International Journal of Data Science and Analytics (2023) 16:301–316
https://doi.org/10.1007/s41060-022-00371-8

REVIEW

Learning to discover medicines

Minh-Tri Nguyen[1] · **Thin Nguyen**[1] · **Truyen Tran**[1]

Received: 9 June 2022 / Accepted: 5 November 2022 / Published online: 18 November 2022

Abstract
Discovering new medicines is the hallmark of the human endeavor to live a better and longer life. Yet the pace of discovery has slowed down as we need to venture into more wildly unexplored biomedical space to find one that matches today's high standard. Modern AI-enabled by powerful computing, large biomedical databases, and breakthroughs in deep learning offers a new hope to break this loop as AI is rapidly maturing, ready to make a huge impact in the area. In this paper, we review recent advances in AI methodologies that aim to crack this challenge. We organize the vast and rapidly growing literature on AI for drug discovery into three relatively stable sub-areas: (a) *representation learning* over molecular sequences and geometric graphs; (b) *data-driven reasoning* where we predict molecular properties and their binding, optimize existing compounds, generate *de novo* molecules, and plan the synthesis of target molecules; and (c) *knowledge-based reasoning* where we discuss the construction and reasoning over biomedical knowledge graphs. We will also identify open challenges and chart possible research directions for the years to come.

Keywords Drug discovery · Artificial intelligence · Machine learning · Biomedical representation learning · Drug discovery reasoning

1 Introduction

The COVID-19 pandemic has triggered an unprecedented rise of investment capital in AI for drug discovery (DD), the process of identifying new medicines for a druggable target [1]. Recent breakthroughs in AI present a great opportunity to break the so-called *Eroom's Law* in DD–the inverse of the well-known Moore's Law–dictating that the rate of FDA drug approval is slowing down despite a huge increase in development cost [2]. Enabled by deep learning advances, powerful computing, and large databases, modern AI is ready to make a huge impact through *in silico* processes to supplement and sometimes replace the *in vitro* counterparts [3] of drug development. For a wide range of problems, from determining the 3D structure of proteins to predicting drug-target binding, to generating synthesizable molecules, AI has helped change the DD landscape in recent years. The reverse also holds: The problems in DD necessitate new advances in AI methodologies to deal with the new scope and complexity typically not seen in traditional application domains of AI such as computer vision and language processing.

To understand the new problems DD brings to AI, we first briefly introduce the DD pipeline [4]. The pipeline of DD involves five main steps: discovery and development, preclinical research, clinical research, drug review and approval, post-market safety monitoring. In the discovery and development phase, the candidate drugs are found by learning about the target disease, existing drugs with newly found effects (drug repurposing), or suggestion from other sources such as AI framework. The first phase has four main steps: target identification and validation, hit discovery and confirmation, hit-to-lead, and lead optimization. Target identification and validation identify the biological causes of the target disease. Hit discovery and confirmation searches for the 'hit' molecules in the drug database. Hit-to-lead optimizes the 'hit' molecules to satisfy more desirable properties. Lead optimization reduces flaws while maintaining the desired properties. Preclinical research phase assesses the dosing and toxicity by in vitro and in vivo testing. Clinical research phase is human trial phase. In the drug review and approval phase,

✉ Minh-Tri Nguyen
tri.nguyen1@deakin.edu.au

Thin Nguyen
thin.nguyen@deakin.edu.au

Truyen Tran
truyen.tran@deakin.edu.au

[1] Applied Artificial Intelligence Institute, Deakin University, Burwood, VIC, Australia

clinical trial data are assessed for drug approval. Post-market safety phase monitors the drug's effectiveness and safety in the marketplace.

From the DD pipeline, *there are three major DD questions AI can help answer*. The first is, given the molecule, what are its chemo-biological and therapeutic properties? Second, for a given target, what kind of molecules will therapeutically modify its functions? Finally, given a molecule, how can we synthesize and optimize the molecule from the available compounds, meaning solving the problems of synthetic tractability and reaction planning?

In this survey, we bring in the AI and reasoning perspectives for answering these questions, with an emphasis on recent developments. Each question poses representation, learning, and reasoning sub-problems. This is because drugs, targets, and the hosting environments need to be represented in machine comprehensible formats. Section 2 shows the current trend of using learned representation to explore the power of computing and the richness of data. Early works in AI for DD use biological sequences with their physical-chemical characteristic features. With the availability of structure information (large database [5] or high-accuracy prediction tool [6]) and powerful and efficient representation learning model [7], recent works focus on learning features from different sources via pre-training. Once the learning has been completed, the next phases of prediction, search, and discovery are performed using reasoning methods that leverage the learned models with data-driven reasoning (Sec. 3) and the vast domain knowledge with knowledge-driven reasoning (Sec. 4). The common tasks for AI model in DD involve answering major DD questions such as properties prediction, interaction prediction, synthesize, and optimization. Before concluding, we will discuss the remaining challenges and opportunities for AI/ML in this important area (Sec. 5). See Fig. 1 for a taxonomy of the problem space, which we follow in the paper.

The recent surveys [8–10] focus on drug discovery AI techniques which learn the target tasks using the biological entities' characteristics. Our survey frames the drug discovery problems in AI as a reasoning process that starts from the representation learning to reason on the data structure (data-driven reasoning) and especially on the learnt knowledge (knowledge-driven reasoning).

The methods chosen to be discussed in this paper have been peer-reviewed in journals and conferences or received more than one hundred citations.

2 Learning representations

The first step in applying AI/ML is to form a computer-readable representation of biomedical entities and concepts. We will primarily focus on the drug-target pairs. A *drug* is a small molecule, while a *target* such as protein is a large (macro) one. Typically, in drugs, we are concerned with atoms and bonds, while in proteins, we are concerned with amino acids. The methods discussed in this section are summarized in Table 1.

2.1 Representing data

2.1.1 Molecular strings

The atoms and bonds of a small molecule can be efficiently represented as a string of ASCII characters. There are several ways to represent the molecule as the ASCII string. Examples of the sequence representation of the alanine molecule are presented in Table 2.

Chemical formula is a sequence representation showing the elements and their proportion in the chemical compound. Empirical formula is the simplest chemical formula that only presents the ratio of elements in the compound. The element is presented as the element symbol, while the ratio is presented as the subscript after the element symbol. Molecular formula is also a chemical formula presentation. Molecular formula is similar to empirical formula but shows the number of atom of each element instead of the elements' ratio. Another two chemical formula types are structural formula, which is the mixture of graphical and sequential representation of the molecular structure, and condensed formula. The structural formula shows how the atoms are organized in 2D space and connected through bonds. The condensed formula is similar to structural formula but the bonds information is omitted or limited to fit in a single line, which cause the structural information losses. The sequence chemical formula (empirical and molecular) is simple and human-readable. Because sequence chemical formula does not contain the chemical structure information, it has the identifiable problem in which it cannot distinguish two compounds sharing the same formula but having different molecular structures.

International Union of Pure and Applied Chemistry (IUPAC) nomenclature of organic chemistry is a convention molecule naming method proposed by the International Union of Pure and Applied Chemistry. Instead of using single letter like formula, IUPAC name uses word to represent elements and functional groups, which is easier for human to pronounce. IUPAC name for a compound may have different version due to the words' order. The Preferred IUPAC Name is a unique IUPAC assigned to each compound.

A popular representation is SMILES (Simplified Molecular-Input Line-Entry System) [11,37,38], which can be decoded back to the molecule structure graph. However, the SMILES string may not be unique as a molecule can have different SMILES forms [12]. Another issue is that SMILES sequence generation is not open-source project. Different SMILES sequence generation software may have different

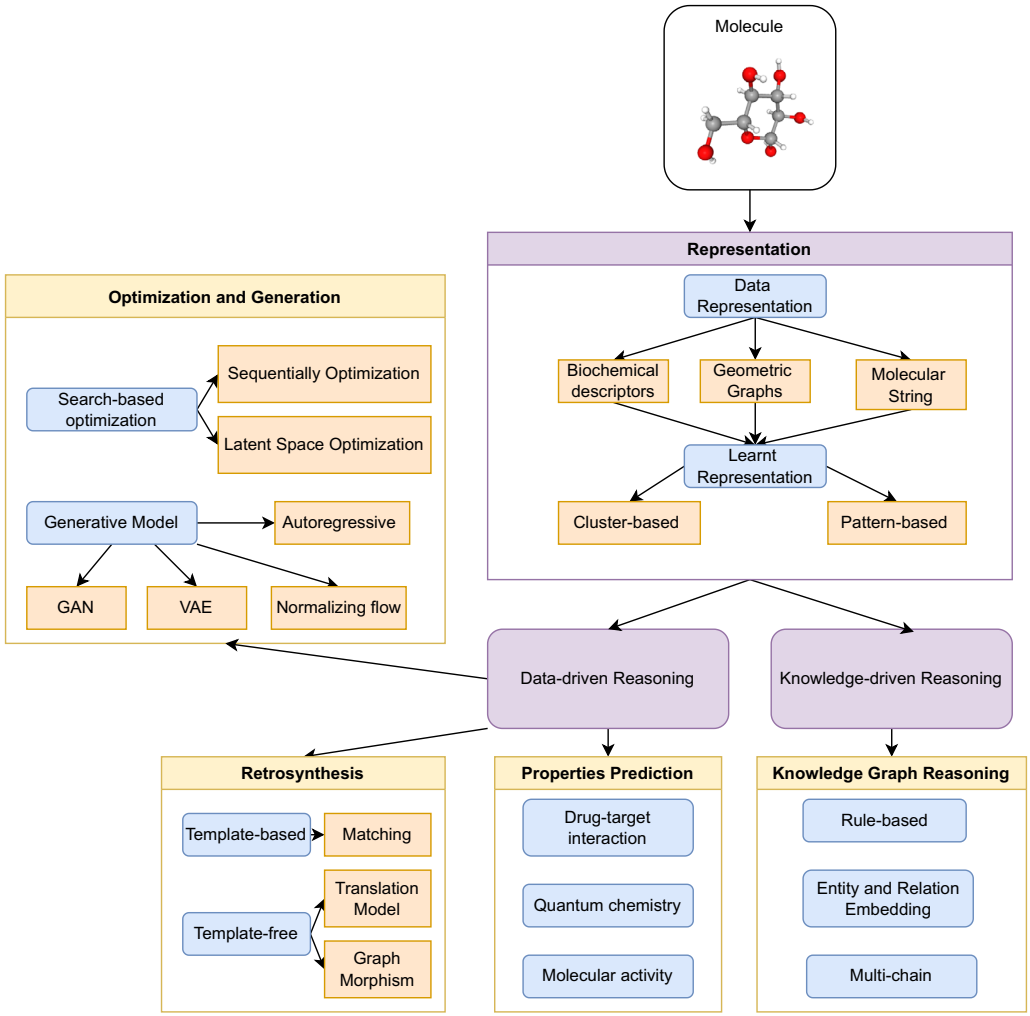

Fig. 1 Three aspects of AI in drug discovery: (i) Transforming the biological data into representations readable by computer; (ii) data-driven reasoning in which models estimated from data are used to infer properties, optimize, generate molecules and plan synthesis; and (iii) reasoning with biomedical knowledge graphs

generation algorithm, thus different SMILES sequences of the same compound. There are some efforts to canonicalize the SMILES sequence such as Universal SMILES [12] and Rdkit SMILES [39]. Validity is also an issue of the SMILES representation. Due to its complex grammar structure, a random SMILES sequence is likely to violate the syntax or physical-chemical constraints [22]. This is a challenging problem for deep learning molecule generative model as it requires the model to learn to generate a valid SMILES sequence. SELF-referencIng Embedded Strings (SELFIES) [22] tries to solve the invalidity problem. A single SELFIES sequence associates with a single valid molecule structure. This characteristic of the SELFIES sequence benefits the deep generative and optimization model. The model can focus on learning to generate the molecule with desired properties without having to learn the validity of generated molecule. However, the SELFIES representation shows no significant difference in prediction downstream task [27].

To overcome the unambiguous and identifiability problems of SMILES sequence, International Chemical Identifier (InChI) is a community-effort project. The main goal of InChI is that each molecule has one unique InChI sequence and each InChI sequence associates with only one molecule. InChI sequence is generated from the layers of molecule structure information. InChI encoding process starts with the main layer, which contains the molecular formula, the skeletal structure, followed by charge layer, stereochemical layer, isotopic layer, fixed-H layer, and reconnected layer. InChI is canonical which allows the unambiguous representation of different molecules. In the latest InChI version 1.06, the issue of two molecules sharing the same InChI string is not observed in the PubChem dataset test. The issue of a molecule having two InChI strings is observed in 547 molecules out of

Table 1 Drug and protein representation methods

Method	Type	Encoded information	Encoding method desgin	Learning objective
Chemical formula	Sequence	Proportion	Manual designed	N/A
Molecular formula	Sequence	Quantity	Manual designed	N/A
Structural formula	Sequence, Graphical	Structure	Manual designed	N/A
Condensed formula	Sequence	Structure	Manual designed	N/A
IUPAC	Sequence	Structure, Functional group type	Manual designed	N/A
MSA	Sequence	Evolutionary Information	Sequence alignment	N/A
SMILES [11], Universal SMILES [12], Rdkit SMILES [13]	Sequence	Structure	Manual designed	N/A
Sandberg et al. method [14]	Descriptor	Physical-chemical characteristic	Manual designed	N/A
ACF [15]	Descriptor	Hydrophobicity + AA frequency	Autocorrelation function	N/A
MACCS [16], PubChem Fingerprint [17]	Descriptor	Structure	Manual designed	N/A
InChI [18]	Sequence	Structure	Manual designed	N/A
Gao et al. method [19]	Descriptor	AA frequency	Manual designed	N/A
Daylight fingerprint [20]	Descriptor	Structure	Manual designed	N/A
ECFPs [21]	Descriptor	Structure	Manual designed	N/A
SELFIES [22]	Sequence	Structure	Manual designed	N/A
sPairs [23]	Descriptor	AA distance map	Manual designed	N/A
Mol2Vec [24], ProtVec [25]	Learnt Representation	Sequence	word2vec	Pattern: unsupervised predict the neighbor token
CNN-fingerprint [26]	Learnt Representation	Structure (fingerprint)	CNN	Pattern: supervised learning
ChemBERTa [27], ESM [28]	Learnt Representation	Sequence	Transformer	Pattern:unsupervised token masking
DeepCDA [29]	Learnt Representation	Sequence	CNN	Pattern: supervised learning
Grover [30]	Learnt Representation	Molecular structure graph	Transformer	Pattern: unsupervised subgraph masking
GCC [31]	Learnt Representation	Molecular structure graph	GIN	Cluster: subgraph similarity
GraphCL [32]	Learnt Representation	Molecular structure graph	GNN	Cluster: graph augmentation
MSA Transformer [33]	Learnt Representation	Evolutionary Information	Transformer	Pattern: MSA random masking
Graphormer [34]	Learnt Representation	Molecular structure graph	GNN, Transformer	Pattern: supervised pre-training
JOAO [35]	Learnt Representation	Molecular structure graph	GNN	Cluster: graph augmentation
HyperAttentionDTI [36]	Learnt Representation	Sequence	CNN, Attention	Pattern: supervised learning

Table 2 The sequence representation of the L-alanine

Representation	Example
Molecular formula	C3H7NO2
IUPAC name	(2S)-2-aminopropanoic acid
InChI	InChI=1S/C3H7NO2/c1-2(4)3(5)6/h2H,4H2,1H3,(H,5,6)/t2-/m0/s1
Canonical SMILES	CC(C(=O)O)N

over 111 million molecules, achieving 0.0005% error rate in PubChem dataset test. InChI representation is complex with multilayer of structure information. The deep learning model may not fully capture the grammar and syntax which leads to lower performance compared to SMILES sequence [40].

Likewise, a protein can also be represented by a string of characters varying in length. Each character represents one of the 20 amino acids. The evolutionary information (EI) is incorporated of the target sequence by searching for related proteins to form multiple sequence alignment (MSA) and extracting evolutionary information. MSA is the alignment of three or more biological sequences such as protein. Aligning the biological sequence reveals EI which tells similarity in structure and function as well as the origin and the evolution of the protein family [41]. The EI indicates the important substructure of the protein remains stable through evolution. The EI has shown its effectiveness in several tasks such as protein folding prediction [6].

Molecular string is a simple and easy-to-store data structure. The sequential representation allows that fast searching and indexing for large collection purpose. However, sequential representation is not a flexible tool to encode the molecule's information for machine learning framework. It is difficult to add or remove additional information such as bonds, atom's weight, and bond angles with sequential representation as the adjustments have to follow the sequence's rigid grammar and syntax. As a result, this limits the usage of molecular string in the ML model.

2.1.2 Geometric graphs

A richer and more precise representation of a molecule is attributed graph. A molecular graph is defined as $G = (V, E)$ where V is the atom set of the molecule and E is the edge set of bondings between atoms. To balance between 3D structural information and simplicity, 2D representation via an attributed graph can be used. For example, in the case of protein, the distance/contact between residues can be predicted [42,43] to form the contact/distance map. The contact/distance map is then used as an adjacency matrix of an attributed graph where each node represents a residue and edges represent the contact/distance between residues. Compared to sequential representation, the geometric graph offers a flexible tool to represent the molecule and biological entities. Additional information about the molecule can be encoded using node features and edge features. Geometric graph can adapt to a wide range of model designs and input data.

Indeed, many biomedical problems are well cast into graph reasoning: Molecule properties prediction as *graph classification/regression*, drug-target binding as *graph-in-graph*, chemical-chemical interaction as *graph-graph pairing*, molecular optimization as *graph edit/translation*, and finally chemical reaction as *graph morphism*.

2.1.3 Biochemical descriptors

For small molecules, *fingerprints* are often used to encode the 2D structure into a vector. One approach is using *structural keys* to encode the structure of the molecule into a bit string, each bit represents the presence or absence of predefined feature such as substructure or fragment [16,44]. The structural keys suffer from the lack of generalization because it depends on the pre-defined fragments and substructure to encode the molecule. The alternative approach is *hashed fingerprints* which encode the counting of molecular fragments into numeric values using a hash function, without relying on a pre-defined library. Based on the fragment enumerating process, the hashed fingerprints can be categorized into path-based [20] and circular types [21]. In [26], manual fingerprint extraction is replaced by a *learnable* hashing function on the convolution over the molecular graph.

For proteins, a set of descriptors are constructed based on the amino acids, their appearance frequency [19], their individuals [14] or autocorrelation [15] physical-chemical characteristic, and sequence-order feature derived from physicochemical distance. With the advance in the protein structure prediction, the distance between residues is also considered in protein description [23].

2.2 Learnt representation

Molecular strings, geometric graphs, and biomedical descriptors are manually designed. These representations are designed to compress the molecule information while trying to preserve its structure and identifiability in the features. However, it is difficult to ensure that these features are useful for the target tasks. Deep learning framework can automat-

ically learn the feature needed for the target task from the input representation.

The string representation of molecules makes it ready for applying language modeling techniques, assuming the existence of statistical sequential patterns, akin to those found in linguistic grammars. Because the structure of atoms in the molecule follows a set of rules such as valence which can be viewed as grammar rules in chemical language. This is not limited to string representation and can be extended to other types of representation such as graphs. Several unsupervised representation learning models have exploited the structural patterns to learn the molecule representations. An early work in sequential representation learning, word2vec [45], learns the representation by using the predicting neighbor tokens as a self-supervised task. Inspired by word2vec, Mol2Vec [24] and ProtVec [25] view substructure in molecule and residues in protein sequence as word in sentence to learn the molecule and protein representation in an unsupervised manner. Word2vec is also used to represent other biological entities such as ion channels [46] or RNA [47]. Recently, Transformers with BERT-like masking strategy [7] has become a popular technique to learn sequential representations, e.g., out of molecular SMILES sequence [27] or protein sequence [28]. The Transformer-based biological sequence representation learning methods mask random tokens (residues or atoms) in the sequence. Then the models are trained by predicting the masked tokens. The main difference between Transformer-based representation is token unit of the biological sequence, which can be single symbol of SMILES sequence [27], single residue of the protein sequence [28], evolutionary information through MSA [33], or molecule and protein substructure [48]. Both word2vec and Transformer-based sequential representation learning can take advantages the large unlabeled biological sequence dataset such as UniRef [49], ZINC [50], and ChEMBL [51]. Convolution neural network (CNN) is a popular network architecture in computer vision. CNN can learn the image local feature via kernel, window which can be applied to molecule substructure feature. The latent vector of sequence input can be learnt using 1D CNN [29,36].

Likewise, the graph representation of molecules allows us to learn the graph structure patterns. For example, DeepWalk [52] learns the node representation of the given graph structure using random walks to learn the pattern of nodes' neighbor. Grover [30] uses subgraph masking as contextual properties prediction and graph motif prediction. Given a molecule graph, the contextual property of a masked atom node v are its neighbor's atom type and edge type. The molecular graph structure allows the DL to use graph representation learning network such as GCN, GAT, and GIN [53]. Conventional GNN such as GCN, GAT, and GIN has the receptive field limited to neighbor nodes. To improve the graph-level representation for graph-level prediction task, Graphormer [34] uses Transformer to have global receptive field.

An increasingly popular strategy is through exploring the local structure in the chemical space by using *contrastive learning* for estimating representations. This works by minimizing an energy-based loss to keep the distance in the embedding space small for similar molecules and large for dissimilar pairs [31]. The main difference between these contrastive losses is the number of positive, negative samples and how the pairs are sampled. The similar molecule pairs are generated by graph augmentation [32] with node dropping, edge perturbation, attribute masking, and subgraph. The augmentation can be tuned manually by ad hoc rules [32] or automatically [35]. However, the molecule graph augmentation should be designed carefully as single change in bond can dramatically change the identity and validity of the molecule [32].

3 Data-driven reasoning

There are three basic questions in drug discovery. The first question is determining **whether the given molecule is drug-like**, meaning having therapeutic effects on druggable targets. The second question is given a biological target, **what are the candidate compounds** that are likely to modify their functions in a desired way? This question addresses two sub-problems: searching and generating. The former is about finding a suitable molecule from an approved list, and the latter is generating a *de novo* molecule tailored to the target. The final question is given a molecule, **how can we make it**? This question addresses the sub-problems of synthetic tractability, reaction planning, and retrosynthesis. The methods in this section are summarized in Table 3

3.1 Molecular property prediction

The first question is predicting the molecule's properties. There are a wide range of prediction tasks, from predicting the drug-likeness, which targets it can modulate, to predicting molecule dynamics/kinetics/effects/metabolism if administered orally or via injection. Molecule property prediction is a fundamental task in many stages of drug discovery.

The task is a specific case of many-body systems where the emergent properties of a group of interacting objects are predicted. The most popular first-principle technique to tackle this general problem is Density Functional Theory (DFT) to approximate the wave function, which describes the quantum state of an isolated quantum system in the many-body system. However, calculating DFT is computationally expensive and can take up to $\mathcal{O}(10^3)$ seconds for a medium-sized molecule, making rapid screening over millions of potential candidates intractable.

Table 3 Data-driven reasoning method in drug discovery

Method	Task	Approach
IterRefLSTM [54]	Molecular properties	LSTM, GCN
enn-s2s [55]	Quantum-chemistry properties	MPNN with atom's feature
D-MPNN [56]	Molecular properties, Drug-target interaction	MPNN with descriptors
SpookyNet [57]	Quantum-chemistry properties	MPNN with atom's states
KronRLS [58]	Drug-target affinity	Kernel function
SimBoost [59]	Drug-target affinity	Similarity matrix
Pafnucy [60]	Drug-targetaffinity	3D convolution
DeepDTA [61]	Drug-target affinity	1D CNN
GraphDTA [53]	Drug-target affinity	GNN, 1D CNN
DGraphDTA [62]	Drug-target affinity	GNN
Drug-VQA [63]	Drug-target interaction	2D CNN, BiLSTM, Attention
GraphDTA [53]	Drug-target affinity	GNN
GEFA [64]	Drug-target affinity	GCN, Attention, Pretrain Transformer
DockTScore [65]	Drug-target affinity	Docking empirical SF, RF, MLR
IIFDTI [66]	Drug-target interaction	GAT, CNN, Attention
GCPN [67]	Molecule optimization	RL (Policy-based)
JT-VAE [68]	Molecule optimization	Junction tree VAE, BO
Rafael et al. [40]	Molecule optimization, Molecular generation	RNN-based autoencoder, BO
MolDQN [69]	Molecule optimization	RL (DQN)
MSO [70]	Molecule optimization	RNN-based autoencoder, Genetic algorithm
FREED [71]	Molecule optimization	RL (Soft-Actor-Critic), Fragment-based, Docking Score
GrammarVAE [72]	Molecular generation	Parse tree, 1D CNN, logic rules
GraphVAE [73]	Molecular generation	GraphVAE with GCN encoder
GraphRNN [74]	Molecular generation	Autoregressive RNN
Mol-CycleGAN [75]	Molecular generation	CycleGAN
GraphAF [76]	Molecular generation	Normalizing flow
Retrosynthesis DHN [77]	Retrosynthesis	Templated-based with Deep Highway Network
GLN [78]	Retrosynthesis	Templated-based with ConditionalGraph Logic Network
Karpov et al. method [79]	Retrosynthesis	Templated-free, Transformer seq2seq
G2Gs [80]	Retrosynthesis	Templated-free, RL graph2graph

A recent approach is to approximate DFT calculations by learning a graph neural network (GNN) over the molecular graphs, which can be trained on a large pre-computed dataset. Once trained GNNs can run many orders of magnitudes faster than precise DFT methods with reasonable accuracy. A popular type of GNNs is the Message Passing Neural Network (MPNN) [55,57] which models the atoms interaction with message passing function, update function, and readout function. MPNN and its cousin, the Graph Convolutional Network (GCN), have since been frequently used in predicting physical chemistry properties (e.g., water solubility, hydrophobicity), physiology (e.g., toxicity), and biophysics (bind affinity) [53,54,56]. Alternatively, the molecule properties prediction can also be formulated as a reasoning process to answer a query (of a specific property), and thus lends itself to more elaborate neural networks such as Graph Memory Networks (GMN) [81].

3.2 Drug-target affinity prediction

Drug-target binding affinity indicates the strength of the binding force between the target protein and its ligand (drug or inhibitor) [82]. There are two main approaches: the structural approach and the non-structural approach [83]. Structural methods utilize the 3D structure of proteins and ligands to run the interaction simulation between proteins and ligands. On the other hand, the non-structural approach relies on ligand and protein features such as sequence, hydrophobic, similarity and other structural information to construct the molecule graph or learn the contextual relationship between atoms and residues [62,64].

Structural approach The structure-based approach usually relies on molecular docking which simulates the post-binding 3D conformation of drug-target complex. As there are several possible conformations, the simulated structure is evaluated using a scoring function (SF). The scoring function can vary from the molecular mechanics' interaction energies [84], empirical scoring function using van der Waals and electrostatic energy terms [65], to machine learning predicted value derived from protein and drug features [85,86], or 3D convolution on 3D structure [60]. Pafnucy [60] uses 3D convolution to learn 3D protein-ligand complex representation. Each atom is represented by its atom types, hybridization, number of bonds, partial charge, and molecule type (ligand or protein).

Non-structural approach The non-structural approach relies on the drug/target similarity, and structural features such as protein sequence or secondary structure without relying on calculating the exact 3D structure of the drug-target complex. Popular among them are kernel-based methods which employ kernel functions to measure the molecule similarity [58]. Alternatively, drug-drug, target-target, and drug-target similarity features are used as input for a classifier/regressor [59]. More recently neural networks have become common as they can learn the drug and target representations instead of handcrafting them. For sequences, 1D convolution [61], BiLSTM [63], or language model feature [64] are used to encode the biological sequence to the latent space. The drawback of sequential features is that they ignore the structural information of the drug and the target which also plays a critical role in the drug-target interaction. Thus, graph neural networks have been applied whenever graph structures are available [62,64,87]. More elaborate techniques use self-attention to model the residue-atom interactions between drug and protein [63,64,66]. IIFDTI [66] uses word2vec and CNN to learn protein representation and GAT [88] to learn molecule graph representation. The drug-protein interaction is learnt using two parallel Transformer-based encoder-decoder.

3.3 Molecular optimization and generation

The second question is what kind of molecules interact with the given target. The traditional combinatorial chemistry approach uses a template as a starting point. From this template, a list of variations is generated with the goals that they should bind to the pocket with good pharmacodynamic, have good pharmacokinetics, and be synthetically accessible. The space of drugs is estimated to be 10^{23} to 10^{80} substances but only 10^{8} substances have been synthesized thus far. Thus, it is practically impossible to model this space fully. The current techniques for graph generations can be search-based, generative, or a combination of both approaches. The search-based approach starts with the template and uses optimization framework such as Bayesian Optimization to improve it over time. This approach does not require a large amount of data but demands a reliable evaluator through expensive computer simulation or lab experiments. A more ambitious approach is building expressive generative models of the entire chemical space, and thus it requires a large amount of data to train.

3.3.1 Search-based optimization

The search-based approach can be formulated as structured machine translation. Search-based methods search for an inverse mapping of the knowledge base and binding properties back to query molecules. In this approach, the template molecule is represented as a graph or a string. The starting molecule is optimized toward desirable properties. There are two common strategies for optimization. The first strategy is *sequential optimization in the discrete chemical space* via atom/bond addition/deletion while maintaining the validity of the molecule. This sequential discrete search fits well to the reinforcement learning frameworks with target molecule properties as rewards [69]. Reinforcement learning can cooperate with the graph representation of the molecule with a graph policy network [67]. FREED [71] explicitly restricts the generation space by using valid fragment from the library instead of implicitly restriction such as QED [89], reducing the invalid molecule space. FREED uses the docking score as the reward function for more straightforward proxy function to estimate the drug efficacy.

The second strategy is *continuous optimization in the latent representation space*. First, the input molecule graphs or strings are encoded into the latent space. The encoder architecture depends on the input molecule representation, varying from sequence-based encoder (e.g., RNN [40] or molecule graph junction tree [68]). To have an embedding space representing a set of specific properties, the encoder is jointly trained with the property prediction task [40]. Then the molecule is optimized in the latent space with Bayesian Optimization (BO) [68] or genetic algorithms [70] before being decoded back to the original molecule space. The bottleneck of this approach is in accurate modeling of the drug latent space.

3.3.2 Generative molecular generation

The molecule optimization can be viewed as inverse function learning where the function that maps the desired outputs to the target structure is learned. Then the generative models can leverage the existing data and query the simulators in an offline manner. In particular, the generative models start with randomly sampled structure x variable. Then the simulators answer the query structure x with the properties y. With a sufficiently large number of (x, y) pairs, the machine learning can learn the inverse function $x \approx g(y)$.

The core idea of generative models is learning and sampling from the density function $p(x)$ of the training data. The main challenges are due to the complexity of the discrete molecular space, unlike those typically seen in continuous domains like computer vision. The most popular generative models to date are variational autoencoder (VAE), generative adversarial networks (GAN), autoregressive models and normalizing flow models.

VAE is a two-stage process: the visible input structure is first encoded into the hidden variable and then decode back to the original structure. The first VAE implemented in modeling the drug space was by mapping the SMILES sequence into the vector space [40]. Then the vector space is explored by optimization methods such Bayesian Optimization (BO) [40] or genetic algorithms [70]. GraphVAE [73] operates directly on the expressive graph representations. Since the iterative generation of discrete structure such as graph is non-differentiable, GraphVAE models the decoded graph as probabilistic fully connected on the restricted k-node domain. Then the decoded graph is compared with the ground truth by a standard graph matching. The main drawbacks of searching in the latent space of VAEs is that it cannot explore the low density regions, where most interesting novel compounds reside. A more intrinsically explorative strategy is through compositionality, where novel combinations can be generated once the compositional rules are learnt. This has been studied under GrammarVAE [72], an interesting method that imposes a set of SMILES grammar rules via parse trees to ensure the validity of the SMILES sequence.

GAN is a powerful alternative to VAEs as it does not require an encoder, and hence it models the compound distribution *implicitly*. GAN has two sub-models: a discriminator and a generator. The discriminator determines whether any two samples come from the same distribution. The generator learns the to generate good samples by trying to fool the discriminator to believe that the generated samples are real training data. Mol-CycleGAN [75] learns the mapping function $G : X \rightarrow Y$ and $G : Y \rightarrow X$ with two discriminators D_X and D_Y where X is the set of input molecule and Y is the molecule set with desired properties. This ensures the generator transform input molecule to the desired properties while retaining the structure.

Autoregressive models factorize the density function $p(x)$ as: hence allowing generation of molecules in a stepwise manner. GraphRNN [74] encodes a sequence of graph states using RNN. Each state represents a step in the graph generation process. GraphRNN uses BFS to reduce the complexity of learning all the possible graph state sequences.

Normalizing flow models explicitly learn the complex density function by transforming the simple distribution through a series of invertible functions. GraphAF [76] defines an invertible function mapping the multivariate Gaussian distribution to a molecular graph structure. Each step of molecule generation samples random variables to map them to atom/bond features.

3.4 Retrosynthesis

The third question is given a molecule graph, how can the target molecule be synthesized? The problem is known as retrosynthesis planning, and it involves determining a chain of reactions to finally synthesize a target molecule with high efficiency and low cost. At each reaction step, a set of reactants needs to be identified for an intermediate molecule. This problem can be viewed as the reverse of chemical reaction prediction. Normally, chemical reaction prediction is predicting the post-reaction products of two or more molecules. However, in retrosynthesis, given the post-reaction product, the task is to search for two or more feasible candidates for chemical reaction. Both reaction prediction and retrosynthesis can be cast as *graph morphism,* where the molecules form a graph of disconnected sub-graphs, each of which is a molecule. Reaction changes the graph edges (dropping bonds and creating new bonds) but keeps the nodes (atoms) intact. A learnable graph morphism was introduced in GTPN [90], a reinforcement learning-based technique to sequentially modify the bonds.

There are two main approaches to solve the retrosynthesis problem: template-based and template-free. The **template-based** approach relies on the set of predefined molecules to construct the target molecules. This approach formulates the retrosynthesis as the subgraph matching problem to match the template to the target molecule. The matching problem is then solved by a variety of techniques, ranging from a simple deep neural network [77] to a more sophisticated framework such as conditional graphical models [78]. The template-based approach suffers from poor generalization on unseen structures as it relies on predefined fragments and template libraries.

The **template-free** approach is proposed to overcome the poor generalization of the template-based approach by inheriting the strong generalization from the (machine) translation model. The Transformer model can effectively solve the retrosynthesis problem formulated as sequence-to-sequence, in which the product molecule SMILES sequence is translated into a set of reactants SMILES sequences [79]. Graph-to-graph is another approach [80], in which at first the reaction center is identified using edge embedding of the graph neural network to break the target molecule into synthons. Then the synthons are translated into reactants using graph translation model.

4 Knowledge-based reasoning

The biomedical community has accumulated a vast amount of domain knowledge over the decades, among them those structured as knowledge graphs are the most useful for learning and reasoning algorithms. We are primarily interested in the knowledge graphs that represent the relationships between biomedical entities such as drug, protein, diseases, and symptoms. Examples of manually curated databases are OMIM [91] and COSMIC [92]. Formally a knowledge graph is a triplet $\mathcal{K} = \langle H, R, T \rangle$ where H and T are the set of entities, R is the set of labeled relationship edge connecting entities of H and T. Entities of H and T can have single or multiple relation, directed or undirected relation.

Biomedical knowledge graphs enable multiple graph reasoning problems for drug discovery. Among them is *drug repurposing*, which aims to find novel uses of existing approved drugs. This is extremely important when the demands for new diseases are immediate, such as COVID-19, when the market is too small to warrant a full *de novo* costly development cycle (e.g., rare, localized diseases). Given a knowledge graph, the drug repurposing is searching for new links to a target from existing drug nodes—a classic *link prediction* problem. This setup is also used in *gene-disease prioritization* in which the relationship between diseases and molecular entities (proteins and genes) is predicted [93]. Another reasoning task is *polypharmacy prediction* of the adverse side effects due to the interaction of multiple drugs. The multi-relation graph with graph convolution neural network can encode the drug-drug interactions [94]. Given a pair of drugs, the drugs are embedded using the encoder and the polypharmacy prediction task is formulated as a link prediction task.

In what follows, we briefly discuss two major AI/ML problems: *graph construction* and *graph reasoning*.

4.1 Automating biomedical knowledge graph construction

Biomedical knowledge graph is constructed using existing databases or a rich source of data from biomedical publications. As manual literature curation is time-consuming, ML has been applied to speed up the process. The usual framework starts with relevant sentences filtering, followed by biomedical entity identification and disambiguation [95]. The biomedical entities relationships are extracted from selected text using rule-based method [96], unsupervised [97], or supervised manner [98]

4.2 Reasoning on biomedical knowledge graphs

Reasoning on knowledge graphs is the process of inferring the relationship between a pair of entities and the logic behind the relationship. Machine learning reasoning applying to this problem can be categorized into rule-based reasoning, embedding-based reasoning, and multi-chain reasoning. The methods discussed in this section are summarized in Table 4.

4.2.1 Rule-based reasoning

Rules-based reasoning uses logic rules or ontology to infer the new triplet from the knowledge graph KG. A logic rule is defined by its head $\mathcal{H}$ and body $B = \{B_1, B_2, ..., B_n\}$:

$$\mathcal{H} \leftarrow B_1 \wedge B_2 \wedge ... \wedge B_n \tag{1}$$

AMIE [106] explores the knowledge graph with the mining scheme similar to association rule mining. Ontology is the formal way to describe the types, categories of entities' structure. Web ontology language (OWL) is a logic-based language to describe the entities and their relationship. OWL can apply to complex structure like biomedical knowledge graphs [99]. OWL is used to discover the associations between traditional Chinese medicine and western medicine from the large and complex knowledge graph compiled from multiple database such as UniProt [107], DrugBank [108], PubMed [1], and Pfam [109] using Ontotext platform [2].

4.2.2 Entity and relation embedding

The logic-based reasoning suffers from the lack of generalization. A more robust technique assumes a distributed representation of entities and relations, typically as embedding vectors in high-dimensional spaces. **Matrix/tensor factorization** projects the high-dimensional/multi-way objects into multiple low dimensional vectors. TriModel [101] learns a low-rank vector representation Θ_E and Θ_R of knowledge entities $\mathbb{E} = \{H \cup T\}$ and relations R. The graph embedding encoder is trained using tensor factorization where each entity is represented by three embedding vectors. The TriModel is evaluated on the Yamanishi_08 [110], DrugBank_FDA [108], and KEGG-based drug targets dataset [101]. The **Distance-based models** exploit the fact that given a triplet (h, r, t), the embedded representation of h and t is in the proximity, translated by the embedded vector of the relationship r. The best known model TransE [111] learns the embedding of entities by minimizing the distance between $h+r$ and t and maximizing the distance between the $h+r$ and t' where (h, r, t') triplet does not hold. The learned embedding of entities can be used directly as drug and protein representation to predict the drug-target interaction [105] or

[1] http://www.pubmedcentral.gov/.

[2] https://www.ontotext.com/products/ontotext-platform/.

Table 4 Knowledge-based reasoning methods in drug discovery

Method	Task	Approach
OWL [99]	Association rules mining	Rule-based reasoning
MINERVA [100]	Drug-target interaction, Drug repositions	Multi-chain reasoning, Policy-based RL
TriModel [101]	Drug-target interaction	Entity and relation embedding with tensor factorization
BioKGLM [102]	Named entity recognition, Relation extraction, Event extraction	Transformer, Distance-based entity and relation embedding
KGE_NFM [103]	Drug-target interaction	Distance-based entity and relation embedding with structural embedding
PoLo [104]	Drug-target interaction, Drug repositions	Multi-chain reasoning, Policy-based RL with logic path
KG-DTI [105]	Drug-target interaction, Drug repositions	Distance-based entity and relation embedding

further injected to the language model to enhance the contextual relationship [102].

Structural information like 2D structure or 3D conformation is also helpful for entity representation learning. It is necessary to integrate heterogeneous information with structural information. The knowledge graph embedding can be combined with the structural embedding using neural factorization machine to form the hybrid representation [103]. The results from Luo's dataset [112], Hetionet [113], Yamanishi_08's dataset [110] and BioKG [114] demonstrate the advantages of the combination of knowledge graph and structural embedding compared to embedding of knowledge embedding.

4.2.3 Multi-chain reasoning

Shallow embedding has achieved remarkable results in reasoning over the biomedical graphs. However, they can fail to reason when presented with multiple complex relationships. Multi-chain reasoning extends the reasoning from a triplet to an extended path of reasoning chain. DeepPath [115] applies reinforcement learning (RL) with fully connected layers as the policy network to find the optimal path of reasoning in the knowledge graph. MINERVA [100] uses LSTM [116] as the policy network instead of fully-connected layers. The RL can be combined with pre-defined logic rules to learn the drug repurposing to achieve explainable reasoning [104]. The experiments [117] conducted on the Hetionet [113], BIKG Hetionet, and BIKG Hetionet+ [117] show the advantages of the multi-chain reasoning methods [100,104] on the drug repurposing task, while knowledge graph entity embedding TransE [111] performs well on the drug-target interaction task. Multi-chain reasoning provides the interpretability to the reasoning model by showing the traversing path of the model through the knowledge graph. The domain expert can verify the reasoning logic of the model, thus making it more trustworthy.

5 Challenges and opportunities

We are now in a position to discuss remaining challenges and chart possible courses to overcome them.

5.1 Large biomedical space

The drug space is estimated to be from 10^{23} up to 10^{60}. Due to the diversity of the molecules in terms of function and structure, and the combinatorial nature of their interaction, unconstrained exploration of the biomedical space—such as molecule optimization and generation—is intractable. Search-based optimization requires an accurate predicting model which maps the generated molecule to the target properties. The generated molecule may have undiscovered properties which leads to an inaccurate predicting model. As a result, the search direction can be misleading. Human implicit and explicit feedback can assist and redirect the optimization to the correct course.

5.2 Data quality

Poor data quality will have a snowball effect in the multi-stage process of discovery. The data error may lead to an inaccurate machine learning model when trained on insufficient data. Factors affecting the data quality are data entry error, hidden bias, and incompleteness due to law and regulations. Machine learning techniques can help enhance the data quality by detecting and removing the hidden bias in the early stage of data collection, data pre-processing, or considering the bias in the model design. One promising direction is to develop a "foundation model" trained on large-scale data and then adapted to a wide range of relevant downstream tasks, similar to what is happening in the space of text and vision [118].

5.3 Large gap between virtual screening and real clinical trials

There is a large gap between clinical trial results and *in silico* results [119]: Clinical trials can fail despite excellent model prediction. For example, machine learning only predicts the interaction between a drug and a protein without factoring in a chain reaction or off-target interaction that reduces the effectiveness of the drug. It is necessary to have a drug discovery framework that takes account of multiple and chain drug-target, drug-drug, and protein-protein interactions.

5.4 Drug effect on the protein functions

The current drug discovery and optimization work on the binding interaction between the target protein and drug molecule. The machine learning molecule generation and optimization work on the principle of targeting a specific set of properties or proteins. The machine learning framework tries to generate or optimize a molecule that is likely to fit to the binding pocket of the protein. However, there is no clear connection between the binding activity predicted by the machine learning framework and the target protein function change. This opens up the direction to cooperate the protein function information from other sources like literature into the optimization model.

5.5 Personalized prescription and drug discovery

Personalized medicine allows efficient and safe treatment by coursing the treatment based on the patient's genomic environments. With the advance in the 3D printing techniques in pharmaceutics [120], a patient-tailored drug delivery system allows safe and efficient usage of drugs. At the same time, with the development of bio-markers in both clinical and biomedical data, the information from bio-markers is getting integrated into the drug discovery loops. From the machine learning point of view, it presents a challenge as well as an opportunity in personalized medicine and drug discovery systems. With the advance in generative models and optimization, the machine learning framework can combine bio-maker data with the high-speed drug screening, optimization and printing techniques to develop a personalized drug discovery system.

5.6 Efficient human-machine co-creation

The end goal of the drug discovery process and the intermediate goal of machine learning systems may not align due to undiscovered knowledge. Having an efficient human-machine ecosystem allows the domain experts to inject prior knowledge, verify and discover the underlying mechanism.

6 Conclusion

We have provided a survey on recent AI advances targeting one of the most impactful areas of our time: drug discovery. While this is a very challenging task, the rewards are huge, and AI is already making solid progress, contributing to the saving of development costs, and speed up the discovery. Reversing Eroom's Law will demand new fundamental advances in AI itself, from learning in the low-data regime, to explore the vast molecular space, to sophisticated reasoning, to robotic automation. AI will need to work alongside humans and help expand the knowledge bases and then benefit from it.

Acknowledgements None.

Author Contributions All authors contribute in writing and reviewing the manuscript equally.

Funding No funding to declare.

Declarations

Conflict of interest All authors have no conflict of interest to report.

References

1. Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B.: et al. The AI Index 2021 Annual Report. (2021) arxiv: 2103.06312
2. Scannell, J.W., Blanckley, A., Boldon, H., Warrington, B.: Diagnosing the decline in pharmaceutical R& D efficiency. Nat. Rev. Drug Discov. **11**(3), 191–200 (2012). https://doi.org/10.1038/nrd3681
3. Yang, X., Wang, Y., Byrne, R., Schneider, G., Yang, S.: Concepts of artificial intelligence for computer-assisted drug discovery. Chem. Rev. **119**(18), 10520–10594 (2019). https://doi.org/10.1021/acs.chemrev.8b00728
4. The Drug Development Process. https://wwwfdagov/patients/learn-about-drug-and-device-approvals/drug-development-process;
5. Wang, R., Fang, X., Lu, Y., Yang, C.Y., Wang, S.: The PDBbind database: methodologies and updates. J. Med. Chem. **48**(12), 4111–4119 (2005). https://doi.org/10.1021/jm048957q
6. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al.: Highly accurate protein structure prediction with AlphaFold. Nature **596**(7873), 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2
7. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, AN.: et al. Attention is all you need. In: Proceedings of Advances in Neural Information Processing Systems. California, USA; 2017. p. 5998–6008
8. Berdigaliyev, N., Aljofan, M.: An overview of drug discovery and development. Fut. Med. Chem. **12**(10), 939–947 (2020). https://doi.org/10.4155/fmc-2019-0307
9. Vijayan, R.S.K., Kihlberg, J., Cross, J.B., Poongavanam, V.: Enhancing preclinical drug discovery with artificial intelligence. Drug Disc. Today. **27**(4), 967–984 (2022). https://doi.org/10.1016/j.drudis.2021.11.023

10. Deng, J., Yang, Z., Ojima, I., Samaras, D., Wang, F.: Artificial intelligence in drug discovery: applications and techniques. Brief. Bioinform. (2012). https://doi.org/10.1093/bib/bbab430
11. Weininger, D.: SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J. Chem. Inf. Model. **28**(1), 31–36 (1988). https://doi.org/10.1021/ci00057a005
12. O'Boyle, N.M.: Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI. J. Cheminf. **4**(1), 22 (2012). https://doi.org/10.1186/1758-2946-4-22
13. RDKit: cheminformatics and machine learning software. Available from: http://www.rdkit.org
14. Sandberg, M., Eriksson, L., Jonsson, J., Sjöström, M., Wold, S.: New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. J. Med. Chem. **41**(14), 2481–2491 (1998). https://doi.org/10.1021/JM9700575/SUPPL_FILE/JM2481.PDF
15. Feng, Zhi-Ping., Zhang, Chun-Ting.: Prediction of membrane protein types based on the hydrophobic index of amino acids. J. Prot. Chem. **19**(4), 269–275 (2000). https://doi.org/10.1023/A:1007091128394
16. Durant, J.L., Leland, B.A., Henry, D.R., Nourse, J.G.: Reoptimization of MDL keys for use in drug discovery. J. Chem. Inform. Comp. Sci. **42**(6), 1273–1280 (2002). https://doi.org/10.1021/CI010132R
17. PubChem Substructure Fingerprint. Available from: ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.pdf
18. Heller, S.R., McNaught, A., Pletnev, I., Stein, S., Tchekhovskoi, D.: InChI, the IUPAC international chemical identifier. J. Cheminf. **7**(1), 23 (2015). https://doi.org/10.1186/s13321-015-0068-4
19. Gao, Q.B., Wang, Z.Z., Yan, C., Du, Y.H.: Prediction of protein subcellular location using a combined feature of sequence. FEBS lett. **579**(16), 3444–3448 (2005). https://doi.org/10.1016/J.FEBSLET.2005.05.021
20. Daylight Theory: fingerprints. Available from: https://www.daylight.com/dayhtml/doc/theory/theory.finger.html
21. Rogers, D., Hahn, M.: Extended-connectivity fingerprints. J. Chem. Inf. Model. **50**(5), 742–754 (2010). https://doi.org/10.1021/CI100050T
22. 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. **1**(4), 045024 (2020). https://doi.org/10.1088/2632-2153/aba947
23. Xu, Y., Verma, D., Sheridan, R.P., Liaw, A., Ma, J., Marshall, N.M., et al.: Deep dive into machine learning models for protein engineering. J. Chem. Infor. Model. **60**(6), 2773–2790 (2020). https://doi.org/10.1021/ACS.JCIM.0C00073/SUPPL_FILE/CI0C00073_SI_001.PDF
24. Jaeger, S., Fulle, S., Turk, S.: Mol2vec: unsupervised machine learning approach with chemical intuition. J. Chem. Inf. Model. **58**(1), 27–35 (2018). https://doi.org/10.1021/acs.jcim.7b00616
25. Asgari, E., Mofrad, M.R.K.: Continuous distributed representation of biological sequences for deep proteomics and genomics. PLOS one **10**(11), e0141287 (2015). https://doi.org/10.1371/journal.pone.0141287
26. Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., Gómez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A.: et al. Convolutional networks on graphs for learning molecular fingerprints. In: Proceedings of advances in neural information processing systems. Montreal, Canada; 2015. p. 2224–2232. Available from: https://arxiv.org/abs/1509.09292v2
27. Chithrananda, S., Grand, G., Ramsundar, B.: ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. In: Machine learning for molecules workshop, NeurIPS. Online; 2020. Available from: https://arxiv.org/abs/2010.09885v2
28. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., et al.: Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceed. Nat. Acad. Sci. (2021). https://doi.org/10.1073/pnas.2016239118
29. Abbasi, K., Razzaghi, P., Poso, A., Amanlou, M., Ghasemi, J.B., Masoudi-Nejad, A.: DeepCDA: deep cross-domain compound-protein affinity prediction through LSTM and convolutional neural networks. Bioinformatics **36**(17), 4633–4642 (2020). https://doi.org/10.1093/bioinformatics/btaa544
30. Rong, Y., Bian, Y., Xu, T., Xie, W., Wei, Y., Huang, W., et al.: Self-supervised graph transformer on large-scale molecular data. In: Proceedings of advances in neural information processing systems. Online; 2020. Available from: https://arxiv.org/abs/2007.02835v2
31. Qiu, J., Chen, Q., Dong, Y., Zhang, J., Yang, H., Ding, M., et al.: GCC: Graph contrastive coding for graph neural network pre-training. In: Proceedings of the international conference on knowledge discovery & data mining. vol. 20. San Diego, CA, USA; 2020. p. 1150–1160. Available from: https://dl.acm.org/doi/10.1145/3394486.3403168
32. You, Y., Chen, T., Sui, Y., Chen, T., Wang, Z., Shen, Y.: Graph contrastive learning with augmentations. In: Proceedings of advances in neural information processing systems. Online; 2020. Available from: https://github.com/Shen-Lab/GraphCL
33. Rao, RM., Liu, J., Verkuil, R., Meier, J., Canny, J., Abbeel, P., et al.: MSA Transformer. In: Proceedings of the international conference on machine learning. PMLR; 2021. p. 8844–8856. Available from: https://proceedings.mlr.press/v139/rao21a.html
34. Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., et al.: Do transformers really perform bad for graph representation? In: Proceedings of the advances in neural information processing systems; 2021. Available from: https://github.com/Microsoft/Graphormer
35. You, Y., Chen, T., Shen, Y., Wang, Z.: Graph contrastive learning automated. In: Proceedings of the international conference on machine learning; 2021. p. 139. Available from: https://github.com/
36. Zhao, Q., Zhao, H., Zheng, K., Wang, J.: HyperAttentionDTI: improving drug-protein interaction prediction by sequence-based deep learning with attention mechanism. Bioinformatics **38**(3), 655–662 (2022). https://doi.org/10.1093/bioinformatics/btab715
37. Weininger, D., Weininger, A., Weininger, J.L.: SMILES. 2. Algorithm for generation of unique SMILES notation. J. Chem. Inf. Comp. Sci. **29**(2), 97–101 (1989). https://doi.org/10.1021/ci00062a008
38. Weininger, D.: SMILES. 3. DEPICT. Graphical depiction of chemical structures. J. Chem. Inf. Model. **30**(3), 237–243 (1990). https://doi.org/10.1021/ci00067a005
39. Schneider, N., Sayle, R.A., Landrum, G.A.: Get Your Atoms in Order-An Open-Source Implementation of a Novel and Robust Molecular Canonicalization Algorithm. J. Chem. Inf. Model. **55**(10), 2111–2120 (2015). https://doi.org/10.1021/acs.jcim.5b00543
40. Gómez-Bombarelli, R., Wei, J.N., Duvenaud, D., Hernández-Lobato, J.M., Sánchez-Lengeling, B., Sheberla, D., et al.: Automatic chemical design using a data-driven continuous representation of molecules. ACS Centr. Sci. **4**(2), 268–276 (2018). https://doi.org/10.1021/acscentsci.7b00572
41. Sofi, M.Y., Shafi, A., Masoodi, K.Z.: Bioinformatics for everyone. Elsevier, Hobroken (2022)
42. Wang, S., Sun, S., Li, Z., Zhang, R., Xu, J.: Accurate de novo prediction of protein contact map by ultra-deep learning model. PLOS Computat. Biol. **13**(1), e1005324 (2017)

43. Rahman, J., Newton, M.A.H., Islam, M.K.B., Sattar, A.: Enhancing protein inter-residue real distance prediction by scrutinising deep learning models. Scient. Reports. **12**(1), 787 (2022). https://doi.org/10.1038/s41598-021-04441-y
44. Bolton, E.E., Wang, Y., Thiessen, P.A., Bryant S.H.: Integrated platform of small molecules and biological activities. PubChem (2008) pp. 217–241
45. Mikolov, T., Chen, K., Corrado, G., Dean J.: Efficient estimation of word representations in vector space. In: Proceedings of the international conference on learning representations, workshop track. Arizona, USA; 2013. Available from: https://arxiv.org/abs/1301.3781v3
46. Zheng, J., Xiao, X., Qiu, W.R.: iCDI-W2vCom: identifying the ion channel-drug interaction in cellular networking based on word2vec and node2vec. Front. Genet. **9**, 12 (2021). https://doi.org/10.3389/fgene.2021.738274
47. Yi, H.C., You, Z.H., Cheng, L., Zhou, X., Jiang, T.H., Li, X., et al.: Learning distributed representations of RNA and protein sequences and its application for predicting lncRNA-protein interactions. Comput. Struct. Biotech. J. **18**, 20–26 (2020). https://doi.org/10.1016/j.csbj.2019.11.004
48. Huang, K., Xiao, C., Glass, L.M., Sun, J.: MolTrans: molecular interaction transformer for drug-target interaction prediction. Bioinformatics **37**(6), 830–836 (2021). https://doi.org/10.1093/bioinformatics/btaa880
49. Suzek, Baris E., Wang, Yuqi, Huang, Hongzhan, McGarvey, Peter B., Cathy, H Wu.: UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics **31**(6), 926–932 (2015). https://doi.org/10.1093/bioinformatics/btu739
50. Irwin, J.J., Shoichet, B.K.: ZINC-A free database of commercially available compounds for virtual screening. J. Chem. Inf. Model. **45**(1), 177–182 (2005). https://doi.org/10.1021/CI049714
51. Gaulton, A., Hersey, A., Nowotka, M., Bento, A.P., Chambers, J., Mendez, D., et al.: The ChEMBL database in 2017. Nucl. Acids Res. **45**(D1), D945–D954 (2017). https://doi.org/10.1093/nar/gkw1074
52. Perozzi, B., Al-Rfou, R., Skiena, S.: DeepWalk: Online learning of social representations. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. **3**, 701–710 (2014). https://doi.org/10.1145/2623330.2623732
53. Nguyen, T., Le, H., Quinn, T.P., Nguyen, T., Le, T.D., Venkatesh, S.: GraphDTA: predicting drug-target binding affinity with graph neural networks. Bioinformatics **37**(8), 1140–1147 (2021). https://doi.org/10.1093/bioinformatics/btaa921
54. Altae-Tran, H., Ramsundar, B., Pappu, A.S., Pande, V.: Low data drug discovery with one-shot learning. ACS Centr. Sci. **3**(4), 283–293 (2017). https://doi.org/10.1021/ACSCENTSCI.6B00367
55. Gilmer, J., Schoenholz, SS., Riley, PF., Vinyals, O., Dahl, GE.: Neural message passing for quantum chemistry. In: Proceedings of the international conference on machine learning. Vienna, Austria; 2017. p. 2053–2070. Available from: https://arxiv.org/abs/1704.01212v2
56. Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., et al.: Analyzing learned molecular representations for property prediction. J. Chem. Inf. Model. **59**(8), 3370–3388 (2019). https://doi.org/10.1021/ACS.JCIM.9B00237/SUPPL_FILE/CI9B00237_SI_001.PDF
57. Unke, O.T., Chmiela, S., Gastegger, M., Schütt, K.T., Sauceda, H.E., Müller, K.R.: SpookyNet: learning force fields with electronic degrees of freedom and nonlocal effects. Nat. Commun. **12**(1), 7273 (2021). https://doi.org/10.1038/s41467-021-27504-0
58. Cichonska, A., Ravikumar, B., Parri, E., Timonen, S., Pahikkala, T., Airola, A., et al.: Computational-experimental approach to drug-target interaction mapping: A case study on kinase inhibitors. PLOS Comput. Biol. **13**(8), e1005678 (2017)
59. He, T., Heidemeyer, M., Ban, F., Cherkasov, A., Ester, M.: SimBoost: a read-across approach for predicting drug-target binding affinities using gradient boosting machines. J. Cheminf. **9**(1), 1–14 (2017)
60. Stepniewska-Dziubinska, M.M., Zielenkiewicz, P., Siedlecki, P.: Development and evaluation of a deep learning model for protein-ligand binding affinity prediction. Bioinformatics **34**(21), 3666–3674 (2018)
61. Öztürk, H., Özgür, A., Ozkirimli, E.: DeepDTA: deep drug-target binding affinity prediction. Bioinformatics **34**(17), 821–829 (2018)
62. Jiang, M., Li, Z., Zhang, S., Wang, S., Wang, X., Yuan, Q., et al.: Drug-target affinity prediction using graph neural network and contact maps. RSC Adv. **10**(35), 20701–20712 (2020). https://doi.org/10.1039/D0RA02297G
63. Zheng, S., Li, Y., Chen, S., Xu, J., Yang, Y.: Predicting drug-protein interaction using quasi-visual question answering system. Nat. Mach. Intell. **2**(2), 134–140 (2020)
64. Nguyen, T.M., Nguyen, T., Le, T.M., Tran, T.: GEFA: early fusion approach in drug-target affinity prediction. IEEE/ACM Trans. Comput. Biol. Bioinf. **19**(2), 718–728 (2022). https://doi.org/10.1109/TCBB.2021.3094217
65. Guedes, I.A., Barreto, A.M.S., Marinho, D., Krempser, E., Kuenemann, M.A., Sperandio, O., et al.: New machine learning and physics-based scoring functions for drug discovery. Scientif. Report. **11**(1), 3198 (2021). https://doi.org/10.1038/s41598-021-82410-1
66. Cheng, Z., Zhao, Q., Li, Y., Wang, J.: IIFDTI: predicting drug-target interactions through interactive and independent features based on attention mechanism. Bioinformatics **38**(17), 4153–4161 (2022). https://doi.org/10.1093/bioinformatics/btac485
67. You, J., Liu, B., Ying, R., Pande, V., Leskovec, J.: Graph convolutional policy network for goal-directed molecular graph generation. In: Proceedings of advances in neural information processing systems. Montreal, Canada; 2018. p. 6410–6421. Available from: https://arxiv.org/abs/1806.02473v3
68. Jin, W., Barzilay, R., Jaakkola, T.: Junction tree variational autoencoder for molecular graph generation. In: Proceedings of the international conference on machine learning. Vienna, Austria; 2018. p. 3632–3648. Available from: https://arxiv.org/abs/1802.04364v4
69. Zhou, Z., Kearnes, S., Li, L., Zare, R.N., Riley, P.: Optimization of molecules via deep reinforcement learning. Scientif. Reports. **9**(1), 1–10 (2019)
70. Winter, R., Montanari, F., Steffen, A., Briem, H., Noé, F., Clevert, D.A.: Efficient multi-objective molecular optimization in a continuous latent space. Chem. Sci. **10**(34), 8016–8024 (2019). https://doi.org/10.1039/c9sc01928f
71. Yang, S., Hwang, D., Lee, S., Ryu, S., Ju Hwang, S.: Hit and Lead Discovery with Explorative RL and Fragment-based Molecule Generation. In: Proceedings of advances in neural information processing systems; 2021
72. Kusner, MJ., Paige, B., Hernández-Lobato, JM.: Grammar variational autoencoder. In: Proceedings of the international conference on machine learning. Sydney, Australia; 2017. p. 3072–3084. Available from: https://arxiv.org/abs/1703.01925v1
73. Simonovsky, M., Komodakis, N.: GraphVAE: towards generation of small graphs using variational autoencoders. In: Proceedings of the international conference on artificial neural networks. Siem Reap, Cambodia; 2018. p. 412–422. Available from: http://link.springer.com/10.1007/978-3-030-01418-6_41
74. You, J., Ying, R., Ren, X., Hamilton, WL., Leskovec, J.: GraphRNN: Generating realistic graphs with deep auto-regressive models. In: Proceedings of the international conference on

machine learning. Stockholm Sweden; 2018. p. 9072–9081. Available from: https://arxiv.org/abs/1802.08773v3
75. Maziarka, L., Pocha, A., Kaczmarczyk, J., Rataj, K., Danel, T., Warchoł, M.: Mol-CycleGAN: a generative model for molecular optimization. J. Cheminf. **12**(1), 2 (2020). https://doi.org/10.1186/s13321-019-0404-1
76. Shi, C., Xu, M., Zhu, Z., Zhang, W., Zhang, M., Tang, J.: GraphAF: A flow-based autoregressive model for molecular graph generation. In: Proceedings of the international conference on learning representations. Addis Ababa, Ethiopia; 2020. Available from: https://arxiv.org/abs/2001.09382v2
77. Baylon, J.L., Cilfone, N.A., Gulcher, J.R., Chittenden, T.W.: Enhancing retrosynthetic reaction prediction with deep learning using multiscale reaction classification. J. Chem. Inf. Model. **59**(2), 673–688 (2019). https://doi.org/10.1021/acs.jcim.8b00801
78. Dai, H., Li, C., Coley, CW., Dai, B., Song, L.: Retrosynthesis prediction with conditional graph logic network. In: Proceedings of advances in neural information processing systems. Vancouver, Canada; 2019. Available from: https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html
79. Karpov, P., Godin, G., Tetko, IV.: A transformer model for retrosynthesis. In: Proceedings of the international conference on artificial neural networks. Munich, Germany; 2019. p. 817–830. Available from: http://link.springer.com/10.1007/978-3-030-30493-5_78
80. Shi, C., Xu, M., Guo, H., Zhang, M., Tang, J.: A graph to graphs framework for retrosynthesis prediction. In: Proceedings of the international conference on machine learning. vol. PartF168147-12. Vienna, Austria; 2020. p. 8777–8786. Available from: https://arxiv.org/abs/2003.12725v3
81. Pham, T., Tran, T., Venkatesh, S.: Graph memory networks for molecular activity prediction. In: Proceedings of the international conference on pattern recognition. Beijing, China; 2018. p. 639–644. Available from: https://arxiv.org/abs/1801.02622v2
82. Ma, W., Yang, L., He, L.: Overview of the detection methods for equilibrium dissociation constant KD of drug-receptor interaction. J. Pharmaceut. Anal. **8**(3), 147–152 (2018)
83. Thafar, M., Raies, A.B., Albaradei, S., Essack, M., Bajic, V.B.: Comparison study of computational prediction tools for drug-target binding affinities. Front. Chem. **11**, 7 (2019). https://doi.org/10.3389/fchem.2019.00782
84. Meng, E.C., Shoichet, B.K., Kuntz, I.D.: Automated docking with grid-based energy evaluation. J. Comput. Chem. **13**(4), 505–524 (1992)
85. Kundu, I., Paul, G., Banerjee, R.: A machine learning approach towards the prediction of protein-ligand binding affinity based on fundamental molecular properties. RSC Adv. **8**(22), 12127–12137 (2018)
86. Gomes, J., Ramsundar, B., Feinberg, EN., Pande, VS: atomic convolutional networks for predicting protein-ligand binding affinity. arXiv preprint arXiv:1703.10603. 2017;
87. Do, K., Tran, T., Nguyen, T., Venkatesh, S.: Attentional multilabel learning over graphs: a message passing approach. Mach. Learn. **108**(10), 1757–1781 (2018)
88. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph attention networks. In: Proceedings of the international conference on learning representations. Toulon, France; 2017. Available from: https://arxiv.org/abs/1710.10903v3
89. Bickerton, G.R., Paolini, G.V., Besnard, J., Muresan, S., Hopkins, A.L.: Quantifying the chemical beauty of drugs. Nat. Chem. **4**(2), 90–98 (2012)
90. Do, K., Tran, T., Venkatesh, S.: Graph transformation policy network for chemical reaction prediction. Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining. **7**, 750–760 (2019). https://doi.org/10.1145/3292500.3330958
91. Amberger, J.S., Bocchini, C.A., Scott, A.F., Hamosh, A.: OMIM.org: leveraging knowledge across phenotype-gene relationships. Nucl. Acid Res. **47**(D1), D1038–D1043 (2019). https://doi.org/10.1093/nar/gky1151
92. Forbes, S.A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., et al.: COSMIC: somatic cancer genetics at high-resolution. Nucl. Acid Res. **45**(D1), D777–D783 (2017). https://doi.org/10.1093/nar/gkw1121
93. Paliwal, S., de Giorgio, A., Neil, D., Michel, J.B., Lacoste, A.M.: Preclinical validation of therapeutic targets predicted by tensor factorization on heterogeneous graphs. Scientif. Reports. **10**(1), 18250 (2020). https://doi.org/10.1038/s41598-020-74922-z
94. Zitnik, M., Agrawal, M., Leskovec, J.: Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics **34**(13), i457–i466 (2018). https://doi.org/10.1093/BIOINFORMATICS/BTY294
95. Weber, L., Sänger, M., Münchmeyer, J., Habibi, M., Leser, U., Akbik, A.: HunFlair: an easy-to-use tool for state-of-the-art biomedical named entity recognition. Bioinformatics **37**(17), 2792–2794 (2021). https://doi.org/10.1093/BIOINFORMATICS/BTAB042
96. Müller, H.M., Van Auken, K.M., Li, Y., Sternberg, P.W.: Textpresso central: a customizable platform for searching, text mining, viewing, and curating biomedical literature. BMC Bioinform. **19**(1), 1–16 (2018). https://doi.org/10.1186/S12859-018-2103-8/FIGURES/11
97. Szklarczyk, D., Gable, A.L., Nastou, K.C., Lyon, D., Kirsch, R., Pyysalo, S., et al.: The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucl. Acid. Res. **49**(D1), D605–D612 (2021). https://doi.org/10.1093/NAR/GKAA1074
98. Li, J., Sun, Y., Johnson, RJ., Sciaky, D., Wei, CH., Leaman, R., et al.: BioCreative V CDR task corpus: a resource for chemical disease relation extraction. Database. 2016 5;2016:baw068. https://doi.org/10.1093/database/baw068
99. Chen, X., Chen, H., Zhang, N., Chen, J., Wu, Z.: OWL reasoning over big biomedical data. In: Proceedings of the international conference on big data. Santa Clara, CA, USA; 2013. p. 29–36. Available from: http://ieeexplore.ieee.org/document/6691755/
100. Das, R., Dhuliawala, S., Zaheer, M., Vilnis, L., Durugkar, I., Krishnamurthy, A., et al.: Go for a walk and arrive at the answer: reasoning over paths in knowledge bases using reinforcement learning. In: Proceedings of the international conference on learning representations, ICLR 2018 - Conference Track Proceedings; 2017. Available from: https://arxiv.org/abs/1711.05851v2
101. Mohamed, S.K., Nováček, V., Nounu, A.: Discovering protein drug targets using knowledge graph embeddings. Bioinformatics **36**(2), 603–610 (2019). https://doi.org/10.1093/bioinformatics/btz600
102. Fei, H., Ren, Y., Zhang, Y., Ji, D., Liang, X.: Enriching contextualized language model from knowledge graph for biomedical information extraction. Brief. Bioinform. (2021). https://doi.org/10.1093/bib/bbaa110
103. Ye, Q., Hsieh, C.Y., Yang, Z., Kang, Y., Chen, J., Cao, D., et al.: A unified drug-target interaction prediction framework based on knowledge graph and recommendation system. Nat. Commun. **12**(1), 6775 (2021). https://doi.org/10.1038/s41467-021-27137-3
104. Liu, Y., Hildebrandt, M., Joblin, M., Ringsquandl, M., Raissouni, R., Tresp, V.: Neural multi-hop reasoning with logical rules on biomedical knowledge graphs. In: Extended semantic web conference; 2021. Available from: https://github.com/liu-yushan/PoLo
105. Wang, S., Du, Z., Ding, M., Rodriguez-Paton, A., Song, T.: KG-DTI: a knowledge graph based deep learning method for

drug-target interaction predictions and Alzheimer's disease drug repositions. Appl. Intell. **52**(1), 846–857 (2022). https://doi.org/10.1007/s10489-021-02454-8
106. Galárraga, LA., Teflioudi, C., Hose, K., Suchanek, F.: AMIE: association rule mining under incomplete evidence in ontological knowledge bases. In: Proceedings of the international conference on world wide web. Rio de Janeiro, Brazil; 2013. p. 413–422. Available from: http://dl.acm.org/citation.cfm?doid=2488388.2488425
107. UniProt Consortium: UniProt: a worldwide hub of protein knowledge. Nucl. Acid Res. **47**(D1), D506–D515 (2019)
108. Wishart, D.S.: DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucl. Acid Res. **34**(90001), D668–D672 (2006). https://doi.org/10.1093/nar/gkj067
109. El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., et al.: The Pfam protein families database in 2019. Nucl. Acids Res. **47**(D1), D427–D432 (2019). https://doi.org/10.1093/nar/gky995
110. Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., Kanehisa, M.: Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. Bioinformatics. **24**(13), i232–i240 (2008). https://doi.org/10.1093/bioinformatics/btn162
111. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: Proceedings of advances in neural information processing systems. Nevada, USA; (2013)
112. Luo, Y., Zhao, X., Zhou, J., Yang, J., Zhang, Y., Kuang, W., et al.: A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. Nat. Comm. **8**(1), 573 (2017). https://doi.org/10.1038/s41467-017-00680-8
113. Himmelstein, DS., Lizee, A., Hessler, C., Brueggeman, L., Chen, SL., Hadley, D., et al.: Systematic integration of biomedical knowledge prioritizes drugs for repurposing. eLife. 2017 9;6. https://doi.org/10.7554/eLife.26726
114. Walsh, B., Mohamed, SK., Nováček, V.: BioKG: A knowledge graph for relational learning on biological data. In: Proceedings of the 29th ACM International conference on information & knowledge management. New York, NY, USA: ACM; 2020. p. 3173–3180
115. Xiong, W., Hoang, T., Wang, WY.: DeepPath: a reinforcement learning method for knowledge graph reasoning. Proceedings of empirical methods in natural language processing. 2017 7; pp. 564–573. https://doi.org/10.18653/v1/d17-1060
116. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997). https://doi.org/10.1162/NECO.1997.9.8.1735
117. Edwards, G., Nilsson, S., Rozemberczki, B., Papa, E.: Explainable biomedical recommendations via reinforcement learning reasoning on knowledge graphs. In: International workshop on machine learning on graphs; 2021. Available from: https://arxiv.org/abs/2111.10625v1
118. Bommasani, R., Hudson, DA., Adeli, E., Altman, R., Arora, S., von Arx S., et al.: On the opportunities and risks of foundation models. (2021) arXiv:2108.07258
119. Viceconti, M., Pappalardo, F., Rodriguez, B., Horner, M., Bischoff, J., Musuamba, Tshinanu F.: In silico trials: verification, validation and uncertainty quantification of predictive models used in the regulatory evaluation of biomedical products. Methods. **1**(185), 120–127 (2021). https://doi.org/10.1016/J.YMETH.2020.01.011
120. Goole, J., Amighi, K.: 3D printing in pharmaceutics: a new tool for designing customized drug delivery systems. Int. J. Pharmac. **499**(1–2), 376–394 (2016). https://doi.org/10.1016/j.ijpharm.2015.12.071