Highlights

A Survey on Deep Learning Architectures for *De Novo* Drug Design

S Falciglia

- The chemical space may contain between 10^{23} to 10^{60} drug-like compounds, which means that exploring this extensive space is computationally impractical. The goal of *de novo* drug design is to create workable molecules, rather than searching for them.
- Although structure-based drug design (SBDD) methods are the most accurate, they are still seen as time-consuming and laborious. Recently, geometric deep learning has been proposed as a way of addressing these issues.
- For deep learning-based models for *de novo* drug design, a new classification based on recent changes in trends, purpose, and usability is proposed. This classification complements the natural categorization that arises from considering the architectures alone.

A Survey on Deep Learning Architectures for *De Novo* Drug Design

S Falciglia[a,1]

[a] *Department of Computer, Control and Management Engineering, Via Ariosto 25, Rome, 00185, Italy,*

ARTICLE INFO

Keywords:
drug discovery
de novo drug design
binding ligand generation
structure-based drug design

ABSTRACT

De novo drug design is one step in the drug discovery process, with the goal of creating binding ligand molecules from the target protein's information, rather than searching through the whole chemical space. This approach helps in generating diverse ligand shapes that complement the pocket without the need to search through the entire chemical space. Several deep learning-based models have been developed recently for *de novo* drug design. This study aims to conduct a comprehensive survey of the latest methods for *de novo* drug design. Two different taxonomies are proposed based on the deep learning architectures and the purpose and usability of the model.

1. Introduction

The process of drug discovery involves identifying active compounds that have therapeutic effects on the intended diseases. Despite its ability to scan thousands of different compounds individually, a high throughput screening technique remains both time-consuming and expensive [1]. On average, it costs US$2.8 billion and takes 15 years to discover and develop a drug [2]. Pharmaceutical chemists, in particular, have encountered significant challenges in selecting and creating prospective medicines for a specific area that meet the criteria for preclinical testing. In order to overcome these challenges, AI techniques have been employed in almost every aspect of drug discovery.

In conjunction with accessible data resources, various deep learning based methodologies are emerging throughout all stages of the drug development process. Predicting target structure, DTI, drug-target binding affinity, and *de novo* design represent significant challenges. *De novo* drug design specifically pertains to the generation of novel drug-like compounds without the aid of a starting template. Whilst conventional structure-based and ligand-based drug design methods have facilitated the identification of small-molecule drug candidates, they rely on knowledge pertaining to the active site of a biological target or the pharmacophores[1] of a known active binder [3], thus impeding their ability to be applied in modern drug discovery. There are estimated to be between 10^{23} and 10^{60} drug-like molecules in the chemical space [4]; therefore, completely exploring this vast chemical space is computationally infeasible. Under this light, *De novo* drug design (molecular generation) seeks to explore chemical space by generating fresh molecules with preferable features from scratch to complement existing chemical libraries.

In recent years, various deep learning based models have been proposed for *de novo* drug design, employing SMILES, fingerprints, molecular graphs, and 3D geometry as input. The aim of this survey is to provide a comprehensive survey on deep learning architectures for *De Novo* Drug Design. The rest of the paper is organized as follows. Section 2 defines the input format, referred to as the molecular descriptors. In Section 3, we provide some indications regarding the evaluation method for the architectures. Sections 4 and 5 describe the state-of-the-art deep learning models for *de novo* drug design. The former focuses on the classification based on different architectures used, while the latter discusses the classification based on recent trend changes, purpose, and usability. Finally, Section 6 draws the conclusions of this study.

2. Molecular descriptors

One of the essential aspects of AI-based drug discovery and analysis is to convert molecules into a computer-readable format, without compromising their intrinsic physicochemical properties [5]. Several drug descriptors have been proposed to represent drugs [6].

The zero-dimensional (0D) descriptor is the most straightforward molecular representation while being obtained with the chemical formula of drugs [7]. It includes fundamental descriptors such as molecular weight, atom number, atom type, and number of heavy atoms. The 0D descriptor is overly simplistic and can exclusively extract shallow information.

The one-dimensional (1D) descriptor encodes drugs based on their substructures, such as the number of rings, functional groups, substituent atoms, and atom-centered fragments, as described in [7]. The elements of the 1D descriptor are usually binary, where, for instance, 1/0 indicates if a substituent atom is present or absent, or are defined based on substructure occurrence frequencies. In addition to the property-based 1D descriptor, there is another type of 1D descriptor - the Simplified Molecular Input Line Entry System (SMILES), described in [8]. A drug's representation in SMILES uses a string of characters. SMILES is dependent on the order of atoms, and because of this, a drug may have multiple SMILES representations.

The two-dimensional (2D) descriptor offers supplementary information to the 1D descriptor by considering atom

falciglia.2015426@studenti.uniroma1.it (S. Falciglia)
ORCID(s): 0009-0009-4871-2553 (S. Falciglia)

[1] Pharmacophores. Pharmacophore is defined as the essential geometric arrangement of atoms or functional groups necessary to produce a given biological response.

adjacency, connectivity, and other kinds of topological features. Thus, 2D descriptors are typically obtained by representing a drug as a graph, where nodes represent atoms and edges represent bonds. To obtain more information, the molecular fingerprint (FP) was proposed for encoding molecules in binary form [9]. FP indicates the presence or absence of specific substructures through a string of fixed length marked with 1 or 0.

The three-dimensional (3D) molecule descriptor represents a molecule in 3D space [10], where each atom is identified by its spatial coordinates on the x, y, and z axes. It comprises both spatial and geometrical configuration information and thus has a high information content. However, traditional structure-based drug design methods, which depend on physical modeling, hand-crafted scoring functions, and enumeration, are still considered time-consuming and laborious. To tackle these issues, geometric deep learning [11] has been recently suggested as a way to accelerate and enhance the structure-based drug design process.

The high information content of a molecule's 3D descriptor makes structure-based drug design (SBDD) [12, 13, 14] an essential tool in the design and optimization of drug candidates through the effective utilization of the three-dimensional geometric information of target proteins. SBDD can be organized into three stages. The first step involves the Binding Site Prediction task, aiming to identify regions of the protein structure that can serve as binding sites for ligand molecules. The second step can consist of three independent tasks. Binding Pose Generation, also known as protein-ligand docking, focuses on predicting the binding conformations of the protein-ligand complex. *De novo* Molecule Generation involves generating binding ligand molecules from scratch using structural information about the target protein. Linker Design involves combining disconnected molecular fragments into a combined ligand molecule conditioned on the target protein. Once the protein-ligand complex structure is obtained, the third step involves the Binding Affinity Prediction task, which aims to predict the affinity between a protein and a ligand-based on their binding structure. It's worth noting that the order and categories of SBDD tasks are not always fixed because SBDD is an iterative process that proceeds through multiple cycles, ultimately leading to optimized drug candidates for clinical trials [15]. Certain techniques may have the ability to accomplish multiple tasks, as EquiBind [16], which allows for the prediction of ligand binding poses without prior knowledge of the binding site, i.e., using blind docking.

3. Evaluation metrics

The most accurate way to assess the quality of the generated molecules is undoubtedly through direct wet laboratory experiments. This is often not feasible due to the large number of molecules that are generated. As a result, certain criteria employed in the field of machine learning (ML) are used to evaluate the quality of generated molecules [17]. Currently utilized metrics have certain limitations. Indeed, a generative model may overemphasize these metrics by producing compounds that lack drug-likeness, such as short/long-chain alkanes and macrocyclic compounds.

Four generation indices, namely validity, uniqueness, novelty, and diversity, are commonly used in studies concerning *de novo* drug design. The generation index is essential in the evaluation of the performance of the DL generator model via sets of generated molecules, rather than evaluating the generated substances as drugs. However, this does not suggest that models with superior generative metrics produce better drugs. For a rapid calculation of the generation index of SMILES data, it is possible to use an open library like RDKit [18], GuacaMol [19], or MOSES [20]. Validity assesses whether a generated compound can exist or not. Uniqueness is a criterion that determines if the generator creates a new compound without duplicating it. Novelty is a measure of non-overlap by comparing the generated set with the existing set of data. Diversity or dissimilarity (or distance) is an indicator that determines the level of dissimilarity and diversity in the resulting compounds when a small number of structures or a small number of atoms are changed. Moreover, controllability is primarily employed in *de novo* models that incorporate condition control functions [21, 22]. It denotes the level of precision with which the output compound's property value is distributed concerning the input condition. It is not portrayed as a specific value in contrast to other metrics, but it is typically depicted using a histogram to evaluate the distribution when compared to the target value. The performance gets better with the decrease in variance.

Finally, a high-quality dataset is essential for applying AI to drug discovery. For instance, the Drug-Gene Interaction Database (DGIdb) [23] supply details on drug-gene interactions as well as genes or gene products that may interact with drugs [24]. To date, DGIdb comprises more than 40,000 genes and 10,000 drugs involved in over 100,000 drug-gene interactions. Other well-known datasets include ChEMBL [25], ChemDB [26], COCONUT (The Collection of Open Natural Products) [27], DrugBank [28], DTC (Drug Target Commons) [29], INPUT (Intelligent Network Pharmacology Platform Unique for Traditional, Chinese Medicine) [30], PubChem [31], SIDER (Side Effect Resource) [32], and STITCH (Search Tool for Interacting Chemicals) [33]. These databases facilitate the transition of drug discovery into the big data era and accelerate the drug discovery process [6].

4. Categorization based on DL models

This section aims to categorize studies on *de novo* drug design based on the deep learning models used. Generally, these models require SMILES, i.e., 1D data, nodes and edges of a graph, i.e., 2D data, or 3D data, i.e., structural data. Structural data availability has increased due to recent advancements in highly accurate protein structure prediction, such as AlphaFold [34]. Models that require structural data as input are part of Geometric Deep Learning (GDL) [11]. This family of neural network architectures incorporates and encodes 3D geometric data, automatically extracting useful 3D structural features, instead of relying on handcrafted feature engineering. As explained in Section 2, which states the three-stage structure of SBDD, these architectures take the pocket 3D-graph as input and then output the ligand 3D-graph [35].

We categorize *de novo* drug design models as Encoder-Decoder based, RNN based, GAN/Flow/Diffusion based, GNN based, and reinforcement learning based.

Encoder-Decoder based CVAE (Conditional VAE) [36] is a molecular generative model specialized to control multiple molecular properties simultaneously by imposing them on a latent space. The condition vector represents the molecular properties that must be controlled and is involved in both encoding and decoding processes. Both the input and output of the model are represented using SMILES codes. The encoder and decoder both consist of a Recurrent Neural Network (RNN) with a Long-Short Term Memory (LSTM) cell. The decoder cells are unrolled 120 times, and a softmax layer is applied to the output of the decoder. ECAAE (entangled conditional adversarial autoencoder) [37] also creates molecular structures based on multiple properties. Its architecture enhances that of the Supervised AAE [37], which had dozens of intricate conditions and thousands of variations in generating substantial objects such as molecules. Here, molecules are represented as SMILES. SQUID (Shape-conditioned eQUIvariant generator for Drug-like molecules) [38] is a novel multimodal 3D generative model that allows shape-conditioned 3D molecular design by encoding molecular shape equivariantly and chemical identity variably. By using autoregressive fragment-based generation with heuristic bonding geometries, the model ensures the local geometric and chemical validity of the generated molecules. This allows the model to give priority to scoring the rotatable bond for the best alignment of the growing conformational structure to the target shape.

RNN based Overall, 2D graph-based models perform better than SMILES-based models in various metrics, particularly in generating valid outputs [39]. It is not surprising that graph-based models produce highly valid output structures since SMILES generation has much stricter rules for output compared to the generation of molecular graphs. MolRNN [39] is a new *de novo* molecular design framework considerably better tailored for molecule production and based on a sort of sequential graph generator that does not employ atom-level recurrent units. A recurrent RNN network with three GRU layers is used. It performs better than its SMILES-based counterpart MolMP [39], with the added advantage of producing highly interpretable output compared to SMILES.

GAN/Flow/Diffusion based LiGANN [40] produces a range of diverse 3D ligand shapes that complement the pocket. Afterward, a shape-captioning network decodes the generated shapes into SMILES strings, which can be employed as input for large virtual screening campaigns. The model comprises a Bicycle GAN for three-dimensional input. The generator utilizes a 3D CNN to create ligand representations that complement the input protein pocket. By generating diverse shapes, and diverse compound scaffolds that are capable of binding to a protein pocket, as well as diverse pharmacophoric models, can be captured. GraphVF [41] is a variational flow-based framework that combines 2D topology and 3D geometry to generate controllable binding 3D molecules. The approach followed here is to transform a simple prior distribution (i.e., the encoding of the topology pattern in 2D molecules) into the complex data distribution of 3D binding molecules by applying a series of invertible transformation functions. DiffSBDD [12] is a 3D-conditional diffusion model that is SE(3) equivariant, generating new ligands conditioned to protein pockets, respecting the symmetries of translation, rotation, and permutation. Various in silico experiments show how novel and different drug-like ligands with high docking scores are generated.

GNN based GraphINVENT [42] is a platform designed to facilitate graph-based molecular design using GNNs. In particular, the gated-GNN model outperforms other GNN-based models. New molecules are created by generating a single bond probabilistically at a time. The generative model comprises of two segments: a GNN block and a global readout block. The GNN block receives the 2D molecular graph representation as input and produces the transformed note feature vectors and the graph embedding as output. The global readout block predicts a global graph property using the previous outputs of the GNN block. This property consists of the action probability distribution (APD) of each graph, which is a vector containing probabilities for all possible actions to grow a graph. MG^2N^2 (Molecule Generative Graph Neural Network) [43] is a sequential molecular graph generator based on a set of GNN modules that sequentially add a node or a group of nodes to the graph, along with their connections. This enables the utilization of the graph output from the preceding step as the network input. The generation process becomes more interpretable due to the sequentiality and modularity of the architecture.

Reinforcement learning based GENTRL (generative tensorial reinforcement learning) [44] optimizes synthetic feasibility, novelty, and biological activity. The molecular representation is given in SMILES notation. It is a two-step generative machine learning algorithm that combines reinforcement learning, variational inference, and tensor decompositions. Firstly, an encoder-decoder based model learns a mapping of the chemical space to a continuous space of 50 dimensions. Then, reinforcement learning is used to explore this space and discover new compounds. GENTRL employs three separate self-organizing maps (SOMs) as reward functions: the trending SOM, the general kinase SOM, and the specific kinase SOM. The method prioritizes the generated structures by utilizing these three SOMs sequentially. ReLeaSE (Reinforcement Learning for Structural Evolution) [45] integrates two deep neural networks – generative and predictive – that are separately trained but jointly used to produce novel chemical libraries with a specific target. The process employs SMILES strings as a representation of molecules. The method is characterized by two phases. During the initial phase, generative and predictive models are separately trained using a supervised learning algorithm. During the second phase, the two models are trained together using the RL approach to control the formation of novel chemical structures toward those having desired physical and/or biological properties. The generative model is leveraged to produce fresh molecules, while the predictive module assesses each generated molecule's feasibility and assigns a numerical reward-valued score to it. The generative model is composed of a stack-augmented RNN (Stack-RNN) [46]. The predictive model is made up of an embedding layer, an LSTM layer, and two dense layers. ORGAN (Objective-Reinforces GAN) [47] uses adversarial training and expert-based rewards with reinforcement learning to direct generated samples' structure and quality to specific domain-specific metrics. The ORGAN training process, which is an extension of SeqGAN (Sequence-based GAN) [48], incorporates domain-specific targets other than the discriminator's reward. The input for ORGAN is in SMILES format. In this instance, the generator consists of an RNN with LSTM cells, while the discriminator is a CNN that has been specifically designed for text classification tasks. DeepLigBuilder [49] produces 3D molecular structures and places them within the binding sites of target proteins. The method relies on the Ligand Neural Network (L-Net) [49] which is an innovative graph generative model used for designing 3D molecules that are chemically and conformationally valid, have high drug-likeness and diversity, combined with Monte Carlo tree search (MCTS) to perform structure-based *de novo* drug design tasks. The MCTS algorithm is responsible for searching for molecules with high binding affinity. At the same time, L-Net is used to encourage the structure to be valid, drug-like, widely diverse, and easy to synthesize.

5. Categorization based on purpose

A classification of *de novo* drug designs based on the DL models may not be sufficient to understand the purpose of the model. In this section, we propose a new classification introduced by [50] to reflect recent changes in trends, purpose, and usability (see Fig. 1). Specifically, we distinguish five classes: (i) chemical latent space, (ii) condition control of compounds, (iii) generation at once or sequentially, (iv) fragment-base generation, and (v) genetic algorithm.

5.1. Chemical Latent Space

This type of method is grounded on manifold learning, which generates a low-dimensional space showcasing the latent features of input data within the original data space [51]. In a lower-dimensional space, it is feasible to explore or optimize molecules [52]. Since similar compounds or proteins are densely represented in a well-trained latent space [53], comparing properties or computing compounds using compound structures is also possible [54].

An instance of a model that belongs to this category is GENTRL, refer to Section 4.

5.2. Condition Control of Compounds

When studying new drug candidates, rather than creating numerous compounds and filtering them through multiple stages, it is feasible to impose conditions or properties that meet the intended goal of the generation [36]. Controllable condition models are beneficial in designing new drugs or refining current ones owing to their ability to alter properties while preserving the primary structural features of the molecule.

One instance of a model in this category is the CVAE, as shown in Section 4, where the molecular property is combined with the input value to the encoder and decoder. Advanced iterations of this model have been developed. These include a semi-supervised VAE (SSVAE) [21], which includes a property predictor that enables more compounds to be used, and an adversarial autoencoder (AAE) [22], in which the latent space is modified by directly predicting properties from the latent vector.

5.3. Generation at Once or Sequentially

Within this class, we make a distinction between two methods for generating compounds. The one-time generation method (refer to AE or GAN) generates a new compound directly in the latent space, whereas the sequential method (refer to RNN) starts from nothing or a specific substructure before gradually completing the compound. The one-time method has a simple structure and is capable of providing more diverse results. Conversely, the sequential method is generated while retaining the active site or core scaffold with its essential attributes, thus it is capable of enhancing the binding score or properties.

5.4. Fragment-Based Generation

This type of approach is based on the observation that compounds have more similar properties at the scaffold level than at the atomic level. Fragment-based DL models are beneficial since they are more likely to produce a product that exists naturally when constructing rather substantial molecules, whereas an atom-based model allows for the fabrication of compounds that may contain, for example, a 10-carbon ring, which is rare in nature.

5.5. Genetic Algorithm

The Genetic Algorithm is a method inspired by biogenetics, primarily used to address optimization problems. The algorithm generates an initial generation random data set, which is then combined to create a new generation. This process is repeated to intersect some data with the highest score to create the next generation and obtain the most optimal results. Recently, researchers have studied the use of a Genetic Algorithm for *de novo* drug design, proposing its combination with deep learning.

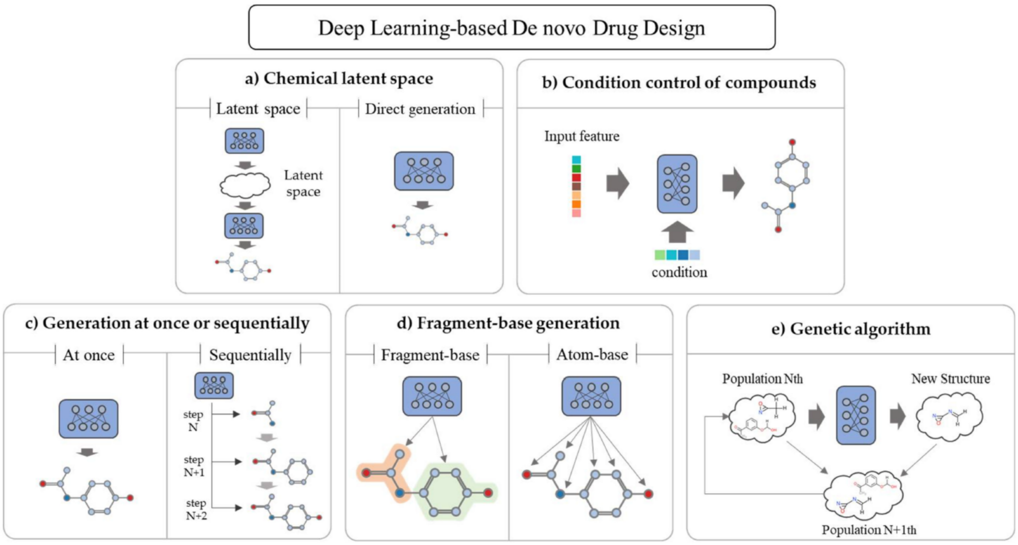

Figure 1: Deep learning based *de novo* drug design, courtesy of [50]

An instance is AutoGrow4 [55], a valuable tool for producing completely new drug-like molecules and refining existing ligands as it employs a genetic algorithm to develop anticipated ligands as required. To commence, the AutoGrow4 program begins with an original (input) group of compounds. This initial group, known as generation 0, comprises a selection of chemically diverse molecular fragments (for an entirely new design) or recognized ligands (for enhancing optimization). By applying the three operations of elitism, mutation, and crossover to the initial population, AutoGrow4 produces the primary generation. Following generations are produced similarly from the compositions of the preceding generation.

6. Conclusions

This survey discusses the application of deep learning-based models for *de novo* drug design. The survey explains the centrality of this single step in the drug discovery process, along with the notation and metrics to work with molecules and neural network architectures. Finally, the survey presents two categorizations of *de novo* drug design studies, based on DL architectures and the purpose and usability of the model.

CRediT authorship contribution statement

S Falciglia: Conceptualization of this study.

References

[1] J. A. DiMasi, H. G. Grabowski, R. W. Hansen, Innovation in the pharmaceutical industry: new estimates of r&d costs, Journal of health economics 47 (2016) 20–33.

[2] N. Fleming, How artificial intelligence is changing drug discovery, Nature 557 (2018) S55–S55.

[3] V. D. Mouchlis, A. Afantitis, A. Serra, M. Fratello, A. G. Papadiamantis, V. Aidinis, I. Lynch, D. Greco, G. Melagraki, Advances in de novo drug design: from conventional to machine learning methods, International journal of molecular sciences 22 (2021) 1676.

[4] P. G. Polishchuk, T. I. Madzhidov, A. Varnek, Estimation of the size of drug-like chemical space based on gdb-17 data, Journal of computer-aided molecular design 27 (2013) 675–679.

[5] L. David, A. Thakkar, R. Mercado, O. Engkvist, Molecular representations in ai-driven drug discovery: a review and practical guide, Journal of Cheminformatics 12 (2020) 1–22.

[6] W. Chen, X. Liu, S. Zhang, S. Chen, Artificial intelligence for drug discovery: Resources, methods, and applications, Molecular Therapy-Nucleic Acids (2023).

[7] F. Grisoni, D. Ballabio, R. Todeschini, V. Consonni, Molecular descriptors for structure–activity applications: a hands-on approach, Computational Toxicology: Methods and Protocols (2018) 3–53.

[8] D. Weininger, Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules, Journal of chemical information and computer sciences 28 (1988) 31–36.

[9] A. Capecchi, D. Probst, J.-L. Reymond, One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome, Journal of cheminformatics 12 (2020) 1–15.

[10] H. Matter, T. Pötter, Comparing 3d pharmacophore triplets and 2d fingerprints for selecting diverse compound subsets, Journal of chemical information and computer sciences 39 (1999) 1211–1225.

[11] K. Atz, F. Grisoni, G. Schneider, Geometric deep learning on molecular representations, Nature Machine Intelligence 3 (2021) 1023–1032.

[12] A. Schneuing, Y. Du, C. Harris, A. Jamasb, I. Igashov, W. Du, T. Blundell, P. Lió, C. Gomes, M. Welling, et al., Structure-based drug design with equivariant diffusion models, arXiv preprint arXiv:2210.13695 (2022).

[13] S. Luo, J. Guan, J. Ma, J. Peng, A 3d generative model for structure-based drug design, Advances in Neural Information Processing Systems 34 (2021) 6229–6239.

[14] C. Isert, K. Atz, G. Schneider, Structure-based drug design with geometric deep learning, Current Opinion in Structural Biology 79 (2023) 102548.

[15] M. Batool, B. Ahmad, S. Choi, A structure-based drug discovery paradigm, International journal of molecular sciences 20 (2019) 2783.

[16] H. Stärk, O. Ganea, L. Pattanaik, R. Barzilay, T. Jaakkola, Equibind: Geometric deep learning for drug binding structure prediction, in: International conference on machine learning, PMLR, 2022, pp. 20503–20521.

[17] M. Wang, Z. Wang, H. Sun, J. Wang, C. Shen, G. Weng, X. Chai, H. Li, D. Cao, T. Hou, Deep learning approaches for de novo drug design: An overview, Current Opinion in Structural Biology 72 (2022) 135–144.

[18] K. Y. Gao, A. Fokoue, H. Luo, A. Iyengar, S. Dey, P. Zhang, et al., Interpretable drug target prediction using deep neural representation., in: IJCAI, volume 2018, 2018, pp. 3371–3377.

[19] N. Brown, M. Fiscato, M. H. Segler, A. C. Vaucher, Guacamol: benchmarking models for de novo molecular design, Journal of chemical information and modeling 59 (2019) 1096–1108.

[20] D. Polykovskiy, A. Zhebrak, B. Sanchez-Lengeling, S. Golovanov, O. Tatanov, S. Belyaev, R. Kurbanov, A. Artamonov, V. Aladinskiy, M. Veselov, et al., Molecular sets (moses): a benchmarking platform for molecular generation models, Frontiers in pharmacology 11 (2020) 565644.

[21] S. Kang, K. Cho, Conditional molecular design with deep generative models, Journal of chemical information and modeling 59 (2018) 43–52.

[22] S. H. Hong, S. Ryu, J. Lim, W. Y. Kim, Molecular generative model based on an adversarially regularized autoencoder, Journal of chemical information and modeling 60 (2019) 29–36.

[23] M. Griffith, O. L. Griffith, A. C. Coffman, J. V. Weible, J. F. McMichael, N. C. Spies, J. Koval, I. Das, M. B. Callaway, J. M. Eldred, et al., Dgidb: mining the druggable genome, Nature methods 10 (2013) 1209–1210.

[24] S. L. Freshour, S. Kiwala, K. C. Cotto, A. C. Coffman, J. F. McMichael, J. J. Song, M. Griffith, O. L. Griffith, A. H. Wagner, Integration of the drug–gene interaction database (dgidb 4.0) with open crowdsource efforts, Nucleic acids research 49 (2021) D1144–D1151.

[25] A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrián-Uhalte, et al., The chembl database in 2017, Nucleic acids research 45 (2017) D945–D954.

[26] J. Chen, S. J. Swamidass, Y. Dou, J. Bruand, P. Baldi, Chemdb: a public database of small molecules and related chemoinformatics resources, Bioinformatics 21 (2005) 4133–4139.

[27] M. Sorokina, P. Merseburger, K. Rajan, M. A. Yirik, C. Steinbeck, Coconut online: collection of open natural products database, Journal of Cheminformatics 13 (2021) 1–13.

[28] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, et al., Drugbank 5.0: a major update to the drugbank database for 2018, Nucleic acids research 46 (2018) D1074–D1082.

[29] J. Tang, B. Ravikumar, Z. Alam, A. Rebane, M. Vähä-Koskela, G. Peddinti, A. J. van Adrichem, J. Wakkinen, A. Jaiswal, E. Karjalainen, et al., Drug target commons: a community effort to build a consensus knowledge base for drug-target interactions, Cell chemical biology 25 (2018) 224–229.

[30] X. Li, Q. Tang, F. Meng, P. Du, W. Chen, Input: An intelligent network pharmacology platform unique for traditional chinese medicine, Computational and Structural Biotechnology Journal 20 (2022) 1345–1351.

[31] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, et al., Pubchem substance and compound databases, Nucleic acids research 44 (2016) D1202–D1213.

[32] M. Kuhn, I. Letunic, L. J. Jensen, P. Bork, The sider database of drugs and side effects, Nucleic acids research 44 (2016) D1075–D1079.

[33] M. Kuhn, C. von Mering, M. Campillos, L. J. Jensen, P. Bork, Stitch: interaction networks of chemicals and proteins, Nucleic acids research 36 (2007) D684–D688.

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

[35] Z. Zhang, J. Yan, Q. Liu, E. Che, A systematic survey in geometric deep learning for structure-based drug design, arXiv preprint arXiv:2306.11768 (2023).

[36] J. Lim, S. Ryu, J. W. Kim, W. Y. Kim, Molecular generative model based on conditional variational autoencoder for de novo molecular design, Journal of cheminformatics 10 (2018) 1–9.

[37] D. Polykovskiy, A. Zhebrak, D. Vetrov, Y. Ivanenkov, V. Aladinskiy, P. Mamoshina, M. Bozdaganyan, A. Aliper, A. Zhavoronkov, A. Kadurin, Entangled conditional adversarial autoencoder for de novo drug discovery, Molecular pharmaceutics 15 (2018) 4398–4405.

[38] K. Adams, C. W. Coley, Equivariant shape-conditioned generation of 3d molecules for ligand-based drug design, arXiv preprint arXiv:2210.04893 (2022).

[39] Y. Li, L. Zhang, Z. Liu, Multi-objective de novo drug design with conditional graph generative model, Journal of cheminformatics 10 (2018) 1–24.

[40] M. Skalic, D. Sabbadin, B. Sattarov, S. Sciabola, G. De Fabritiis, From target to drug: generative modeling for the multimodal structure-based ligand design, Molecular pharmaceutics 16 (2019) 4282–4291.

[41] F. Sun, Z. Zhan, H. Guo, M. Zhang, J. Tang, Graphvf: Controllable protein-specific 3d molecule generation with variational flow, arXiv preprint arXiv:2304.12825 (2023).

[42] R. Mercado, T. Rastemo, E. Lindelöf, G. Klambauer, O. Engkvist, H. Chen, E. J. Bjerrum, Graph networks for molecular design, Machine Learning: Science and Technology 2 (2021) 025023.

[43] P. Bongini, M. Bianchini, F. Scarselli, Molecular generative graph neural networks for drug discovery, Neurocomputing 450 (2021) 242–252.

[44] A. Zhavoronkov, Y. A. Ivanenkov, A. Aliper, M. S. Veselov, V. A. Aladinskiy, A. V. Aladinskaya, V. A. Terentiev, D. A. Polykovskiy, M. D. Kuznetsov, A. Asadulaev, et al., Deep learning enables rapid identification of potent ddr1 kinase inhibitors, Nature biotechnology 37 (2019) 1038–1040.

[45] M. Popova, O. Isayev, A. Tropsha, Deep reinforcement learning for de novo drug design, Science advances 4 (2018) eaap7885.

[46] A. Joulin, T. Mikolov, Inferring algorithmic patterns with stack-augmented recurrent nets, Advances in neural information processing systems 28 (2015).

[47] G. L. Guimaraes, B. Sanchez-Lengeling, C. Outeiral, P. L. C. Farias, A. Aspuru-Guzik, Objective-reinforced generative adversarial networks (organ) for sequence generation models, arXiv preprint arXiv:1705.10843 (2017).

[48] L. Yu, W. Zhang, J. Wang, Y. Yu, Seqgan: Sequence generative adversarial nets with policy gradient, in: Proceedings of the AAAI conference on artificial intelligence, volume 31, 2017.

[49] Y. Li, J. Pei, L. Lai, Structure-based de novo drug design using 3d deep generative models, Chemical science 12 (2021) 13664–13675.

[50] J. Kim, S. Park, D. Min, W. Kim, Comprehensive survey of recent drug discovery using deep learning, International Journal of Molecular Sciences 22 (2021) 9983.

[51] C. Fefferman, S. Mitter, H. Narayanan, Testing the manifold hypothesis, Journal of the American Mathematical Society 29 (2016) 983–1049.

[52] B. Sanchez-Lengeling, A. Aspuru-Guzik, Inverse molecular design using machine learning: Generative models for matter engineering, Science 361 (2018) 360–365.

[53] D. Koge, N. Ono, M. Huang, M. Altaf-Ul-Amin, S. Kanaya, Embedding of molecular structure using molecular hypergraph variational autoencoder with metric learning, Molecular informatics 40 (2021) 2000203.

[54] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, A. Aspuru-Guzik, Automatic chemical design using a data-driven continuous representation of molecules, ACS central science 4 (2018) 268–276.

[55] J. O. Spiegel, J. D. Durrant, Autogrow4: an open-source genetic algorithm for de novo drug design and lead optimization, Journal of cheminformatics 12 (2020) 1–16.