augmented the self-attention between atoms with chemical structural information: (1) adjacency on molecular graph and (2) inter-atomic distances. MolAT outperformed the SMILES-based models in predicting various molecular properties, such as water solubility, BBBP, and metabolic stability prediction. ST-KD [142] proposed an end-to-end SMILES transformer by knowledge distillation in a graph transformer without pre-training. Knowledge distillation can transfer knowledge from a teacher model to a student model. In ST-KD, the teacher model (graph transformer) is trained first, and then the output of the graph transformer and the real labels of the data are used to train the student model (SMILES transformer). Student model (SMILES transformer) can learn structural information of molecules since the hidden representations and attention weights of the distillation are focused on the information learned by the teacher from the molecular graphs. ST-KD showed outstanding performances on QM datasets against graph-based and SMILES-based models. It demonstrated that efficient chemical representations learned by knowledge distilled transformer can capture more global information than graph-based representations. It also indicated that global information is more appropriate for QM datasets. Meanwhile, on FreeSolv and HIV datasets, ST-KD showed competitive results with graph-based models, indicating that these tasks are more likely to focus on local graph structures.

3.4. Graph Neural Networks

Graph Neural Networks (GNNs) are most well suited for processing graph data. GNNs can be used to predict relationship among the members in social networks [143], and can inference biomarkers using biomedical networks [144,145] (e.g., protein–protein networks). Message passing operation, a core operation in GNN, is to learn feature information propagated on graphs. Specifically, GNN aggregates information from neighboring nodes $u \in \mathcal{N}(v)$ of node v and uses the information to update the representation of node v. To represent a graph G, a readout function is then used to summarize all updated node representations to an 1d vector representing the graph.

Since chemical compounds are complex 3D structures of atoms and bonds, it is natural to represent them as graphs. To represent chemical compounds in GNN, popular approaches [102,107,146,147] are to design the message passing algorithm to learn node/edge representations and aggregate them as a molecule representation. However, learning molecule representations this way often results in summarizing local proximity information. Recent studies [104,148–150] have begun to use message passing non-locally. Some works [148,149] addressed this problem by message passing on r-radius subgraphs to learn a more global representation for molecule graph. Others [104,150] explicitly extract knowledge-guided subgraphs from the molecules and make representations of them, indicating that domain knowledge can be used to reflect global molecular properties for better graph-level representations on chemical property prediction tasks.

Depending on view of chemical structures, proposed GNN based approaches can be divided into three categories: *atom/bond-level*, *subgraph-level*, or *graph-level*. Given a molecular graph, *atom/bond-level* GNNs [102,107,146,147] aggregate information on a target atom with adjacent atoms and bonds. MPNN [147] used long-range interactions on gated graph neural networks [151] for molecule prediction tasks using QM9 dataset and outperformed existing strong baselines without explicit feature engineering. In follow-up studies, CMPNN [107] used a node-edge interaction module where information can pass between node and edge. CMPNN guided model learn topological relationship among elements in molecules and outperformed baseline methods in absorption, distribution and toxicity tasks. In absorption task, CMPNN achieved Root-Mean-Squared-Error (RMSE) s of 0.23 and 0.82 for the ESOL and FreeSolv datasets, respectively. In distribution task, CMPNN outperformed and achieved AUC-ROC of 0.963 for BBBP dataset, and also achieved AUC-ROC of 0.856 and 0.933 for Tox21 and ClinTox datasets in toxicity task.

To learn better atom representations, there are also methods improve GNN in various ways: Gilmer et al. [147], Feinberg et al. [103], Yang et al. [102] used Gated Recurrent Unit (GRU) to improve long-range message propagation over the molecule graph.

Geometric distance-based methods that successfully captured chemical properties in QM9 and MD17 datasets [152,153]. Attention mechanism-based methods [105,154] focused on element-level representation learning to obtain better molecule representation so they outperformed the baseline in ADT and made some interpretable visualizations for these tasks. MT-PotentialNet [54] not only propagated message differently depending on different edge types but learned a unified featurizer for multiple tasks using multi-task learning. Empirical experimental results achieved unprecedented accuracy in ADMET property prediction tasks and revealed that potentialNet with multi task learning on ADMET dataset not only interpolated but also extrapolated to new chemical space.

Subgraph-level GNNs can overcome the limitation of atom/bond-level GNNs by explicitly utilizing subgraphs extracted from molecular structure. A subgraph can be conveniently defined as *r*-radius subgraphs [148,149] or as domain knowledge guided functional groups [104,150]. FraGAT [104] fragmented the graph by the unit of bonds. FraGAT cut all acyclic single bonds to make basic fragments, then iteratively connected them to form fragment-pair, which can regenerate original molecule after one bond-ligation reaction. FraGAT conducted experiments on 14 benchmark datasets including absorption, distribution and toxicity tasks. FraGAT outperformed baseline with RMSEs of 0.48 and 0.54 for ESOL and FreeSolv datasets in absorption task, and achieved AUC-ROC of 0.933 for BBBP dataset in distribution task. The AUC-ROCs of 0.863 and 0.969 for the Tox21 and ClinTox datasets in toxicity task also outperformed other methods. Experimental results showed that FraGAT can achieve the state-of-the-art predictive performance in most cases, especially in absorption tasks (e.g., FreeSolv) and toxicity tasks compared to AttentiveFP [155] with 0.736 RMSE for FreeSolv and 94% AUC-ROC for ClinTox dataset. All results demonstrated the effectiveness of using three-level hierarchical structural information of molecules in FraGAT.

Similar to *subgraph-level*, *graph-level* GNNs are to model global structure of chemicals. For example, MGCN [156] achieved the best performance in predicting 11 out of 13 properties on QM9 dataset. This is because MGCN explicitly used hierarchical multi-level connection layers (point-wise, pair-wise, triple-wise, etc) to encode global interactions in chemical graph.

3.5. Reinforcement Learning

Reinforcement learning (RL) is a computational framework where trained agents or machine learning models make a sequence of decisions in an environment to achieve a specific goal. When the agent performs a certain action in its current state, the environment gives a computed reward to the agent for that action. RL learns policies to make the optimal choices that maximize total reward by exploring possible states and actions in a trial and error manner. Because of this learning framework of RL, it is widely used in various domains such as the game of GO [157], autonomous driving [158], and protein structure prediction [159].

Using RNN as an agent architecture, Olivecrona et al. [160] first adopted RL framework for generating SMILES string of molecules which are valid and bind to dopamine receptor type 2 (DRD2) receptor. Starting from a single atom SMILES string, the agent recursively determines the element and the direction of the atom to be attached; the reward is given whether the current molecular structure is active against DRD2, is chemically valid, and avoids sulfur. Iterative additions of atoms by the agent expands the scope of the chemical space, while the reward is navigating the agent not to be distracted to inactive or invalid chemical space. Ståhl et al. [161] improved the RL framework by incorporating substructural information and chemical properties such as Tanimoto structural similarity [162] and Levenshtein string distance [163]. However, due to the inherent linearity of SMILES, it is difficult to satisfy the Markov decision process (MDP) assumption in SMILES-based RL approaches. By defining a state as a form of graph, GCPN [164] fully overcame an invalid MDP assumption problem; because graphs can better capture molecular topology, chemical rules can be better reflected in transition dynamics. Although GCPN performed better than the SMILES-based RL models, exploration of chemical space is limited using a predefined set of scaffold subgraphs. By defining atom/bond-level actions, MolDQN [165] can explore non-restricted vast chemical space without scaffold subgraphs.

In general, because the reward should be calculated from arbitrary molecules which the model generates, simple molecular properties such as logP and QED is often used. A recently published MoleGuLAR [166] used multi-objective scheme for generating drug like molecules with high binding affinity to novel targets along with desired logP: −6.76 kcal/mol mean binding affinity, 2.9 mean logP, 0.42 mean QED for targeting 2.5 logP and 1 QED, respectively. The authors described that their switching reward functions rather than the sum of rewards improved optimization quality because an alternating reward takes the model into the better local chemical space where one property is optimal when optimization for the conflicting property is started. REACTOR [167] generated average 77 actives for DRD2 objectives while following chemical reactions for generating 'synthesizable' derivatives - it outperformed previous like MolDQN [165] or JTVAE [168], which added additional synthesizability term as a negative reward and generated 9.667 and 4.0 actives, respectively.

4. Data Augmentation: How to Extend Our Knowledge on Chemical Space?

There are a number of databases that collect chemical compounds with various experimental results. Two main datasets are ChEMBL, ZINC in drug discovery. ChEMBL [24] is a database of molecules with bioactivity data. It contains 2.1 million compound information in the latest update. ZINC [169] is a database of commercially available molecules. It contains 1.3 billion compound information in v20. Although the number of compounds in existing databases seems large, they are quite small in the whole chemical space. Thus, we need more powerful computational techniques to explore chemical space. There are extensive reviews on building a chemical space by using a corpus of unlabeled chemical data for pre-training of deep learning architectures [170–172]. The main point in these deep learning-based methods for chemical space exploration is *embedding* where chemical compounds are represented in the alternative form of embedding vectors. Embedding is the mapping of high-dimensional feature vectors from the raw data space into a relatively low-dimensional space. To learn complex interactions between the raw features, deep learning models try to embed data using various architectures such as CNNs, RNNs, or GNNs. Because they learn potential characteristics of the data, embedding spaces have two advantages:

(1) It is easier to compute similarities between data points or to identify common features that are difficult to capture in the raw data space.
(2) The data is transformed into vectors regardless of its original format, allowing machine learning and deep learning models utilize the vector representations.

A question arises here: How can we make a more effective embedding space given chemical tasks? More effective chemical embeddings should have good generalizability potential that allows the structural properties of compounds to be tailored for

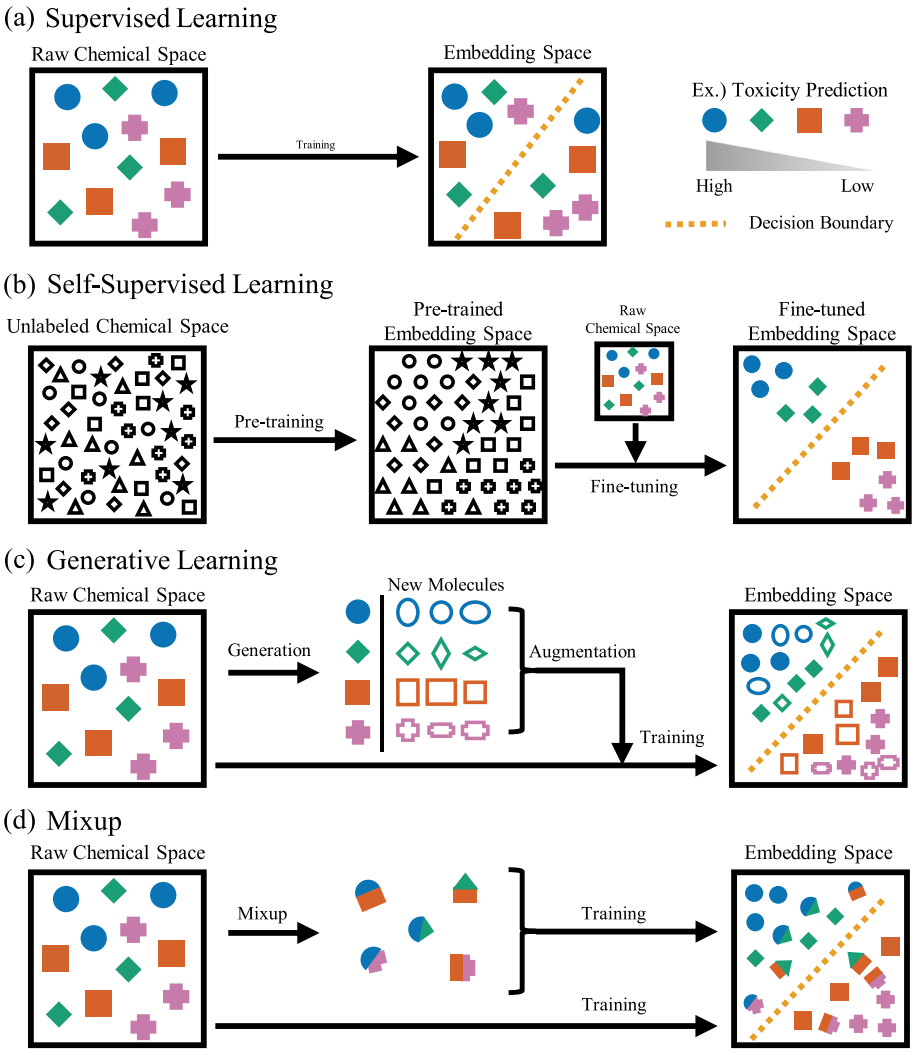

Fig. 2. Approaches to address lack of data.

given benchmark tasks [173,174]. In particular, most benchmark chemical datasets have a small number of labeled samples (e.g. 475 and 642 drugs in DILI and hydration free energy datasets, respectively), resulting in insufficient structural diversity only with the labeled samples. Thus, deep learning models employ various self-supervised learning and generative strategies into their framework to increase the structural diversity. In other words, deep learning is a search algorithm that makes exploration from the raw chemical space to an embedding space with desired properties. The purpose of the search algorithm is to explore from the start state to the goal state through intermediate states by transitions, where the raw and desired chemical spaces are a start and goal states, respectively. Here, we defined 'transition' as an optimized procedure by an objective function of deep learning models.

A major obstacle to make efficient transitions is the lack of labeled data (Fig. 2 (a)). Various attempts have been made to address this problem, and this section focuses on *Data Augmentation*. Data augmentation is to provide additional data to deep learning models to help guide searches. If data becomes sufficient after data augmentation, the deep learning model can better achieve the goal state because the objective function can consider various aspects of the current embedding space. The main issue is how can we provide more data when the labeled data is not enough. To overcome the data insufficiency issue, there are three widely used approaches: (1) self-supervised learning, (2) generative learning, and (3) mixup.

All three methods are similar in that they provide additional data to the deep learning model, but each method has its own characteristics.

(1) self-supervised learning (SSL), as the name suggests, is a method of generating labels from the data itself, so it can take advantage of many unlabeled chemical compounds. Recently, SSL has been spotlighted in various fields such as computer vision, natural language processing, and graph learning. It is a good technique for fine-tuning with small labeled data after constructing a uniform and distinguishable embedding space from a large number of unlabeled data (Fig. 2 (b)).
(2) Generative learning is a method of creating new chemicals with similar properties from known chemicals (Fig. 2 (c)). Generative models such as generative adversarial network (GAN) and variational autoendocer (VAE) can be used, and data in various formats such as SMILES and graph can be generated and used for learning downstream tasks.
(3) Mixup is a technique mainly used in supervised learning that mixes up two or more data representations and label information to create new data. This has the effect of interpolating the space in the sense of filling the space between the labeled data (Fig. 2 (d)).

Table 3
Graph learning methods for building a chemical space.

Data level	Year	Self-supervied Framework	
		predictive	contrastive
Node/Edge-level	~ 2019	EdgePred AttrMasking ContextPred N-Gram-Graph	Infomax
	2020	GPT-GNN	GRACE InfoGraph GMI
	2021	MGSSL MolGNet	GraphCL JOAO MolCLR MoCL
	2022		
Subgraph-level	~ 2019		
	2020	Grover	GCC
	2021	MGSSL MolGNet	GraphCL JOAO GraphLoG Sugar MolCLR MoCL MICRO-Graph MolCLE
	2022		
Graph-level	~ 2019		
	2020		InfoGraph
	2021		GraphLoG MoCL
	2022	D-SLA	KCL

4.1. Applications of Self-supervised Learning

Self-supervised learning (SSL) tries to learn the structural diversity and general semantics of unlabeled data to create an embedding space that can be used as an initial value in the process of fine-tuning. In particular, it is used to build embedding spaces to learn the semantic information of compounds in various computational forms such as SMILES or molecular graphs.

4.1.1. SSL with SMILES

SMILES-BERT [139] and ChemBERTa [95] utilized BERT (or transformer) architecture that is widely in the field of NLP for text data SSL for its outstanding performance. The two BERT models used ZINC [68] or PubChem [175] database as a pre-trained dataset, by masking a portion of tokens in each SMILES, the pre-training procedure predicts the masked symbols, which is to learn hidden semantics of the SMILES representation. The space created by the pre-trained BERT encoder partially reflects to the semantic information of the chemical compounds. Thus, the space serves as a useful intermediate state, and even with a small amount of label data, an informative goal state, i.e., the embedding space, can be constructed through a fine-tuning process.

4.1.2. SSL with Molecular Graph

Self-supervised learning using graphs has become recently more prevalent for chemical embeddings. We divide graph SSL into *predictive* and *contrastive* according to the loss function. *Predictive* methods generate labels related to data, and predict them using cross-entropy loss. *Contrastive* methods use InfoNCE or NT-Xent loss to determine the distance of positive and negative samples in the embedding space. Due to the nature of the loss functions, the predictive methods guide search towards constructing a more distinguishable embedding space and, in contrast, the contrastive methods construct a uniform embedding space.

As examples of the predictive methods, [176] developed node attributes and context prediction tasks, called AttrMasking and ContextPred, respectively. These methods train an encoder that predicts the attribute and structural information of the graph to determine efficient intermediate states. Similarly, N-Gram-Graph [177] utilized the word2vec scheme to estimate node attributes. GPT-GNN [178] focused on node attributes prediction and global topology prediction tasks. On the other hand, the above methods are based on "graph-driven labels" using the structural characteristics of the graph. There are also methods of generating "knowledge-guided labels" based on the characteristics of the molecular graph. Grover [179] generated substructure-based labels based on the type and number of atoms/bonds in the k-hop neighborhood. MGSSL [180] performed pre-training as motif-level graph generation process, and it expects to build a more suitable chemical space for molecular graphs using functional motif-based subgraph information and generation process about the overall structure of the graph. MolGNet [181] and Kim et al. [182] tried to avoid negative transfer by designing a chemical space that can learn chemical validity to reflect chemical stability rather than specific properties.

As examples of the contrastive methods, graph topology-based approaches perform contrastive scheme by same- or cross-view of the node/edge-, subgraph- or graph-level comparison (Table 3). For example, GCC [183] performed subgraph-to-subgraph contrast, which focuses on generalization of chemical space in terms of chemical subgraphs so that the latent representation well reflect molecular properties arising from functional groups. SUGAR [184] performed subgraph-to-graph contrast, which is to explore the interpretability and semantic connections between substructures and molecular graphs. In addition, graph-to-graph contrast methods [185–187] tried to learn semantics between the augmented graphs in the given dataset. On the other hand, MolCLR [98], MoCL [97], KCL [100], and MICRO-GRAPH [96] leveraged multi-level chemical knowledge where atoms, bond, subgraphs, or graphs can pose in developing chemical properties. The key to success of these methods is to focus on the semantic information shared by chemical graphs by knowledge guided augmentation that excavate meaningful subgraphs as well as embeddings.

The key advantage of SSL is that downstream deep learning models can learn as more diversified molecular structures as possible even without provided labels. Thus, most graph-based pre-training methods demonstrated how much contribution their pre-training strategy contributed to downstream tasks. A seminal work introducing 'ContextPred' by Hu et al. [176] in most tasks achieved an average of 7.2% mean improvement in eight benchmark data sets by pre-training on ZINC15 database compared to the non-pretrained vanilla GIN model. Recent chemistry-inspired methods MolGNet [181] outperformed existing tools (mostly GNN methods) on both the classification tasks (SIDER, ClinTox, BACE, Tox21, and ToxCast) and the regression tasks (solubility, hydration free energy, and lipophilicity). MGSSL [180] also demonstrated the usefulness of chemistry guided pre-training strategy on the set of benchmark GNN models (GCN, GIN, RGCN, DAGNN, and GraphSage) by an average margin of 7.56% on eight different benchmark data sets.

4.2. Applications of Generative Learning

4.2.1. Generation of Molecules with Desired Properties

The purpose of generative learning is to generate new, unknown data with properties similar to the given data. The goal is to learn a latent distribution from the given data and then generate similar data from the distribution. For this, we can adopt the generator/discriminator or the encoder/decoder framework with CNN, RNN, and GNN. In particular, in the field of computer vision, various structures such as GAN [188], VAE [189], and RL [190] are used as generative models, and research on chemical generation is

also in progress to address lack of data in specific tasks such as prediction of pIC_{50} or logP.

JT-VAE [168] utilized a VAE for molecule generation, which directly uses molecular graphs instead of SMILES. Given a molecular graph, it was converted into a junction tree format with a vocabulary of valid chemical substructures. Based on the junction tree, a VAE encoded the tree structure into a latent space, and decoded the input tree from the latent space. While converting the reconstructed tree into the molecular graph, JT-VAE guided the decoder using the graph embedding learned by a GNN encoder. Another work, Mol-CycleGAN [191] was a CycleGAN[192]-based model that generates new molecules with high structural similarity to the input molecules. Based on the latent space from JT-VAE, Mol-CycleGAN optimized the generator that learns desired chemical properties by discriminating two different molecule sets (e.g., active/inactive or high/low of IC_{50}).

Generative models explicitly sample modified molecules, then evaluates how much the generated molecule is optimized. Thus, GL is often used in molecular optimization tasks. The generated molecules are mostly evaluated by both synthesizability and numerical properties. A recently developed Mol-CycleGAN [191] and GCPN [164] represented VAE/GAN and RL methods, respectively. Under penalized logP optimization task of drug-like molecules with similarity constraints ($\delta \geqslant 0/0.2/0.4/0.6$), Mol-CycleGAN outperformed GCPN in the mean improvement of the property. However, in terms of the success rate, Mol-CycleGAN has lower success rates for the more stringent constraints (δ = 0.4, 0.6). Regardless of constraints, GCPN showed robust success rates and comparable improvements to Mol-CycleGAN in the stringent constraints.

One of the latest research trends is 3D structure-based molecule generation. Since the late 2010s, several models have been proposed to discover novel compounds with target properties, including QM9 dataset-based quantum mechanics information. cG-SchNet [193], a conditional generative neural network for inverse design of molecules, enabled joint targeting of multiple properties including HOMO–LUMO gap and energy. Another 3D compound generative model, MOLGYM [194] constructed an RL environment for molecule design in Cartesian coordinates. Using rewards provided through fast quantum-chemical calculations, the agent was not only able to generate 3D molecules but also placed water molecules around a compound, predicting solvation state and separating inter-atom interactions from intra-atom forces.

4.2.2. Target-specific lead identification & optimization

In a drug discovery perspective, generating a compound with desired chemical properties is important, but its interaction with a desired protein target may be more valuable information. Several generative models have been proposed to meet the needs of lead identification and optimization using deep learning framework, based on given targets.

GENTRL [67] was a deep generative model that developed potential DDR1 kinase inhibitors in 21 days. To guide the search, GENTRL utilized a VAE with a rich prior distribution in the latent space. The rich prior distribution was obtained by tensor train decomposition with chemical properties, including MCE-18, pIC_{50}, and a binary indicator of passing medicinal chemistry filters (MCFs). Then, with the trained encoder, GENTRL learned the generation process of DDR1 kinase inhibitors by using RL framework with kinase related reward functions.

MORLD [195], a docking score reward-based reinforcement learning framework, generated and optimized lead compounds fit to query protein structure without intense screening on large bioassay libraries. The authors claimed that their proposed method speed up the DDR1 kinase inhibitor discovery time from 21 days by GENTRIL down to 2 by their method.

An interesting model proposed by Méndez-Lucio et al. [196] is a two-staged GAN-based model that generates molecules that fit the desired gene expression profiles. The discriminator calculates the probability of whether a generated molecule is a valid molecule, and a conditional neural network predicts whether the molecule fits the given desired expression profile.

4.3. Applications of Mixup

In supervised learning, when data is insufficient, decision boundaries can be constructed too tightly, which leads to overfitting the training data. Mixup [197,198] performs interpolation of both input data and label information to smooth decision boundaries and infer information between the boundaries. Let x is a input data and y is a input label. Mixup of two data generates new data $x\prime$ and $y\prime : x\prime, y\prime := \mathrm{Mix}_\lambda(f(x_1), f(x_2)), \mathrm{Mix}_\lambda(y_1, y_2)$, where $\mathrm{Mix}_\lambda(a, b) = \lambda \cdot a + (1 - \lambda) \cdot b$ and λ is mixing ratio. $f(x)$ is a function that maps the input x into the latent space to interpolate. For example, $f = \mathrm{Identity}(\cdot)$ is input mixup [199], and $f = \mathrm{ENCODER}(\cdot)$ is manifold mixup [200]. Then, the deep learning model is to be trained to learn the mixed data and predict the corresponding *mixed labels*, rather than original class labels. As of now, researches on mixup are mainly in the field of images. This is because the mixing technique is suitable for grid-structured data and the labels of the interpolated data may not be smooth. For example, SMILES augmented by mixing two SMILES may exhibit neither of the two chemicals. Also, irregular graph sizes and connectivity of graph are major challenge of graph-level mixup.

Even so, some studies for graph-level mixup have been conducted recently. Wang et al. [197] suggested two mixup schemes for node- and graph-level classification. Node-level mixup scheme consists of two-branch graph convolution and two-stage mixup framework for considering receptive field of nodes and preventing unintended mixed representations. Graph-level mixup is performed in the embedding space of graph representations, which is equivalent to the manifold mixup. Graph Transplant [198] also suggested graph-level mixup based on meaningful subgraphs related to labels. To obtain the informative subgraphs, it utilizes node saliency information and adaptively determines the labels.

5. Additional Features Required Beyond Chemical Compound Information

In addition to feature engineering by deep learning methods from molecular structures, features that cannot be directly inferred from the molecular structures can provide structurally diversified chemical information make tasks-specific chemical spaces [201–203]. In this section, we discuss two major features, molecular descriptors and pharmacogenomics profiles. First, chemical heuristics are arranged as a general feature set. While ECFP fingerprints solely produce subgraph-based binary information, PubChem fingerprints, a 881-long bit vector, include rules such as the number of aromatic rings, or the number of unsaturated bonds [79]. Second, recent advances in gene expression measurement technologies, especially next generation sequencing, can provide multi-level and extensive features in addition to chemical features [204]. For example, in terms of pharmacogenomics, RNA sequencing measures more than 20,000 genes in human samples.

Molecular Descriptors. logP (partition coefficient) and TPSA (total polar surface area) are related to the solubility of a compound in aqueous solutions and the presence of specific structural features such as the number of rings is related to carcinogenesis [205]. TOP [112] leveraged logP, molecular weight, and TPSA to the independent fully connected layers on the word embeddings of SMILES along with biGRU to learn chemical structures. Addition of

physico-chemical properties selected by genetic algorithm to biGRU featurization provided 0.195 of improvement in AUC-ROC to toxicity prediction tasks. Tharwat et al. [206] used 31 molecular descriptors in prediction of four toxicity tasks in combination with multiple data sampling strategies to build an ensemble learning framework. In this way, they achieved the best performance on the four different toxicity tasks by an entropy-based feature selection method [207]. Leveraging chemical fingerprints also provided additional improvement to the text-based modeling of chemical compounds in predictions of solubility, hydration free energy and lipophilicity [106].

Gene Expression Profiles. Gene expression profile is also an important information to featurize chemical structures as the metabolic dynamics of the drugs in biological systems can be inferred from the changes in gene expression profiles [51,208,209]. In pharmacogenomics, perturbations in gene expression can also elicit the mechanistic clue of toxicity as a responsive to drug administration [210–213]. In a recent study, a model is also being developed that uses data other than chemistry to generate a desired chemical compound. In particular, gene expression profile was used to generate a candidate small-molecule drug for cancer [196,214].

Biological Assays. The purpose of bioassay experiments on various drugs and target organisms can be used to narrow down the potential drug targets. From the computational perspective, these experiments can be considered as drug-target interaction (DTI) problem. DTI information can be explicitly fed into ML models that predict cellular responses upon drug treatment [215,216]. Besides, due to experimental costs and limitations, DTI itself is one of the most active research areas in the machine learning communities [217].

6. Discussions

In this section, we discuss current paradigms and future directions for deep learning based chemical embeddings.

6.1. Graph-based Chemical Embeddings

As shown in Table S1, most of the recent works try to utilize molecular graphs as is, as opposed to the linear representation like SMILES. Graph-based chemical embedding methods have advantages on four aspects: (1) It is natural to represent chemical compounds as graphs. Since atoms and bonds are represented by nodes and edges of the graph, interaction information of compounds can be efficiently reflected. (2) Because graphs can model local- and long-range interactions, we can model rich information from structural motif [104,150] to global proprieties of the chemicals [147,156]. (3) Depending on the goal of the task, we can learn the molecular graphs using a variety of techniques including RNN [106], GNN [153], Transformer [179], and so on. (4) Domain knowledge also can be expressed in the form of graphs [218], thus it can be used as prior knowledge of the model to construct the effective embedding space. Because of these advantages, graph-based chemical embedding is an interesting research direction, but it may need more time and efforts because intrinsic properties of compounds, such as the spatial location of atoms, are yet to be well characterized.

6.2. Exploration on Motif-level Learning

Substructures, also known as motifs, are fundamental components that determine characteristics or properties of a compound [219]. To reflect this into machine learning heuristics, as described in Section 4.1, recent studies consider molecules as a set/tree of substructures. While the previous approaches use set of k-hop neighbors (local-level) in a graph to make a graph-level representation, recent approaches put more efforts into leveraging chemical prior knowledge. Predictive models including Grover [179], MGSSL [180], and MolGNet [181] tried to learn the chemistry level semantics of a chemistry graph to effectively construct an entire graph from combinations of subgraphs. In case of contrastive methods, data augmentation strategy is a crucial issue. Rather than atom/bond-level add/deletion schemes, motif-based data augmentation makes it easier to create a valid chemical and retain the properties of a given original compound. Meanwhile, the rich information of motifs is also useful in chemical generation. As shown in Zafirlukast's example for lead compound optimization [220], fatty acid mimetics is one of the most widely used techniques for lead optimization in medicinal chemistry. Deep generative models may also support this effectiveness. JT-VAE [168] is a seminal approach to effectively reconstruct tree-like structure of molecular graphs, while successive efforts have been made by other studies to reflect chemical prior into their policy network [164]. Likewise, substructure based chemical modification by RL on molecular graphs will be helpful to find a desired chemical compound in a chemical space with guaranteed stability that comes from a structure similar to the existing fatty acid-like antagonist [221]. As such motif-level learning methods are still premature, how to investigate the power of motifs remains a challenging and attractive problem.

6.3. Pre-training for chemical space

Pre-training through self-supervised learning or generative models has been increasingly important to compensate for insufficient chemistry data for specific tasks. Building a chemical embedding space using pre-trained methods has the following advantages: (1) As pre-training methods are developed in various domains including computer vision or natural language processing, we can use various data formats, such as graphs [97,98] or SMILES strings [95,139] for pre-training. (2) It uses unlabeled data, but learns a lot of information about the chemical compounds. Structural diversity grows exponentially when learning a large amount of chemical data, even though the label information is not given. It increases the generalizability of the trained model and achieves good performance in a variety of downstream tasks. (3) Even on small labeled datasets, it is possible to train task-specific embedding spaces using fine-tuning or few-shot learning techniques [222]. However, the following disadvantages should also be kept in mind: (1) Little or no correlation between pre-trained tasks and downstream tasks can result in negative transfer that degrades model performances [176]. (2) The lack of a theoretical foundation for pre-training techniques can make it difficult to interpret which properties of the compound have been learned by the generated embedding space. Future studies will require more descriptive and robust pre-training methods to compensate for this point.

6.4. Importance of using Negative Data

The selection of negative data is an important issue. It plays a crucial role in determining decision boundaries in the embedding space. For example, to accurately and rapidly generate DDR1 kinase inhibitors, GENTRL [67] utilized molecules that act on non-kinase targets as negative data. In the mode-of-action, [223] improved in silico target prediction by utilizing negative bioactivity data held in chemogenomic repositories. For drug-target interactions, [224] proposed a systematic method to select reliable negative samples, which significantly improved the prediction accuracy for protein targets of small molecule drugs. From a technical point of view, self-supervised learning is a useful technique for building pre-trained models, but it is difficult to select useful

negative data for using unlabeled compounds. For contrastive learning, the concept of attract/repel between anchor data to positive/negative data is important, so the selection of negative data is very important. We expect that new techniques will develop and use more efficient negative data selection strategies, which results in more effectively navigation of chemical space.

6.5. Potential Risk of Overfitting

Due to the lack of sufficient data labels in chemical benchmarks, deep learning models are prone to overfitting because deep learning models use a large number of parameters [225–227]. It is not possible to avoid the overfitting issue in chemical property prediction tasks, but deep learning models have been trying to address the overfitting issue in two different perspectives: data and computational perspectives.

Computational perspective. In terms of DL models using a large number of trainable parameters, major achievement in reducing overfitting was made by stochastically dropping out the trained weights on randomly selected neurons [225] or Bayesian approaches [228,229]. To fulfill the out-of-sample generalizability of ML models on independent data, train/valid/test data splitting strategies are introduced in terms of scaffold/random/temporal features, etc [11]. Among the splits, it is reported that scaffold/temporal splits produce less bias compared to the random split data preparations [102,230]. In the meantime, TDC benchmark recommends different types of data splitting, performance measure, and modeling strategies on benchmark data sets for fair comparisons [12].

Data perspective. When it comes to HTS BioAssay data, there has been a paper (LIT-PCBA [231]) that developed an unbiased PCBA data set with reduced number of protein targets (p = 15) to avoid potential overestimation by machine learning models. In a recent paper (FP-GNN; [232]), FP-GNN was compared to ML models (NB, SVM, RF and XGBoost) and DL models (DNN, GCN and GAT) on the LIT-PCBA dataset. Given mixed fingerprints (MACCS, PubChem, and Pharmacophore ErG) as input features, the four ML models achieved on average 0.672 accuracy while the three DL models (DNN, GCN, and GAT) and FP-GNN achieved improvement in accuracy to 0.729 and 0.739, respectively. It seems that deep learning models have been evolved to improve on the overfitting issue.

7. Conclusion

In this article, we surveyed how deep learning technologies can model and utilize chemical compound information in a task-oriented way by utilizing annotated properties and assay data in the chemical compounds databases. We first compiled what kind of tasks are tried to be accomplished by machine learning methods (Section 2). Then, we surveyed deep learning technologies to show their modeling power and current applications for accomplishing drug related tasks. Section 3 surveyed deep learning techniques to address the insufficiency issue of annotated data for more effective navigation of chemical space. In Section 5, we surveyed what kind of information, such as assay and gene expression data, can be used to improve the prediction power of deep learning models. Final section surveyed four important newly developed techniques that are yet to be fully incorporated into computational analysis of chemical information.

Recently developed deep learning methods has demonstrated their ability to increase the efficiency of lead compound optimization, and various self-supervised graph learning methods are also being developed based on databases such as ZINC to address the problem of the insufficiency of labeled data. By incorporating newly developed techniques, deep learning models can be more powerful to explore chemical space in search of new compounds or new properties of existing compounds, which can accelerate drug discovery process.

CRediT authorship contribution statement

Sangsoo Lim: Conceptualization, Investigation, Writing - original draft. **Sangseon Lee:** Conceptualization, Investigation, Writing - original draft. **Yinhua Piao:** Investigation, Writing - original draft. **MinGyu Choi:** Investigation, Writing - original draft. **Dongmin Bang:** Investigation. **Jeonghyeon Gu:** Investigation. **Sun Kim:** Conceptualization, Writing - review & editing, Supervision, Project administration, Funding acquisition.

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.

Acknowledgements

This research was supported by the Collaborative Genome Program for Fostering New Post-Genome Industry of the National Research Foundation (NRF) funded by the Ministry of Science and ICT (MSIT) (No. NRF-2014M3C9A3063541), by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) & funded by the Korean government (MSIT) (No. 2022M3E5F3085677), by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [NO. 2021–0-01343, Artificial Intelligence Graduate School Program (Seoul National University)], and by a grant (No. DY0002259501) from Ministry of food and Drug Safety.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.csbj.2022.07.049.

References

[1] Lipinski C, Hopkins A. Navigating chemical space for biology and medicine. Nature 2004;432:855–61.
[2] Medina-Franco JL, Martínez-Mayorga K, Giulianotti MA, Houghten RA, Pinilla C. Visualization of the chemical space in drug discovery. Curr Comput Aided Drug Des 2008;4:322–33.
[3] López-Vallejo F, Giulianotti MA, Houghten RA, Medina-Franco JL. Expanding the medicinally relevant chemical space with compound libraries. Drug Discovery Today 2012;17:718–26.
[4] Reymond J-L, Van Deursen R, Blum LC, Ruddigkeit L. Chemical space as a source for new drugs. MedChemComm 2010;1:30–8.
[5] Polishchuk PG, Madzhidov TI, Varnek A. Estimation of the size of drug-like chemical space based on gdb-17 data. J Computer-Aided Mol Des 2013;27:675–9.
[6] Reymond J-L. The chemical space project. Acc Chem Res 2015;48:722–30.
[7] Reymond J-L, Awale M. Exploring chemical space for drug discovery using the chemical universe database. ACS Chem Neurosci 2012;3:649–57.
[8] Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Delivery Rev 1997;23:3–25.
[9] Doak BC, Over B, Giordanetto F, Kihlberg J. Oral druggable space beyond the rule of 5: insights from drugs and clinical candidates. Chem Biol 2014;21:1115–42.
[10] B.C. Doak, J. Kihlberg, Drug discovery beyond the rule of 5-opportunities and challenges, 2017.
[11] Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, Leswing K, Pande V. Moleculenet: a benchmark for molecular machine learning. Chem Sci 2018;9:513–30.

[12] K. Huang, T. Fu, W. Gao, Y. Zhao, Y. Roohani, J. Leskovec, C.W. Coley, C. Xiao, J. Sun, M. Zitnik, Therapeutics data commons: machine learning datasets and tasks for therapeutics, arXiv e-prints (2021) arXiv–2102.
[13] Ferreira LL, Andricopulo AD. Admet modeling approaches in drug discovery. Drug Discovery Today 2019;24:1157–65.
[14] Loving KA, Lin A, Cheng AC. Structure-based druggability assessment of the mammalian structural proteome with inclusion of light protein flexibility. PLoS Comput Biol 2014;10:e1003741.
[15] Cheng AC, Eksterowicz J, Geuns-Meyer S, Sun Y. Analysis of kinase inhibitor selectivity using a thermodynamics-based partition index. J Med Chem 2010;53:4502–10.
[16] Van De Waterbeemd H, Gifford E. Admet in silico modelling: towards prediction paradise? Nature Rev Drug Discovery 2003;2:192–204.
[17] Wang N-N, Dong J, Deng Y-H, Zhu M-F, Wen M, Yao Z-J, Lu A-P, Wang J-B, Cao D-S. Adme properties evaluation in drug discovery: prediction of caco-2 cell permeability using a combination of nsga-ii and boosting. J Chem Inf Model 2016;56:763–73.
[18] Hou T, Wang J, Zhang W, Xu X. Adme evaluation in drug discovery. 7. prediction of oral absorption by correlation and classification. J Chem Inform Modeling 2007;47:208–18.
[19] Broccatelli F, Carosati E, Neri A, Frosini M, Goracci L, Oprea TI, Cruciani G. A novel approach for predicting p-glycoprotein (abcb1) inhibition using molecular interaction fields. J Med Chem 2011;54:1740–51.
[20] Ma C-Y, Yang S-Y, Zhang H, Xiang M-L, Huang Q, Wei Y-Q. Prediction models of human plasma protein binding rate and oral bioavailability derived by using ga–cg–svm method. J Pharmaceutical Biomed Anal 2008;47:677–82.
[21] Sorkun MC, Khetan A, Er S. Aqsoldb, a curated reference set of aqueous solubility and 2d descriptors for a diverse set of compounds. Sci Data 2019;6:1–8.
[22] Mobley DL, Guthrie JP. Freesolv: a database of experimental and calculated hydration free energies, with input files. J Computer-Aided Mol Des 2014;28:711–20.
[23] Martins IF, Teixeira AL, Pinheiro L, Falcao AO. A bayesian approach to in silico blood-brain barrier penetration modeling. J Chem Inform Modeling 2012;52:1686–97.
[24] A. Hersey, ChEMBL Deposited Data Set-AZ_dataset, Technical Report, Technical report, EMBL-EBI, 2015. https://www. ebi. ac. uk/chembl/doc..., 2015.
[25] Lombardo F, Jing Y. In silico prediction of volume of distribution in humans. extensive data set and the exploration of linear and nonlinear methods coupled with molecular interaction fields descriptors. J Chem Inf Model 2016;56:2042–52.
[26] Veith H, Southall N, Huang R, James T, Fayne D, Artemenko N, Shen M, Inglese J, Austin CP, Lloyd DG, et al. Comprehensive characterization of cytochrome p450 isozyme selectivity across chemical libraries. Nature Biotechnol 2009;27:1050–5.
[27] Carbon-Mangels M, Hutter MC. Selecting relevant descriptors for classification by bayesian estimates: a comparison with decision trees and support vector machines approaches for disparate data sets. Mol Informatics 2011;30:885–95.
[28] F. Cheng, W. Li, Y. Zhou, J. Shen, Z. Wu, G. Liu, P.W. Lee, Y. Tang, admetsar: a comprehensive source and free tool for assessment of chemical admet properties, 2012.
[29] Obach RS, Lombardo F, Waters NJ. Trend analysis of a database of intravenous pharmacokinetic parameters in humans for 670 drug compounds. Drug Metab Dispos 2008;36:1385–405.
[30] Di L, Keefer C, Scott DO, Strelevitz TJ, Chang G, Bi Y-A, Lai Y, Duckworth J, Fenner K, Troutman MD, et al. Mechanistic insights from comparing intrinsic clearance values between human liver microsomes and hepatocytes to guide drug design. Eur J Med Chem 2012;57:441–8.
[31] Zhu H, Martin TM, Ye L, Sedykh A, Young DM, Tropsha A. Quantitative structure- activity relationship modeling of rat acute toxicity by oral exposure. Chem Res Toxicol 2009;22:1913–21.
[32] Wang S, Sun H, Liu H, Li D, Li Y, Hou T. Admet evaluation in drug discovery. 16. predicting herg blockers by combining multiple pharmacophores and machine learning approaches. Mol Pharmaceutics 2016;13:2855–66.
[33] Du F, Yu H, Zou B, Babcock J, Long S, Li M. hergcentral: a large database to store, retrieve, and analyze compound-human ether-a-go-go related gene channel interactions to facilitate cardiotoxicity assessment in drug development. Assay Drug Dev Technol 2011;9:580–8.
[34] Xu C, Cheng F, Chen L, Du Z, Li W, Liu G, Lee PW, Tang Y. In silico prediction of chemical ames mutagenicity. J Chem Inform Modeling 2012;52:2840–7.
[35] Xu Y, Dai Z, Chen F, Gao S, Pei J, Lai L. Deep learning for drug-induced liver injury. J Chem Inform Modeling 2015;55:2085–93.
[36] Alves VM, Muratov E, Fourches D, Strickland J, Kleinstreuer N, Andrade CH, Tropsha A. Predicting chemically-induced skin reactions. part i: Qsar models of skin sensitization and their application to identify potentially hazardous compounds. Toxicol Appl Pharmacology 2015;284:262–72.
[37] Lagunin A, Filimonov D, Zakharov A, Xie W, Huang Y, Zhu F, Shen T, Yao J, Poroikov V. Computer-aided prediction of rodent carcinogenicity by pass and cisoc-psct. QSAR Combinatorial Sci 2009;28:806–10.
[38] Huang R, Xia M, Nguyen D-T, Zhao T, Sakamuru S, Zhao J, Shahane SA, Rossoshek A, Simeonov A. Tox21challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. Front Environ Sci 2016;3:85.
[39] Richard AM, Judson RS, Houck KA, Grulke CM, Volarath P, Thillainadarajah I, Yang C, Rathman J, Martin MT, Wambaugh JF, et al. Toxcast chemical landscape: paving the road to 21st century toxicology. Chem Res Toxicol 2016;29:1225–51.
[40] Gayvert KM, Madhukar NS, Elemento O. A data-driven approach to predicting successes and failures of clinical trials. Cell Chem Biol 2016;23:1294–301.
[41] Waring MJ. Lipophilicity in drug discovery. Expert Opin Drug Discov 2010;5:235–48.
[42] Savjani KT, Gajjar AK, Savjani JK. Drug solubility: importance and enhancement techniques. Int Scholarly Res Notices 2012;2012.
[43] Smith DA. Evolution of adme science: where else can modeling and simulation contribute? Mol Pharmaceutics 2013;10:1162–70.
[44] M.L. Amin, P-glycoprotein inhibition for optimal drug delivery, Drug target insights 7 (2013) DTI–S12519.
[45] Sambuy Y, De Angelis I, Ranaldi G, Scarino M, Stammati A, Zucco F. The caco-2 cell line as a model of the intestinal barrier: influence of cell and culture-related factors on caco-2 cell functional characteristics. Cell Biol Toxicol 2005;21:1–26.
[46] Abbott NJ, Patabendige AA, Dolman DE, Yusof SR, Begley DJ. Structure and function of the blood–brain barrier. Neurobiol Disease 2010;37:13–25.
[47] J. Aslam, Utilization of big data analysis in biomedicinal chemistry, chemistry 4 (2019).
[48] W. Lindup, M. Orme, Clinical pharmacology: plasma protein binding of drugs., British medical journal (Clinical research ed.) 282 (1981) 212.
[49] McDonnell AM, Dang CH. Basic review of the cytochrome p450 system. J Adv Practitioner Oncol 2013;4:263.
[50] Teh LK, Bertilsson L. Pharmacogenomics of cyp2d6: molecular genetics, interethnic differences and clinical importance. Drug metabolism and pharmacokinetics 2011. 1112190300–1112190300.
[51] Zanger UM, Schwab M. Cytochrome p450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. Pharmacol Therapeutics 2013;138:103–41.
[52] Kirchmair J, Göller AH, Lang D, Kunze J, Testa B, Wilson ID, Glen RC, Schneider G. Predicting drug metabolism: experiment and/or computation? Nature Rev Drug Discovery 2015;14:387–404.
[53] Ma J, Sheridan RP, Liaw A, Dahl GE, Svetnik V. Deep neural nets as a method for quantitative structure–activity relationships. J Chem Inform Modeling 2015;55:263–74.
[54] Feinberg EN, Joshi E, Pande VS, Cheng AC. Improvement in admet prediction with multitask deep featurization. J Med Chem 2020;63:8835–48.
[55] Benet LZ, Zia-Amirhosseini P. Basic principles of pharmacokinetics. Toxicologic Pathol 1995;23:115–23.
[56] Toutain P-L, Bousquet-mélou A. Plasma clearance. J Veterinary Pharmacology Therapeutics 2004;27:415–25.
[57] Durairaj C, Shah JC, Senapati S, Kompella UB. Prediction of vitreal half-life based on drug physicochemical properties: quantitative structure–pharmacokinetic relationships (qspkr). Pharmaceutical Res 2009;26:1236–60.
[58] Kramer JA, Sagartz JE, Morris DL. The application of discovery toxicology and pathology towards the design of safer pharmaceutical lead candidates. Nature Rev Drug Discovery 2007;6:636–49.
[59] Walters WP, Murcko M. Assessing the impact of generative ai on medicinal chemistry. Nature Biotechnol 2020;38:143–5.
[60] Polykovskiy D, Zhebrak A, Sanchez-Lengeling B, Golovanov S, Tatanov O, Belyaev S, Kurbanov R, Artamonov A, Aladinskiy V, Veselov M, et al. Molecular sets (moses): a benchmarking platform for molecular generation models. Front Pharmacol 2020;11:565644.
[61] Brown N, Fiscato M, Segler MH, Vaucher AC. Guacamol: benchmarking models for de novo molecular design. J Chem Inform Modeling 2019;59:1096–108.
[62] Lane SJ, Eggleston DS, Brinded KA, Hollerton JC, Taylor NL, Readshaw SA. Defining and maintaining a high quality screening collection: the gsk experience. Drug Discovery Today 2006;11:267–72.
[63] Papadatos G, Davies M, Dedman N, Chambers J, Gaulton A, Siddle J, Koks R, Irvine SA, Pettersson J, Goncharoff N, et al. Surechembl: a large-scale, chemically annotated patent document database. Nucl Acids Res 2016;44:D1220–8.
[64] Baell JB, Holloway GA. New substructure filters for removal of pan assay interference compounds (pains) from screening libraries and for their exclusion in bioassays. J Med Chem 2010;53:2719–40.
[65] Ruddigkeit L, Van Deursen R, Blum LC, Reymond J-L. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. J Chem Inform Modeling 2012;52:2864–75.
[66] Ramakrishnan R, Dral PO, Rupp M, Von Lilienfeld OA. Quantum chemistry structures and properties of 134 kilo molecules. Sci Data 2014;1:1–7.
[67] Zhavoronkov A, Ivanenkov YA, Aliper A, Veselov MS, Aladinskiy VA, Aladinskaya AV, Terentiev VA, Polykovskiy DA, Kuznetsov MD, Asadulaev A, et al. Deep learning enables rapid identification of potent ddr1 kinase inhibitors. Nature Biotechnol 2019;37:1038–40.
[68] Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG. Zinc: a free tool to discover chemistry for biology. J Chem Inform Modeling 2012;52:1757–68.
[69] Merk D, Friedrich L, Grisoni F, Schneider G. De novo design of bioactive small molecules by artificial intelligence. Mol Inform 2018;37:1700153.
[70] Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrián-Uhalte E, et al. The chembl database in 2017. Nucl Acids Res 2017;45:D945–54.

[71] Germain P, Chambon P, Eichele G, Evans RM, Lazar MA, Leid M, De Lera AR, Lotan R, Mangelsdorf DJ, Gronemeyer H. International union of pharmacology. lxiii. retinoid x receptors. Pharmacol Rev 2006;58:760–72.
[72] Michalik L, Auwerx J, Berger JP, Chatterjee VK, Glass CK, Gonzalez FJ, Grimaldi PA, Kadowaki T, Lazar MA, O'Rahilly S, et al. International union of pharmacology. lxi. peroxisome proliferator-activated receptors. Pharmacol Rev 2006;58:726–41.
[73] Rohrer SG, Baumann K. Maximum unbiased validation (muv) data sets for virtual screening based on pubchem bioactivity data. J Chem Inform Modeling 2009;49:169–84.
[74] Wang Y, Bryant SH, Cheng T, Wang J, Gindulyte A, Shoemaker BA, Thiessen PA, He S, Zhang J. Pubchem bioassay: 2017 update. Nucl Acids Res 2017;45 (2017):D955–63.
[75] Holbeck S. Update on nci in vitro drug screen utilities. Eur J Cancer 2004;40:785–93.
[76] Case DA, Cheatham III TE, Darden T, Gohlke H, Luo R, Merz Jr KM, Onufriev A, Simmerling C, Wang B, Woods RJ. The amber biomolecular simulation programs. J Comput Chem 2005;26:1668–88.
[77] Hanwell MD, Curtis DE, Lonie DC, Vandermeersch T, Zurek E, Hutchison GR. Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. J Cheminformatics 2012;4:1–17.
[78] Brooks BR, Brooks III CL, Mackerell Jr AD, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, et al. Charmm: the biomolecular simulation program. J Comput Chem 2009;30:1545–614.
[79] E.E. Bolton, Y. Wang, P.A. Thiessen, S.H. Bryant, Pubchem: integrated platform of small molecules and biological activities, in: Annual reports in computational chemistry, volume 4, Elsevier, 2008, pp. 217–241.
[80] Smith JS, Isayev O, Roitberg AE. Ani-1: an extensible neural network potential with dft accuracy at force field computational cost. Chem Sci 2017;8:3192–203.
[81] Chmiela S, Tkatchenko A, Sauceda HE, Poltavsky I, Schütt KT, Müller K-R. Machine learning of accurate energy-conserving molecular force fields. Science advances 2017;3:e1603015.
[82] Chmiela S, Sauceda HE, Poltavsky I, Müller K-R, Tkatchenko A. sgdml: Constructing accurate and data efficient molecular force fields using machine learning. Comput Phys Commun 2019;240:38–45.
[83] Heifetz A. Quantum mechanics in drug discovery. Springer; 2020.
[84] Li H, Sze K-H, Lu G, Ballester PJ. Machine-learning scoring functions for structure-based drug lead optimization. Wiley Interdisciplinary Rev: Comput Mol Sci 2020;10:e1465.
[85] Temml V, Kutil Z. Structure-based molecular modeling in sar analysis and lead optimization, Computational and Structural. Biotechnol J 2021;19:1431–44.
[86] de Souza Neto LR, Moreira-Filho JT, Neves BJ, Maidana RLBR, Guimarães ACR, Furnham N, Andrade CH, Silva Jr FP. In silico strategies to support fragment-to-lead optimization in drug discovery. Front Chem 2020;8:93.
[87] Weininger D. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. J Chem Inform Computer Sci 1988;28:31–6.
[88] Stork C, Chen Y, Sicho M, Kirchmair J. Hit dexter 2.0: machine-learning models for the prediction of frequent hitters. J Chem Inform Modeling 2019;59:1030–43.
[89] Krenn M, Häse F, Nigam A, Friederich P, Aspuru-Guzik A. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. Mach Learn: Sci Technol 2020;1:045024.
[90] Cereto-Massagué A, Ojeda MJ, Valls C, Mulero M, Garcia-Vallvé S, Pujadas G. Molecular fingerprint similarity search in virtual screening. Methods 2015;71:58–63.
[91] Morgan HL. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. J Chem Doc 1965;5:107–13.
[92] Durant JL, Leland BA, Henry DR, Nourse JG. Reoptimization of mdl keys for use in drug discovery. J Chem Inform Computer Sci 2002;42:1273–80.
[93] Garg D, Gandhi T, Mohan CG. Exploring qstr and toxicophore of herg k+ channel blockers using gfa and hypogen techniques. J Mol Graph Model 2008;26:966–76.
[94] Singh PK, Negi A, Gupta PK, Chauhan M, Kumar R. Toxicophore exploration as a screening technology for drug design and discovery: techniques, scope and limitations. Arch Toxicol 2016;90:1785–802.
[95] S. Chithrananda, G. Grand, B. Ramsundar, Chemberta: Large-scale self-supervised pretraining for molecular property prediction, arXiv preprint arXiv:2010.09885 (2020).
[96] S. Zhang, Z. Hu, A. Subramonian, Y. Sun, Motif-driven contrastive learning of graph representations, arXiv preprint arXiv:2012.12533 (2020).
[97] M. Sun, J. Xing, H. Wang, B. Chen, J. Zhou, Mocl: Contrastive learning on molecular graphs with multi-level domain knowledge, arXiv preprint arXiv:2106.04509 (2021).
[98] Y. Wang, J. Wang, Z. Cao, A.B. Farimani, Molclr: molecular contrastive learning of representations via graph neural networks, arXiv preprint arXiv:2102.10056 (2021).
[99] G.B. Goh, N.O. Hodas, C. Siegel, A. Vishnu, Smiles2vec: An interpretable general-purpose deep neural network for predicting chemical properties, arXiv preprint arXiv:1712.02034 (2017).
[100] Y. Fang, Q. Zhang, H. Yang, X. Zhuang, S. Deng, W. Zhang, M. Qin, Z. Chen, X. Fan, H. Chen, Molecular contrastive learning with chemical element knowledge graph, arXiv preprint arXiv:2112.00544 (2021).
[101] Karpov P, Godin G, Tetko IV. Transformer-cnn: Swiss knife for qsar modeling and interpretation. J Cheminformatics 2020;12:1–12.
[102] Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, Guzman-Perez A, Hopper T, Kelley B, Mathea M, et al. Analyzing learned molecular representations for property prediction. J Chem Inform Modeling 2019;59:3370–88.
[103] Feinberg EN, Sur D, Wu Z, Husic BE, Mai H, Li Y, Sun S, Yang J, Ramsundar B, Pande VS. Potentialnet for molecular property prediction. ACS Central Sci 2018;4:1520–30.
[104] Zhang Z, Guan J, Zhou S. Fragat: a fragment-oriented multi-scale graph attention model for molecular property prediction. Bioinformatics 2021;37:2981–7.
[105] Li P, Li Y, Hsieh C-Y, Zhang S, Liu X, Liu H, Song S, Yao X. Trimnet: learning molecular representation from triplet messages for biomedicine. Briefings Bioinform 2021;22:bbaa266.
[106] Lv Q, Chen G, Zhao L, Zhong W, Yu-Chian Chen C. Mol2context-vec: learning molecular representation from context awareness for drug discovery. Briefings Bioinform 2021;22:bbab317.
[107] Y. Song, S. Zheng, Z. Niu, Z.-H. Fu, Y. Lu, Y. Yang, Communicative representation learning on attributed molecular graphs., in: IJCAI, volume 2020, 2020, pp. 2831–2838.
[108] Jo J, Kwak B, Choi H-S, Yoon S. The message passing neural networks for chemical property prediction on smiles. Methods 2020;179:65–72.
[109] Pope PE, Kolouri S, Rostami M, Martin CE, Hoffmann H. Explainability methods for graph convolutional neural networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society; 2019. p. 10764–73.
[110] Jeon W, Kim D. Fp2vec: a new molecular featurizer for learning molecular properties. Bioinformatics 2019;35:4979–85.
[111] Kruskal JB. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 1964;29:1–27.
[112] Peng Y, Zhang Z, Jiang Q, Guan J, Zhou S. Top: a deep mixture representation learning method for boosting molecular toxicity prediction. Methods 2020;179:55–64.
[113] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Adv Neural Inform Process Syst 2012;25.
[114] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. p. 1–9.
[115] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. p. 7132–41.
[116] Xie S, Girshick R, Dollár P, Tu Z, He K. Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. p. 1492–500.
[117] Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-first AAAI conference on artificial intelligence.
[118] Hirohara M, Saito Y, Koda Y, Sato K, Sakakibara Y. Convolutional neural network based on smiles representation of compounds for detecting chemical motif. BMC Bioinformatics 2018;19:83–94.
[119] I. Wallach, M. Dzamba, A. Heifets, Atomnet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery, arXiv preprint arXiv:1510.02855 (2015).
[120] G.B. Goh, C. Siegel, A. Vishnu, N.O. Hodas, N. Baker, Chemception: a deep neural network with minimal chemistry knowledge matches the performance of expert-developed qsar/qspr models, arXiv preprint arXiv:1706.06689 (2017).
[121] Goh GB, Siegel C, Vishnu A, Hodas N, Baker N. How much chemistry does a deep neural network need to know to make accurate predictions? In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE; 2018. p. 1340–9.
[122] Ragoza M, Hochuli J, Idrobo E, Sunseri J, Koes DR. Protein–ligand scoring with convolutional neural networks. J Chem Inform Modeling 2017;57:942–57.
[123] Lim S, Lee YO. Predicting chemical properties using self-attention multi-task learning based on smiles representation. In: 2020 25th International Conference on Pattern Recognition (ICPR). IEEE; 2021. p. 3146–53.
[124] S. Honda, S. Shi, H.R. Ueda, Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery, arXiv preprint arXiv:1911.04738 (2019).
[125] Zheng S, Yan X, Yang Y, Xu J. Identifying structure–property relationships through smiles syntax analysis with self-attention mechanism. J Chem Inform Modeling 2019;59:914–23.
[126] Chen J-H, Tseng YJ. A general optimization protocol for molecular property prediction using a deep learning network. Briefings in Bioinformatics 2022;23:bbab367.
[127] P. Ertl, R. Lewis, E. Martin, V. Polyakov, In silico generation of novel, drug-like chemical matter using the lstm neural network, arXiv preprint arXiv:1712.07449 (2017).
[128] Li C, Feng J, Liu S, Yao J. A novel molecular representation learning for molecular property prediction with a multiple smiles-based augmentation. Comput Intell Neurosci 2022;2022.
[129] Kimber TB, Gagnebin M, Volkamer A. Maxsmi: maximizing molecular property prediction performance with confidence estimation using smiles augmentation and deep learning. Artif Intell Life Sci 2021;1:100014.
[130] Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inform Modeling 2010;50:742–54.

[131] Degen J, Wegscheid-Gerlach C, Zaliani A, Rarey M. On the art of compiling and using'drug-like'chemical fragment spaces. , : Chemistry Enabling 2008;3:1503–7.
[132] D. Sumner, J. He, A. Thakkar, O. Engkvist, E.J. Bjerrum, Levenshtein augmentation improves performance of smiles based deep-learning synthesis prediction (2020).
[133] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Adv Neural Inform Processing Syst 2017;30.
[134] J.D.M.-W.C. Kenton, L.K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, 2019, pp. 4171–4186.
[135] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36:1234–40.
[136] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2020.
[137] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. p. 10012–22.
[138] Ł. Maziarka, T. Danel, S. Mucha, K. Rataj, J. Tabor, S. Jastrzebski, Molecule attention transformer, arXiv preprint arXiv:2002.08264 (2020).
[139] Wang S, Guo Y, Wang Y, Sun H, Huang J. Smiles-bert: large scale unsupervised pre-training for molecular property prediction. In: Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics. p. 429–36.
[140] D. Xue, H. Zhang, D. Xiao, Y. Gong, G. Chuai, Y. Sun, H. Tian, H. Wu, Y. Li, Q. Liu, X-mol: large-scale pre-training for molecular understanding and diverse molecular analysis, bioRxiv (2021) 2020–12.
[141] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[142] W. Zhu, Z. Li, L. Cai, G. Song, Stepping back to smiles transformers for fast molecular representation inference, arXiv preprint arXiv:2112.13305 (2021).
[143] R. v. d. Berg, T.N. Kipf, M. Welling, Graph convolutional matrix completion, arXiv preprint arXiv:1706.02263 (2017).
[144] Fout A, Byrd J, Shariat B, Ben-Hur A. Protein interface prediction using graph convolutional networks. Adv Neural Inform Processing Systems 2017;30.
[145] Sun M, Zhao S, Gilvary C, Elemento O, Zhou J, Wang F. Graph convolutional networks for computational drug development and discovery. Briefings Bioinform 2020;21:919–35.
[146] Kearnes S, McCloskey K, Berndl M, Pande V, Riley P. Molecular graph convolutions: moving beyond fingerprints. J Computer-Aided Mol Des 2016;30:595–608.
[147] Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. Neural message passing for quantum chemistry. In: International conference on machine learning. PMLR; 2017. p. 1263–72.
[148] Tsubaki M, Tomii K, Sese J. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. Bioinformatics 2019;35:309–18.
[149] Yang Q, Ji H, Lu H, Zhang Z. Prediction of liquid chromatographic retention time with graph neural networks to assist in small molecule identification. Anal Chem 2021;93:2200–6.
[150] Meng M, Wei Z, Li Z, Jiang M, Bian Y. Property prediction of molecules in graph convolutional neural network expansion. In: 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS). IEEE; 2019. p. 263–6.
[151] Y. Li, D. Tarlow, M. Brockschmidt, R. Zemel, Gated graph sequence neural networks, arXiv preprint arXiv:1511.05493 (2015).
[152] J. Klicpera, J. Groß, S. Günnemann, Directional message passing for molecular graphs, in: International Conference on Learning Representations, 2019.
[153] Klicpera J, Becker F, Günnemann S. Gemnet: Universal directional graph neural networks for molecules. Adv Neural Inform Process Syst 2021;34.
[154] S. Ryu, J. Lim, S.H. Hong, W.Y. Kim, Deeply learning molecular structure-property relationships using attention-and gate-augmented graph convolutional network, arXiv preprint arXiv:1805.10988 (2018).
[155] Xiong Z, Wang D, Liu X, Zhong F, Wan X, Li X, Li Z, Luo X, Chen K, Jiang H, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. J Med Chem 2019;63:8749–60.
[156] C. Lu, Q. Liu, C. Wang, Z. Huang, P. Lin, L. He, Molecular property prediction: A multilevel quantum interactions modeling perspective, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 1052–1060.
[157] Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, Hubert T, Baker L, Lai M, Bolton A, et al. Mastering the game of go without human knowledge. Nature 2017;550:354–9.
[158] Sallab AE, Abdou M, Perot E, Yogamani S. Deep reinforcement learning framework for autonomous driving. Electronic Imaging 2017;2017:70–6.
[159] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. Highly accurate protein structure prediction with alphafold. Nature 2021;596:583–9.
[160] Olivecrona M, Blaschke T, Engkvist O, Chen H. Molecular de-novo design through deep reinforcement learning. J Cheminformatics 2017;9:1–14.
[161] Ståhl N, Falkman G, Karlsson A, Mathiason G, Bostrom J. Deep reinforcement learning for multiparameter optimization in de novo drug design. J Chem Inform Modeling 2019;59:3166–76.
[162] Boström J, Hogner A, Schmitt S. Do structurally similar ligands bind in a similar fashion? J Med Chem 2006;49:6716–25.
[163] Levenshtein VI et al. Binary codes capable of correcting deletions, insertions, and reversals. Soviet physics doklady, volume 10. Soviet Union; 1966. p. 707–10.
[164] You J, Liu B, Ying Z, Pande V, Leskovec J. Graph convolutional policy network for goal-directed molecular graph generation. Adv Neural Inform Process Syst 2018;31.
[165] Zhou Z, Kearnes S, Li L, Zare RN, Riley P. Optimization of molecules via deep reinforcement learning. Sci Rep 2019;9:1–10.
[166] Goel M, Raghunathan S, Laghuvarapu S, Priyakumar UD. Molegular: Molecule generation using reinforcement learning with alternating rewards. J Chem Inf Model 2021;61:5815–26.
[167] Horwood J, Noutahi E. Molecular design in synthetically accessible chemical space via deep reinforcement learning. ACS Omega 2020;5:32984–94.
[168] Jin W, Barzilay R, Jaakkola T. Junction tree variational autoencoder for molecular graph generation. In: International conference on machine learning. PMLR; 2018. p. 2323–32.
[169] Irwin JJ, Tang KG, Young J, Dandarchuluun C, Wong BR, Khurelbaatar M, Moroz YS, Mayfield J, Sayle RA. Zinc20—a free ultralarge-scale chemical database for ligand discovery. J Chem Inform Modeling 2020;60:6065–73.
[170] Coley CW. Defining and exploring chemical spaces. Trends Chem 2021;3:133–45.
[171] Öztürk H, Özgür A, Schwaller P, Laino T, Ozkirimli E. Exploring chemical space using natural language processing methodologies for drug discovery. Drug Discovery Today 2020;25:689–705.
[172] Maragakis P, Nisonoff H, Cole B, Shaw DE. A deep-learning view of chemical space designed to facilitate drug discovery. J Chem Inf Model 2020;60:4487–96.
[173] Glavatskikh M, Leguy J, Hunault G, Cauchy T, Da Mota B. Dataset's chemical diversity limits the generalizability of machine learning predictions. J Cheminformatics 2019;11:1–15.
[174] F. Broccatelli, R. Trager, M. Reutlinger, G. Karypis, M. Li, Benchmarking accuracy and generalizability of four graph neural networks using large in vitro adme datasets from different chemical spaces, arXiv preprint arXiv:2111.13964 (2021).
[175] Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, et al. Pubchem 2019 update: improved access to chemical data. Nucl Acids Res 2019;47:D1102–9.
[176] W. Hu, B. Liu, J. Gomes, M. Zitnik, P. Liang, V. Pande, J. Leskovec, Strategies for pre-training graph neural networks, arXiv preprint arXiv:1905.12265 (2019).
[177] Liu S, Demirel MF, Liang Y. N-gram graph: Simple unsupervised representation for graphs, with applications to molecules. Adv Neural Inform Process Syst 2019;32.
[178] Hu Z, Dong Y, Wang K, Chang K-W, Sun Y. Gpt-gnn: Generative pre-training of graph neural networks. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. p. 1857–67.
[179] Rong Y, Bian Y, Xu T, Xie W, Wei Y, Huang W, Huang J. Self-supervised graph transformer on large-scale molecular data. Adv Neural Inform Process Syst 2020;33:12559–71.
[180] Zhang Z, Liu Q, Wang H, Lu C, Lee C-K. Motif-based graph self-supervised learning for molecular property prediction. Adv Neural Inform Process Syst 2021;34.
[181] Li P, Wang J, Qiao Y, Chen H, Yu Y, Yao X, Gao P, Xie G, Song S. An effective self-supervised framework for learning expressive molecular global representations to drug discovery. Briefings Bioinformatics 2021;22: bbab109.
[182] D. Kim, J. Baek, S.J. Hwang, Graph self-supervised learning with accurate discrepancy learning, arXiv preprint arXiv:2202.02989 (2022).
[183] Qiu J, Chen Q, Dong Y, Zhang J, Yang H, Ding M, Wang K, Tang J. Gcc: Graph contrastive coding for graph neural network pre-training. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. p. 1150–60.
[184] Sun Q, Li J, Peng H, Wu J, Ning Y, Yu PS, He L. Sugar: Subgraph neural network with reinforcement pooling and self-supervised mutual information mechanism. In: Proceedings of the Web Conference 2021. p. 2081–91.
[185] H. Hafidi, M. Ghogho, P. Ciblat, A. Swami, Graphcl: Contrastive self-supervised learning of graph representations, arXiv preprint arXiv:2007.08025 (2020).
[186] J. Zeng, P. Xie, Contrastive self-supervised learning for graph classification, arXiv preprint arXiv:2009.05923 (2020).
[187] Ren Y, Bai J, Zhang J. Label contrastive coding based graph neural network for graph classification. In: International Conference on Database Systems for Advanced Applications. Springer; 2021. p. 123–40.
[188] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. p. 4401–10.
[189] Razavi A, Van den Oord A, Vinyals O. Generating diverse high-fidelity images with vq-vae-2. Adv Neural Inform Process Syst 2019;32.
[190] Rao K, Harris C, Irpan A, Levine S, Ibarz J, Khansari M. Rl-cyclegan: Reinforcement learning aware simulation-to-real. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition. p. 11157–66.
[191] Maziarka Ł, Pocha A, Kaczmarczyk J, Rataj K, Danel T, Warchoł M. Mol-cyclegan: a generative model for molecular optimization. J Cheminformatics 2020;12:1–18.
[192] Zhu J-Y, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. p. 2223–32.
[193] Gebauer NW, Gastegger M, Hessmann SS, Müller K-R, Schütt KT. Inverse design of 3d molecular structures with conditional generative neural networks. Nature communications 2022;13:1–11.
[194] Simm G, Pinsler R, Hernández-Lobato JM. Reinforcement learning for molecular design guided by quantum mechanics. In: International Conference on Machine Learning. PMLR; 2020. p. 8959–69.
[195] Jeon W, Kim D. Autonomous molecule generation using reinforcement learning and docking to develop potential novel inhibitors. Scientific reports 2020;10:1–11.
[196] Méndez-Lucio O, Baillif B, Clevert D-A, Rouquié D, Wichard J. De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. Nature Commun 2020;11:1–10.
[197] Wang Y, Wang W, Liang Y, Cai Y, Hooi B. Mixup for node and graph classification. In: Proceedings of the Web Conference 2021. p. 3663–74.
[198] J. Park, H. Shim, E. Yang, Graph transplant: Node saliency-guided graph mixup with local structure preservation, arXiv preprint arXiv:2111.05639 (2021).
[199] H. Zhang, M. Cisse, Y.N. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, in: International Conference on Learning Representations, 2018.
[200] Verma V, Lamb A, Beckham C, Najafi A, Mitliagkas I, Lopez-Paz D, Bengio Y. Manifold mixup: Better representations by interpolating hidden states. In: International Conference on Machine Learning. PMLR; 2019. p. 6438–47.
[201] Lusci A, Pollastri G, Baldi P. Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. J Chem Inform Modeling 2013;53:1563–75.
[202] K. Swanson, Message passing neural networks for molecular property prediction, Ph.D. thesis, Massachusetts Institute of Technology, 2019.
[203] J. Chen, S. Zheng, Y. Song, J. Rao, Y. Yang, Learning attributed graph representations with communicative message passing transformer, arXiv preprint arXiv:2107.08773 (2021).
[204] Lindpaintner K. The impact of pharmacogenetics and pharmacogenomics on drug discovery. Nat Rev Drug Discovery 2002;1:463–9.
[205] Rengarajan T, Rajendran P, Nandakumar N, Lokeshkumar B, Rajendran P, Nishigaki I. Exposure to polycyclic aromatic hydrocarbons with special focus on cancer. Asian Pacific J Tropical Biomed 2015;5:182–9.
[206] Tharwat A, Moemen YS, Hassanien AE. A predictive model for toxicity effects assessment of biotransformed hepatic drugs using iterative sampling method. Sci Rep 2016;6:1–13.
[207] Chen Y, Miao D, Wang R. A rough set approach to feature selection based on ant colony optimization. Pattern Recogn Lett 2010;31:226–33.
[208] Xu C, Li CY-T, Kong A-NT. Induction of phase i, ii and iii drug metabolism/transport by xenobiotics. Arch Pharmacal Res 2005;28:249–68.
[209] De Longueville F, Surry D, Meneses-Lorente G, Bertholet V, Talbot V, Evrard S, Chandelier N, Pike A, Worboys P, Rasson J-P, et al. Gene expression profiling of drug metabolism and toxicology markers using a low-density dna microarray. Biochem Pharmacol 2002;64:137–49.
[210] Fielden MR, Eynon BP, Natsoulis G, Jarnagin K, Banas D, Kolaja KL. A gene expression signature that predicts the future onset of drug-induced renal tubular toxicity. Toxicol Pathol 2005;33:675–83.
[211] Alexander-Dann B, Pruteanu LL, Oerton E, Sharma N, Berindan-Neagoe I, Módos D, Bender A. Developments in toxicogenomics: understanding and predicting compound-induced toxicity from gene expression data. Molecular omics 2018;14:218–36.
[212] Wu Y, Wang G. Machine learning based toxicity prediction: from chemical structural description to transcriptome analysis. Int J Mol Sci 2018;19:2358.
[213] Vo AH, Van Vleet TR, Gupta RR, Liguori MJ, Rao MS. An overview of machine learning and big data for drug toxicity evaluation. Chem Res Toxicol 2019;33:20–37.
[214] Kaitoh K, Yamanishi Y. Triomphe: Transcriptome-based inference and generation of molecules with desired phenotypes by machine learning. J Chem Inf Model 2021;61:4303–20.
[215] Zhang F, Wang M, Xi J, Yang J, Li A. A novel heterogeneous network-based method for drug response prediction in cancer cell lines. Sci Rep 2018;8:1–9.
[216] Güvenç Paltun B, Mamitsuka H, Kaski S. Improving drug response prediction by integrating multiple data sources: matrix factorization, kernel and network-based approaches. Briefings Bioinform 2021;22:346–59.
[217] Lim S, Lu Y, Cho CY, Sung I, Kim J, Kim Y, Park S, Kim S. A review on compound-protein interaction prediction methods: data, format, representation and model, Computational and Structural. Biotechnol J 2021;19:1541–56.
[218] Menon A, Krdzavac NB, Kraft M. From database to knowledge graph—using data in chemistry. Current Opinion Chem Eng 2019;26:33–7.
[219] Lu J, Niu B, Liu L, Lu W-C, Cai Y-D. Prediction of small molecules' metabolic pathways based on functional group composition. Protein Pept Lett 2009;16:969–76.
[220] Brown FJ, Yee YK, Cronk LA, Hebbel KC, Krell RD, Snyder DW. Evolution of a series of peptidoleukotriene antagonists: Synthesis and structure-activity relationships of 1, 6-disubstituted indoles and indazoles. J Med Chem 1990;33:1771–81.
[221] Proschak E, Heitel P, Kalinowsky L, Merk D. Opportunities and challenges for fatty acid mimetics in drug discovery. J Med Chem 2017;60:5235–66.
[222] Wang Y, Abuduweili A, Yao Q, Dou D. Property-aware relation networks for few-shot molecular property prediction. Adv Neural Inform Process Syst 2021;34.
[223] Mervin LH, Afzal AM, Drakakis G, Lewis R, Engkvist O, Bender A. Target prediction utilising negative bioactivity data covering large chemical space. J Cheminformatics 2015;7:1–16.
[224] Liu H, Sun J, Guan J, Zheng J, Zhou S. Improving compound–protein interaction prediction by building up highly credible negative samples. Bioinformatics 2015;31:i221–9.
[225] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15:1929–58.
[226] C. Zhang, O. Vinyals, R. Munos, S. Bengio, A study on overfitting in deep reinforcement learning, arXiv preprint arXiv:1804.06893 (2018).
[227] M. Hardt, B. Recht, Y. Singer, Train faster, generalize better: Stability of stochastic gradient descent, in: International conference on machine learning, PMLR, 2016, pp. 1225–1234.
[228] Xiong HY, Barash Y, Frey BJ. Bayesian prediction of tissue-regulated splicing using rna sequence and cellular context. Bioinformatics 2011;27:2554–62.
[229] Salakhutdinov R, Mnih A. Bayesian probabilistic matrix factorization using markov chain monte carlo. In: Proceedings of the 25th international conference on Machine learning. p. 880–7.
[230] Sheridan RP. Time-split cross-validation as a method for estimating the goodness of prospective prediction. J Chem Inform Modeling 2013;53:783–90.
[231] Tran-Nguyen V-K, Jacquemard C, Rognan D. Lit-pcba: An unbiased data set for machine learning and virtual screening. J Chem Inform Modeling 2020;60:4263–73.
[232] H. Cai, H. Zhang, D. Zhao, J. Wu, L. Wang, Fp-gnn: a versatile deep learning architecture for enhanced molecular property prediction, arXiv preprint arXiv:2205.03834 (2022).
[233] Bjerrum EJ, Sattarov B. Improving chemical autoencoder latent space and molecular de novo generation diversity with heteroencoders. Biomolecules 2018;8:131.