European Journal of Medicinal Chemistry 295 (2025) 117825

Contents lists available at ScienceDirect

European Journal of Medicinal Chemistry

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

Review article

How generative Artificial Intelligence can transform drug discovery?

Ainin Sofia Jusoh [a,b], Muhammad Akmal Remli [a,b,*], Mohd Saberi Mohamad [c,d,e], Tristan Cazenave [f], Chin Siok Fong [g]

[a] Institute for Artificial Intelligence and Big Data, Universiti Malaysia Kelantan, Kota Bharu, 16100, Kelantan, Malaysia
[b] Faculty of Data Science and Computing, Universiti Malaysia Kelantan, Kota Bharu, 16100, Kelantan, Malaysia
[c] Health Data Science Lab, Department of Genetics and Genomics, College of Medical and Health Sciences, United Arab Emirates University, Al Ain, 15551, United Arab Emirates
[d] Faculty of Engineering and Technology, Multimedia University, 75450, Melaka, Malaysia
[e] Department of Biosystems Engineering, Faculty of Agricultural Technology, Universitas Brawijaya, 65145, Malang, East Java, Indonesia
[f] LAMSADE, Université Paris Dauphine - PSL, Paris, France
[g] UKM Medical Molecular Biology Institute (UMBI), 56000, Kuala Lumpur, Malaysia

ARTICLE INFO

Keywords:
Generative artificial Intelligence
Drug discovery
Protein-protein interactions
Drug-target interactions
Database
Performance metrics
Molecules representations

ABSTRACT

Generative Artificial Intelligence (Generative AI) is transforming drug discovery by enabling advanced analysis of complex biological and chemical data. This review explores key Generative AI models, including Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), flow-based models and Transformer-based models, with Transformers gaining prominence due to the abundance of text-based biological data and the success of language models like ChatGPT. The paper discusses molecular representations, performance evaluation metrics, and current trends in Generative AI-driven drug discovery, such as protein-protein interactions (PPIs), drug-target interactions (DTIs) and de-novo drug design. However, these approaches face significant challenges, including applicability domain issues, lack of interpretability, data scarcity, novelty, scalability, computational resource limitations, and the absence of standardized evaluation metrics. These challenges hinder model performance, complicate decision-making, and limit the generation of novel and viable drug candidates. To address these issues, strategies such as hybrid models, integration of multiomics datasets, explainable AI (XAI) techniques, data augmentation, transfer learning, and cloud-based solutions are proposed. Additionally, a curated list of databases supporting drug discovery research is provided. The review concludes by emphasizing the need for optimized AI models, robust validation methods, interdisciplinary collaboration, and future academic efforts to fully realize the potential of Generative AI in advancing drug discovery.

1. Introduction

Drug discovery consists of 4 phases, each with a different process as illustrated in Fig. 1. The 4 phases are the research and development (R&D) phase, preclinical studies phase, clinical trial phase, review and approval phase [1]. These phases will take over 10 years and require significant financial investment due to their complexity and high failure rates.

The first phase, the R&D phase, involves five (5) processes. It starts with target identification. Target identification is when scientists or researcher study and understand the disease, disrupted pathways, proteins, enzymes, receptors, or genes associated with the disease. The primary goal of this process is to discover a potential target linked to the disease that a drug can modulate to achieve a therapeutic effect. Once the target has been identified, scientists and researchers will proceed to the target validation process. Target validation involves proving or confirming that modulating the identified target will produce the desired therapeutic outcome. This step provides strong evidence that the target is druggable and critical to the disease, justifying further research.

Following target validation, the process moves to hit generation, where early drug molecules (hits) are identified. These hits exhibit measurable activity against the target or disease. After generating the hits, researchers proceed to lead identification, selecting the most promising molecules for further study. Next, the lead optimization

* Corresponding author. Institute for Artificial Intelligence and Big Data, Universiti Malaysia Kelantan, Kota Bharu, 16100, Kelantan, Malaysia.
E-mail addresses: tgaininsofia98@gmail.com (A.S. Jusoh), akmal@umk.edu.my (M.A. Remli), saberi@uaeu.ac.ae (M.S. Mohamad), cazenave@lamsade.dauphine.fr (T. Cazenave), chinsiokfong@ppukm.ukm.edu.my (C.S. Fong).

https://doi.org/10.1016/j.ejmech.2025.117825
Received 24 February 2025; Received in revised form 6 May 2025; Accepted 26 May 2025
Available online 27 May 2025

process is conducted. At this stage, the identified lead molecules are chemically modified to improve their efficacy, safety, selectivity, and pharmacokinetic properties. Once optimized, the drug candidates undergo preclinical testing using animal models. This stage aims to evaluate the drug's efficacy, toxicity, and ADME (absorption, distribution, metabolism, and excretion) properties.

If the drug candidate demonstrates stability and safety, regulatory permission must be obtained to proceed with clinical trials. Scientists submit an Investigational New Drug (IND) application to regulatory authorities such as the Food and Drug Administration (FDA) or European Medicines Agency (EMA). Clinical trials are typically divided into three phases, mainly to test the dosage and safety monitory. Phase 1 involves 20–80 participants, Phase 2 involves 100–300 participants and Phase 3 involves 1000–3000 participants. Upon successful completion of clinical trials, a New Drug Application (NDA) is submitted for approval. If approved, the drug moves into the post-approval phase. In this final stage, scientists and researchers monitor the long-term safety and potential side effects of the approved drug through post-marketing surveillance.

All of these processes will take up to more than 10 years and cost anywhere from US$161 million to US$4.54 billion. Despite these significant investments of time and money, the majority of potential drug candidates fail during clinical trials for various reasons, such as poor pharmacokinetic properties, insufficient clinical effectiveness, and adverse effects. However, with the emergence of Generative Artificial Intelligence (Generative AI) technology, this lengthy and costly R&D process can be accelerated and made more cost-effective. Besides, this advancement also can reduce the need for the preclinical studies phase and introduce the non-invasive step of producing new drugs. Let's go through how the advancement of Generative AI shaping a new future for the drug discovery industry.

1.1. How generative AI accelerates drug discovery?

INS018_055 is the first Generative AI-based drug generated that successfully passed Phase 1 clinical trials [2] and is currently undergoing Phase 2 clinical trials, with results expected this year [3]. INS018_055 is designed for the treatment of Idiopathic Pulmonary Fibrosis (IPF), a chronic lung disease, and was discovered by Insilico Medicine in 2020 [4]. The company utilized its published Generative pipeline, PandaOmics, to identify the target for this disease. PandaOmics is a Generative AI drug discovery platform that employs the Generative Pretrained Transformer (GPT) model, similar to the model underlying ChatGPT. Remarkably, the process from target identification (under the R&D phase) to the preclinical phase took only 18 months, demonstrating the efficiency and potential of their Generative AI drug discovery approach. Fig. 2 shows the overview of the current process to generate INS018_055 for IPF using Generative AI.

In order to find therapeutic targets for a variety of illnesses, including IPF, Insilico Medicine used biological network analysis, text data from scientific publications, and multi-omic datasets (such as gene expression profiles and pulmonary fibrosis datasets). Using factors including disease-agnostic qualities, accessibility by therapeutic antibodies or small compounds, novelty, druggability, crystal structure availability, and protein-receptor kinase interactions, PandaOmics produced a ranked list of the top five targets. TNIK was found to be the most important target among them. TNIK was chosen because of its comparatively high results in random walks on heterogeneous graphs, causal inference, routes, network neighbors, interactome community, and negative matrix factorization investigations.

After TNIK was identified, Insilico Medicine verified the target to see if a medication may affect TNIK. Single-cell gene expression datasets from both healthy and IPF patients were used in this validation. The findings showed that, in comparison to normal tissue, TNIK expression is substantially higher in cytotoxic T cells, myofibroblasts, and club cells in damaged tissue. These results emphasized the necessity of developing targeted TNIK inhibitors.

To design the inhibitors, Insilico Medicine used the TNIK kinase domain's accessible crystal structures as a blueprint. The activity of the kinase and its active binding sites were described in depth by these structures. Chemical structures that were customized to the unique characteristics of TNIK were then produced using the Chemistry42 structure-based drug-design AI methodology. The goal was to develop inhibitors that might precisely suppress kinase activity without causing off-target effects by binding to the TNIK active site and establishing hydrogen bonds with the Cys108-NH group in the hinge region.

In the hinge region, INS018_055's carboxyl oxygen and Cys108-NH create a hydrogen bridge. Then, the drug candidates are evaluated by

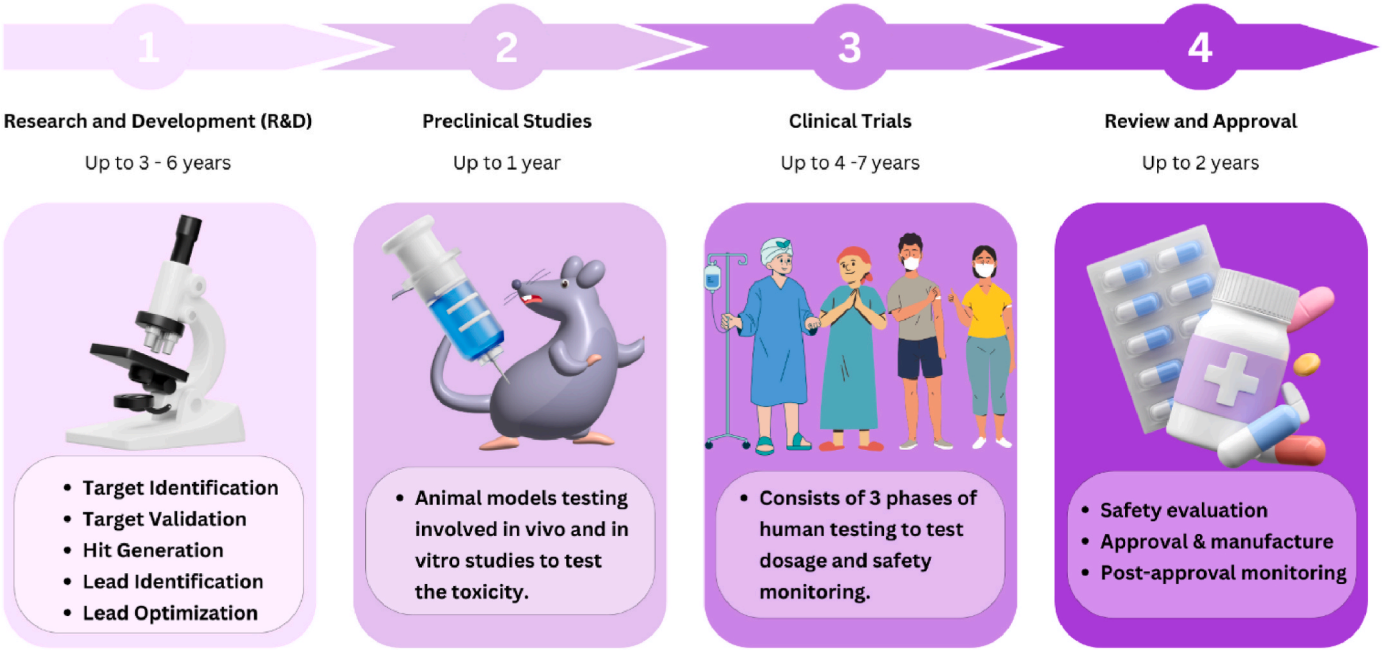

Fig. 1. The process of traditional drug discovery in the wet lab.

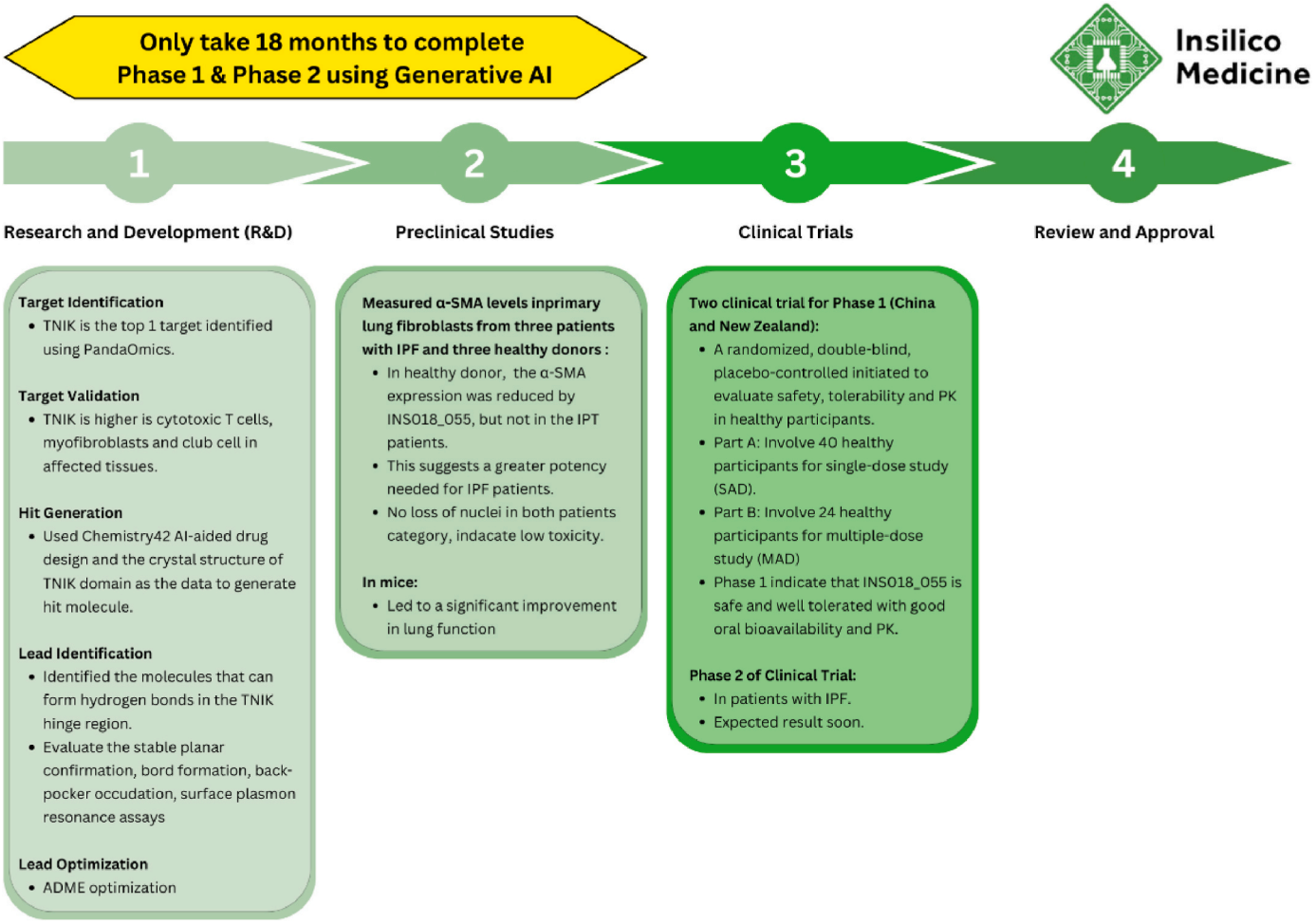

Fig. 2. Overview of the current process to generate INS018_055 for IPF using Generative AI.

their stable planar conformation, bond formation, back-pocket occupation, and surface plasmon resonance assays. In the last lead optimization, this phase prioritizes the improvement in absorption, distribution, metabolism and excretion (ADME) of the drug candidates. The new generated drug candidates were chosen based on the basis of their chemistry, novelty, and synthetic accessibility. This process resulted in a potent candidate, with INS018_055 exhibiting an affinity characterized by a dissociation constant (Kd) value of 4.32 nM. To date, INS018_055 has demonstrated safety in healthy volunteers and is expected to deliver Phase 2 clinical trial results this year. This initiative has shown how advanced these Generative AI technologies are in drug discovery.

2. What is generative AI?

Before going through the advancements of Generative AI in drug discovery, it is essential to understand what Generative AI entails. Generative AI, which underpins applications such as ChatGPT, is an advanced AI model that can analyze and create new data based on the user's prompt [5], as illustrated in Fig. 3. Generative AI can generate new data by understanding the pattern of their training. The new data might be similar data, or improved data based on the user's prompt. Compared to traditional AI, which uses data to analyze and learn from patterns in the data, Generative AI is more sophisticated. Beyond applications such as writing assignments or answering test questions, Generative AI has been utilized to create new drug molecules. In order

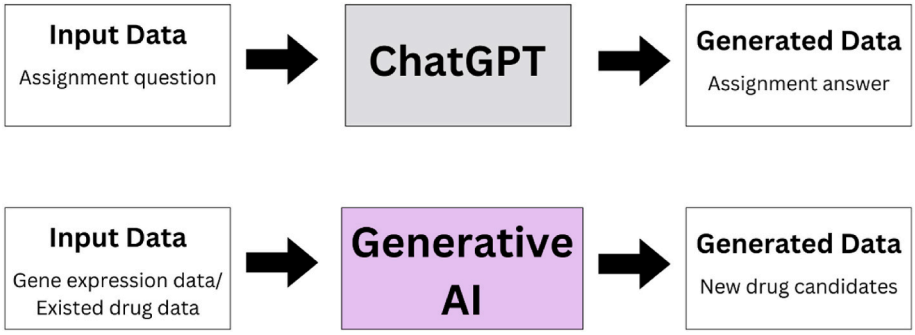

Fig. 3. The similarity between ChatGPT and Generative AI used for drug discovery.

for the Generative AI model to generate new drug candidates based on the input data, researchers must provide enough data for the model to analyze and learn from. Therefore, Generative AI is a model designed to generate or imitate the properties of the original data.

Generative AI is a subset of Deep Learning (DL), Machine Learning (ML), and Artificial Intelligence (AI) [6], as illustrated in Fig. 4. Unlike predictive or classification models, it possesses the ability to generate new data. This capability to create synthetic data opens numerous possibilities across various fields. Generative Adversarial Networks (GANs) were first presented by Ian Goodfellow in 2017, making him the pioneer of Generative AI. GANs are a foundational method that has since become central to the field [7]. In creative industries, Generative AI is used to produce art, music, and voice synthesis, offering new tools for artists and composers. In business, it supports automated content creation, customer service chatbots, and data augmentation. In scientific research, Generative AI is leveraged to generate hypotheses, design experiments and identify novel biomarkers [8,9]. Currently, Generative AI is widely recognized for its transformative impact on drug discovery [10,11]. It has enormous potential to advance pharmaceutical innovation by drastically cutting down on the time and expense needed to develop new drug molecules.

Designing new drug candidates can be accelerated with the advancement of Generative AI technology. However, it remains a complex process that must satisfy predefined criteria related to physical properties, chemical characteristics, and biological measures [12]. These criteria ensure that the generated drug molecules possess favorable pharmacokinetic properties and have a high likelihood of success in clinical trials. Unlike traditional methods, where chemists must manually select and validate safe and effective candidate molecules from a vast chemical space, Generative AI technologies have gained popularity for their ability to automatically generate biologically relevant and synthesizable drug candidates within a significantly shortened timeframe.

3. Tools for generative AI-based drug discovery

To effectively leverage Generative AI technology for drug discovery, several critical factors must be considered. These encompass the types of data required, the desired molecular representations, the choice of appropriate models or algorithms, the selection of suitable performance evaluation metrics and the usability and accessibility of the tools. This section will provide an overview of commonly used datasets in drug discovery research, various molecular representation techniques available to date, widely adopted Generative AI models, the performance evaluation metrics used within this domain and the usability and accessibility of Generative AI-based drug discovery tools.

3.1. Dataset for generative AI drug discovery research

The omics field is experiencing a rapid surge in data generation, driven by advancements in high-throughput sequencing technologies. This explosion of data presents valuable opportunities for predictive modeling in precision medicine, particularly in understanding complex diseases like cancer. To meet the necessary physical, chemical, and biological criteria for drug discovery, Generative AI models depend on adequate and well-validated data. Several widely used chemical and bioinformatics databases provide such datasets, enabling model training, validation, and testing within the drug discovery community.

For virtual (in silico) screening, one well-known resource is the ZINC database (https://zinc.docking.org/), an open-source platform that contains over 750 million commercially available and purchasable chemicals. This includes over 230 million molecules provided in ready-to-dock 3D formats. The vast dataset offered by ZINC is particularly useful for pre-training Generative AI models, which allow them to uncover hidden patterns in both new and old pharmacological compounds. However, ZINC primarily focuses on synthesizability and commercial availability, which may limit the biological activity of the generated molecules. As a result, models trained on ZINC tend to produce molecules that are easy to synthesize but may lack the therapeutic potential required for drug development.

To design drug molecules with enhanced bioactivity, models must also consider chemical interactions with biological targets. A key resource for this purpose is ChEMBL (https://www.ebi.ac.uk/chembl/), which includes over 2.5 million bioactive compounds with drug-like qualities. This database provided detailed information on biological targets and their bioactivity, making it ideal for training models to produce compounds with particular characteristics. However, ChEMBL does not prioritize synthesizability, which can result in molecules that are biologically promising but difficult or expensive to produce. This trade-off highlights the importance of carefully selecting datasets based on the specific goals of the drug discovery project.

The largest open-source chemical information repository, PubChem (https://pubchem.ncbi.nlm.nih.gov/), is another important resource. With over 119 million chemical compounds PubChem provides

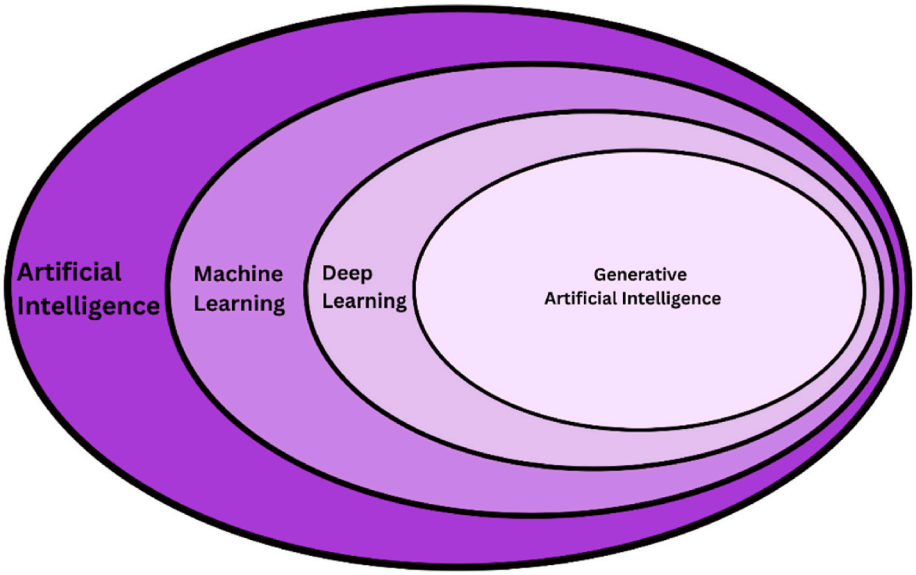

Fig. 4. Generative AI is the subset of Deep Learning, Machine Learning and Artificial Intelligence.

comprehensive details on molecular formulas, structures, properties, biological activities, safety, and toxicity. These datasets allow models to predict molecular behavior under experimental settings. For information on approved and experimental drugs, DrugBank (https://go.drugbank.com/) is a valuable resource. It offers detailed information on drug targets, interactions, and pathways within the human body, making it highly useful for applications such as drug repurposing and interaction studies. Lastly, the LINCS database (https://clue.io/releases/datadashboard) is widely used to train generative models on the properties of biological targets. It contains data on gene expression levels in human cells affected by specific diseases, as well as changes in gene expression in unhealthy cells treated with drugs. This enables models to understand the biological and chemical properties of target cells.

While these databases are widely recognized, there are numerous other datasets available that are suitable for drug discovery, as summarized in Table 1. Collectively, these resources provide immense opportunities for leveraging Generative AI in drug discovery, including de novo drug design, drug repurposing, and other applications. However, the choice of dataset significantly influences the applicability domain of Generative AI models, as discussed in Section 5.0, under subsection 5.1 Applicability Domain Issues: Balancing Synthesizability and Biological Activity. For example, models trained on ZINC may generate molecules with high synthesizability but limited biological activity, while those trained on ChEMBL may produce biologically active molecules that are challenging to synthesize. This trade-off underscores the need for advanced tools and strategies to balance these priorities, as explored in later sections.

Table 1
List of databases widely used for drug discovery.

Database	Description	Link	Usage Example
ZINC	A public database of commercially available compounds.	https://zinc.docking.org/	Virtual screening
ChEMBL	A database of bioactive compounds that resemble drugs. properties.	https://www.ebi.ac.uk/chembl/	De novo drug design
PubChem	A public repository of chemical molecules and their activities against biological assays.	https://pubchem.ncbi.nlm.nih.gov	Toxicity prediction
DrugBank	A comprehensive resource of drug data, including detailed drug targets, interactions, and pathways.	https://go.drugbank.com/	Drug repurposing
LINCS	A database of gene expression profiles from human cells treated with drugs.	https://clue.io/releases/datadashboard	Drug mechanism analysis
QM9	A quantum chemistry dataset with 134,000 molecules for property prediction.	quantum-machine-9-akac-qm9	Quantum property prediction
MOSES	A benchmark dataset for training and evaluating molecular generation models.	https://github.com/molecularsets/moses	Generative model evaluation
UniProt	Provides protein sequence and annotation data.	https://www.uniprot.org/	Protein function prediction
BindingDB	A database of measured binding affinities for DTI.	https://www.bindingdb.org/bind/index.jsp	Drug-target interaction (DTI)
RCSB PDB	A repository of 3D structural data for large biological molecules, including drug-target complexes.	https://www.rcsb.org/	Protein-protein interaction (PPI)
SureChEMBL	Provides compounds extracted from patent literature.	https://www.surechembl.org/	De novo drug design

3.2. The representation of molecules

Generative AI models rely heavily on molecular representations to effectively design and understand molecular properties The ability of Generative AI models to identify and predict molecular behaviour and properties is strongly influenced by the molecular representation [13]. As illustrated in Fig. 5, there are 3 widely used types of molecular representations are: (1) 1D sequence-based, (2) graph-based, and (3) 3D structure-based (Yang & Cheng, 2025).

The concept of representing molecules as sequences originates from the success of Natural Language Processing (NLP) (Ofer et al., 2021). This approach draws on the similarity between the semantics and grammar of biological structures and human language. Consequently, biological and chemical molecules can be encoded as sequences. The Simplified Molecular Input Line Entry System (SMILES) is the most widely used sequence-based representation [13–15]. It converts the structure of a molecule into a string of characters. SMILES relies on five fundamental syntax rules, allowing molecules to be transformed into vectors that generative models can process efficiently. Another widely adopted approach for molecular representation is the graph-based approach. In this method, nodes represent atoms, and edges correspond to the bonds between them [16–18]. This method is more direct and flexible compared to sequence-based representations. While sequence-based representations are memory-efficient and facilitate easier searching, they lack 3D structural information. Similarly, graph-based approaches also fail to capture spatial details, which are crucial for understanding the functional properties of molecules. The 3D structural representation provides detailed information about a molecule's spatial configuration [17,19,20] including crystallization properties and molecular shapes. This data is invaluable for designing molecules that precisely fit specific biological targets, making it critical for drug discovery. However, obtaining accurate 3D structural data is challenging [21]. It often relies on time-consuming and resource-demanding experimental methods, such as X-ray crystallography or cryo-electron microscopy. To address this limitation, computational tools like AlphaFold2 have emerged as transformative alternatives. AlphaFold2 is an open-access platform that offers the ability to predict high-quality 3D molecular structures directly from amino acid sequences [22,23]. Building on this foundation, the recent release of AlphaFold3 further expands its capabilities. Now, AlphaFold3 can predict high-quality 3D molecular structures and its interactions with DNA, RNA, small molecules, and other proteins [24a,b]. This breakthrough improvement significantly accelerates access to 3D structural data, hence enabling more efficient drug design processes. Additionally, several other molecular representation methods are used in drug discovery, as summarized in Table 2.

Among the various representations listed in Tables 2 and 1D sequence-based representations, particularly SMILES and its derivatives (Canonical SMILES, DeepSMILES, SELFIES), remain the most widely adopted in modern generative AI models due to their simplicity, compactness, and compatibility with deep learning architectures inspired by NLP [25,26,27]. Graph-based representations are gaining popularity [25,28,27], especially with the advancement of Graph Neural Networks (GNNs), which allow more direct modeling of atomic relationships and molecular structures. These models often outperform sequence-based ones in property prediction tasks due to their ability to capture connectivity and substructure-level information. Meanwhile, 3D structure-based representations are the most informative, particularly for tasks involving binding affinity or protein-ligand interaction modelling [29].

Several benchmarking studies have been conducted to compare the predictive performance of these representations in virtual screening for drug discovery. For example [30], compared the performance of 1D sequence-based, 2D graph-based and 3D structure-based

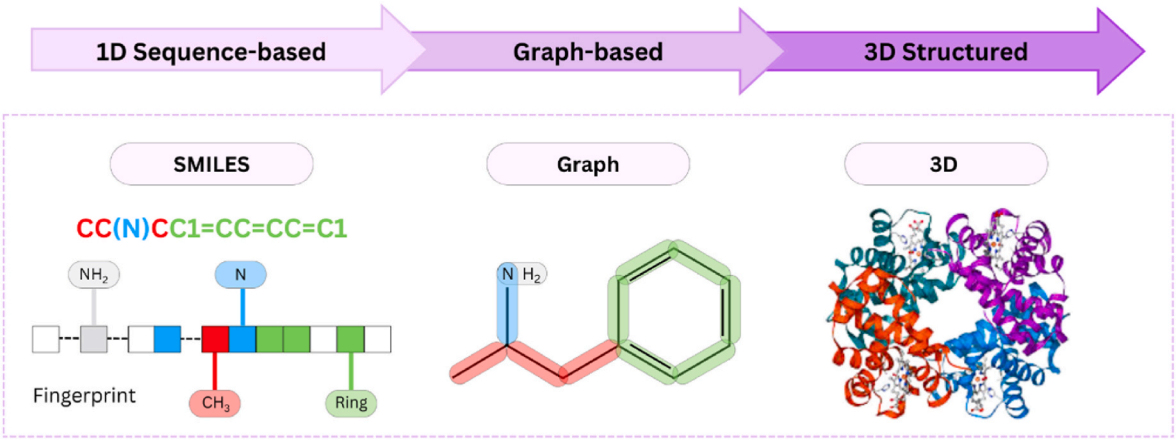

Fig. 5. Overview of molecular representations.

Table 2
Example of commonly used molecular representation methods.

Representations	Descriptions
1D Sequence-Based	
SMILES	Encodes molecular structures as text strings based on basic syntax rules and connectivity.
Canonical SMILES	A standardized version of SMILES that ensure the molecule's uniqueness.
Generic SMILES	A simplified version of SMILES focusing only on basic connectivity.
Isomeric SMILES	An extension of SMILES with connectivity, stereochemistry and isotopic information.
DeepSMILES	A variation of SMILES designed for compatibility with deep learning models.
SELFIES	An alternative to SMILES designed to avoid syntax errors during encoding.
InChI	A hierarchical text-based with connectivity, hydrogen and stereochemistry layers.
InChI Key	A hashed derived from InChI for quick searches and indexing.
2D Graph-Based	
MACC Key	Encodes molecules using 166 binary keys for molecular similarity.
Circular	Encodes molecules using overlapping circular substructures around each atom with local structural information.
Path	Encodes linear paths of atoms and bonds, capturing sequential information within a molecule.
Tree	Represent molecules as hierarchical tress with branching patterns and nested structural relationship.
Atom Pair	Encodes pair of atoms and shortest path distances, with relational and spatial connectivity.
Graph	Represents molecules as graphs with atoms as nodes and bonds as edges
3D Structured-Based	
3D Spatial	Represents molecules based on spatial configuration and coordinates of atom in 3D.
Crystallography	Experimentally determined 3D structures.
AlphaFold2	Provides high-quality 3D structures of molecules from amino acid sequence.
AlphaFold3	An enhanced version of AlphaFold2 that incorporates interactions with DNA, RNA, small molecules, and other proteins.
Voxel	Encodes molecular structures into 3D grids or voxel spaces with electron density.

representations, finding that four representative 3D structure-based models often outperform the others terms of accuracy and computational speed. Moreover, 3D structure-based approaches can more effectively predict real biological interactions and simulate the actual binding process [30]. More recently, 3D-aware models like 3DGT-DDI [31] and DimeNet [32] have shown superior performance on benchmarks such as DrugBank and QM9, especially when 3D information is available. Nevertheless, the choice of representation often depends on the specific application, available data, and computational resources. Thus, an integrated approach that combines multiple representations is increasingly being explored to leverage the strengths of each format.

3.3. *Generative AI models*

Generative AI are categorized as autoregressive [24a,b] and non-autoregressive models [33]. Autoregressive Generative AI models produce output sequentially, generating each part step by step. These models predict the next output based on previously generated outputs. For example, ChatGPT generates one word at a time, progressively constructing complete sentences by connecting these words. In contrast, non-autoregressive Generative AI models generate output simultaneously, without relying on previously produced outputs. This independence allows non-autoregressive models to produce results faster than autoregressive models.

Among the various Generative AI architectures, the earliest and oldest model is the Energy-based Model (EBM) [34]. EBMs have their roots in statistical mechanics and work by associating an energy function to each data point, where lower energy corresponds to more likely configurations [34]. Despite their long history, EBMs are not as commonly used in modern applications due to their computational challenges, such as the need for expensive sampling techniques like Markov Chain Monte Carlo (MCMC) to approximate the true data distribution [35]. However, they remain valuable for tasks such as generating molecular structures and optimizing protein-ligand interactions, where their flexibility can be harnessed. Although not widely used in mainstream drug discovery compared to newer models, EBMs set the foundation for more advanced generative techniques.

Currently, GANs, Variational Autoencoders (VAEs), Transformers and flow-based model are the most widely used generative models in various fields, including drug discovery. These models can be implemented in either autoregressive or non-autoregressive ways, depending on the task and design. For instance, Transformers are often used in autoregressive settings [36] but can also be adapted for non-autoregressive tasks [37]. Similarly, GANs and VAEs are typically non-autoregressive, as they generate outputs in a single forward pass, but they can be combined with autoregressive components for specific applications [38,39]. In contrast, flow-based models also operate in a non-autoregressive manner, learning invertible mappings between simple and complex distributions, which allows for both efficient sampling and precise control over the data generation process [40]. Understanding these models and their categorization helps in selecting the right approach for different generative tasks.

GANs and VAEs are increasingly utilized in the development of new therapeutic compounds. Meanwhile, Transformers represent more

advanced and complex Generative AI models compared to GANs and VAEs. However, all these models incorporate key elements such as GNNs, Graph Convolutional Networks (GCNs), and Convolutional Neural Networks (CNNs) to process chemical or biological data represented as graphs or images. On the other hand, Recurrent Neural Networks (RNNs) are employed to process molecular representations in sequence-based formats, such as SMILES. Additionally, Reinforcement Learning (RL) is a crucial element in Generative AI applications. RL is used to train models to produce outputs that are optimized to achieve specific objectives. In the context of drug discovery, RL helps optimize molecular properties to meet desired criteria. For instance, conditional VAEs leverage RL to generate molecules that satisfy predefined biological and chemical requirements [41].

3.3.1. GANs

The concept of GANs was introduced by Goodfellow [42]. Goodfellow conceived the idea during a discussion with colleagues, where he proposed the framework of two neural networks: a generator and a discriminator. These networks compete against each other to enhance the quality of generated data. In drug discovery, where the chemical space for molecule selection is vast, GANs can generate new molecules that not only meet the biological criteria but also adhere to the chemical requirements of the target. The generator functions to generate new molecules by taking in the random noise, while the discriminator attempts to differentiate between synthetic/generated data and real data. As illustrated in Fig. 6, the architecture of GANs involves these two networks working in opposition. The adversarial process continues until the discriminator can no longer distinguish between the generated data and the real data, resulting in highly realistic synthetic outputs. These promising approaches show that GANs is suitable to design new therapeutic molecules with improved characteristics.

MolGAN, ORGAN, and ORGANIC are among the first examples of GAN-based models designed for generating new therapeutic molecules. ORGAN and ORGANIC depend entirely on input data provided in the SMILES sequence format. When the model depends solely on SMILES sequences, it must employ the seq2seq method to analyze and process the input data. However, the seq2seq method has a notable limitation: it often produces SMILES outputs that are inconsistent and not entirely accurate. This issue arises because seq2seq models can generate SMILES strings that are syntactically invalid or fail to correspond to real molecules.

Training, performance, and achieving convergence in GANs are unpredictable, cumbersome, unstable, and slow. To achieve optimal performance, hyperparameters must be accurately tuned. Convergence occurs when both components of the GAN model (the generator and the discriminator) have been sufficiently improved and stabilized. This is challenging because it involves training two neural networks simultaneously. If GANs fail to achieve convergence, the models encounter a persistent issue known as mode collapse. Mode collapse occurs when the generator produces limited and unvaried outputs. Theoretically, this issue is difficult to avoid [43].

To improve and achieve better model performance, regularization techniques and the integration of various algorithms can be employed. For instance, ORGAN combines the Wasserstein GAN (WGAN) and SeqGAN frameworks in its final model [44]. Additionally, ORGAN incorporates an objective reinforcement mechanism into the reward function of the RNNs generator. This enhancement is designed to guide the RNNs to produce molecules that meet the desired criteria. Building on the original ORGAN concept, the ORGANIC algorithm was developed as an optimized version of ORGAN [45]. This optimization improves the molecular space representation, enabling the production of more accurate molecules and better overall model performance.

MolGAN, on the other hand, employs a graph-based GAN approach. It is notable for its ability to generate molecules that adhere to desired chemical characteristics while maintaining low computational demands [46]. With the increasing recognition of GANs as a powerful Generative AI model for creating new therapeutic molecules, their adoption has grown significantly. Many studies have leveraged GANs with tailored improvements to advance drug discovery research [47–49].

3.3.2. VAEs

Similar to GANs, which are designed with two primary components, namely, the generator and the discriminator. Autoencoders also consist of two primary components which are the encoder and the decoder. The encoder compresses input data into a low-dimensional latent space, while the decoder reconstructs the low-dimensional data from the latent space back to its original dimension. The main purpose of the autoencoder is to generate new data as accurately as possible by capturing the relationships within the data during the compression process and storing important features in the latent space. The latent space generated by a basic autoencoder is deterministic, meaning it consists of fixed points. For instance, the encoder produces a fixed point in the latent space when input data is fed into it. The encoder will reliably generate the same latent vector if the same data is input again. This deterministic nature of

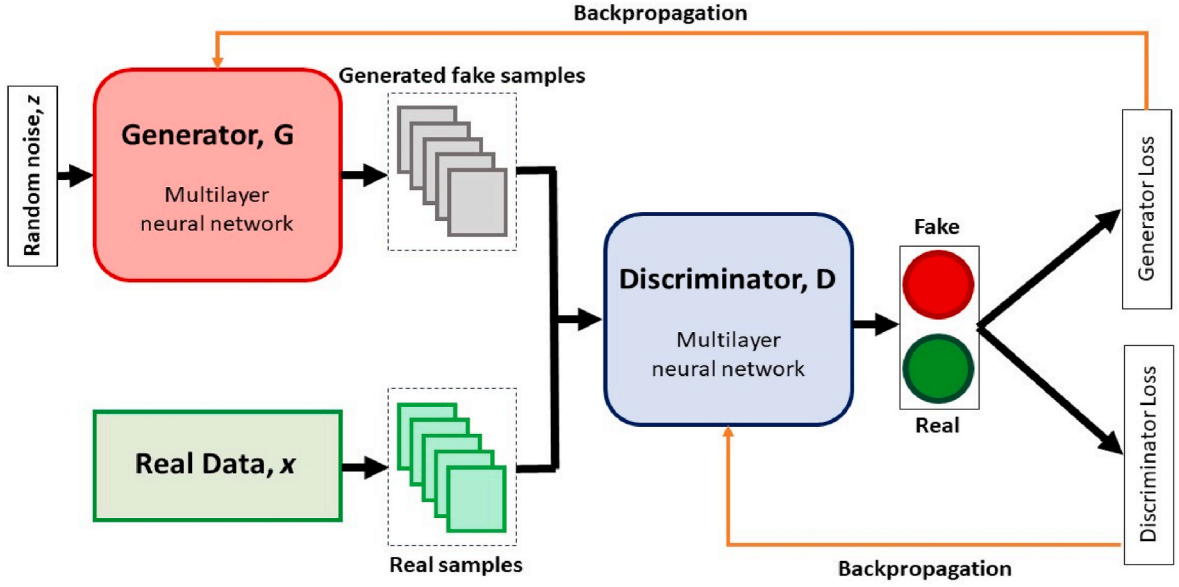

Fig. 6. The architecture of Generative Adversarial Networks (GANs).

the latent space may lead to overfitting, as it eliminates variability and randomness. Overfitting occurs because the neural network only reproduces seen data rather than generating new data by analyzing and understanding the underlying patterns. As a result, the ability of basic autoencoder models to generate novel data is limited and diminished. To address this limitation, VAEs, as shown in Fig. 7 were introduced.

The encoder creates a probabilistic latent space, which is how VAEs vary from simple autoencoders. VAEs compress input data into a lower-dimensional representation in the latent space by mapping it to a probability distribution rather than to fixed points. Specifically, the encoder represents the latent space as a distribution and generates the mean and standard deviation for every input. This probabilistic method promotes continuity in the latent space, ensuring that data points with similar features are located in adjacent areas rather than isolated fixed points. Subsequently, the input data is reconstructed to its original dimensions by the decoder.

This continuity enables VAEs to analyze and identify molecular data more accurately and effectively. Compared to simple autoencoders, the VAE model is better suited for drug discovery applications since it can represent and interpolate within the latent space [50]. This advantage demonstrates that VAEs can learn and optimize parameters to produce new molecules that are identical to the input molecules. Simultaneously, the reconstruction loss is reduced during training.

The capabilities of VAEs have made them one of the most widely recognized and utilized Generative AI models in drug discovery research. Numerous studies have implemented VAEs for this purpose [51–56], with the common framework consisting of an encoder and a decoder. Depending on the type of data representation, three main types of VAEs are commonly used: SMILES-VAE, Graph-VAE, and 3D grid-VAE.

The SMILES-VAE model is designed to handle sequence or string data, typically utilizing stacked GRUs (Gated Recurrent Units) or LSTM (Long Short-Term Memory) networks for its encoder and decoder components. While SMILES-VAEs demonstrate the ability to produce molecular sequences, they still face challenges that hinder the performance of seq2seq models, such as the generation of invalid molecules [57]. These challenges emphasize that molecular graph-based generators are capable of delivering outputs that are entirely valid and practical.

For instance, a study. utilized a conditional graph generator to achieve multi-objective de novo molecule generation [58]. Similarly, they introduced the Junction Tree VAE, where nodes in the junction tree correspond to molecular components or single atoms from the original molecule. The Junction Tree VAE operates in two stages The first stage constructs a scaffold using chemical fragments structured in the form of a junction tree, while the second stage employs a graph-based message-passing network (MPN). MPN is a model derived from Graph Convolutional Networks (GCNs). It aims to combine discrete chemical units into a fully formed molecule. While this method provides precise molecular representations, it faces difficulties in encoding tree structures. Furthermore, the process of decoding latent vectors into new trees remains a challenge, restricting the model's effectiveness in molecular production.

Another notable study proposed a scaffold-based Graph-VAE to generate novel structural or derivative molecules [59]. This approach maintains a specific scaffold as the core element of the molecular substructure while modifying other parts of the molecule. This ensures that the newly generated molecules retain the essential biological and chemical properties of drugs. However, this model has limitations in addressing novel protein targets or unfamiliar proteins requiring new drugs, as it relies on predefined scaffolds. While this method enables the production of diverse molecules, its application to varied molecular data remains restricted.

A study by Ref. [60] is another notable contribution to the drug discovery community. The study employs a three-dimensional (3D) grid-based VAEs method, inspired by image recognition tasks. The method leverages 3D grids to analyze and organize data, facilitating the modeling of spatial arrangements in 3D space. In this 3D grid-based VAE, CNNs are utilized as both the encoder and decoder, as CNNs are specifically designed to analyze data in 3D form. The latent space in this study encodes the spatial configuration of atoms and the chemical characteristics of the input molecules.

In drug discovery, using 3D data is crucial due to its chemical and biological property information [61], which helps determine molecular interactions and binding compatibility with target sites. For instance, 3D data can indicate whether generated molecules share the same physical properties as the binding site, which is essential for effective binding. Additionally, the bioactivity of molecules can be assessed. However, utilizing the 3D grid-based VAE model requires 3D molecular data, which is more challenging to acquire compared to data for models such as SMILES-VAEs (requiring sequence data) or Graph-based VAEs (requiring graph-structured data). The scarcity of 3D molecular data limits the training and performance of these models. Moreover, most existing 3D molecular datasets primarily provide information on least-energy conformations rather than bioactive conformations [62]. These two conformations differ significantly and can lead to the generation of different drug molecules. Least-energy conformations describe the most stable 3D molecular shape under normal conditions. In contrast, bioactive conformations represent the specific 3D shape a molecule adopts when interacting with a biological target. Identifying

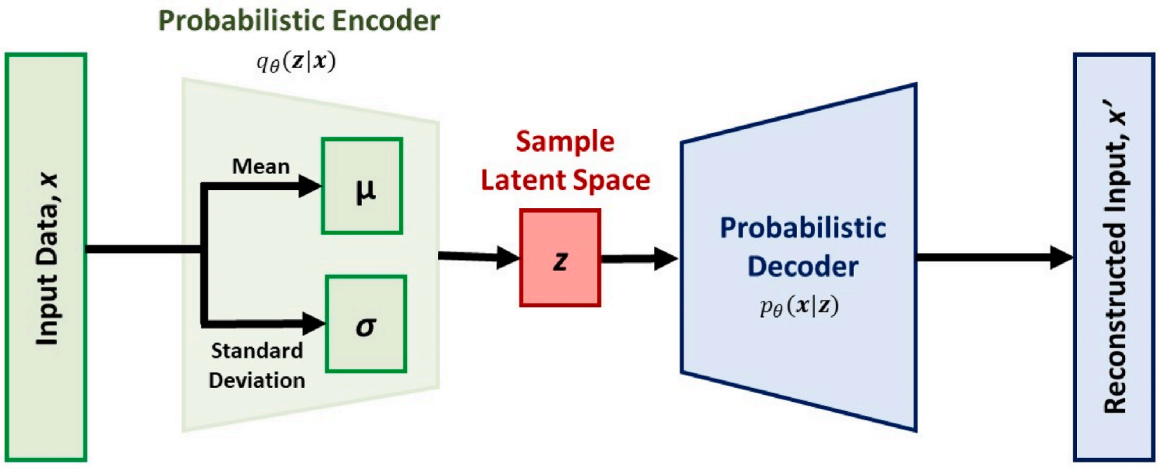

Fig. 7. Architecture of variational autoencoders (VAEs).

bioactive conformations often requires further research, which can be time-consuming and costly.

A study [63] proposed an enhancement to the 3D grid-based VAE model by developing a tool called Libmolgrid. This tool improves the data preparation by representing molecular or atomic data as Gaussian-like densities. This approach ensures smoother and more continuous atomic presence across the 3D grid. Additionally, Libmolgrid assigns each atom type to its channel or layer, simplifying the differentiation of atom types and their interactions. Unlike the original 3D grid-based VAE, where atoms are represented as binary indicators on a grid, Libmolgrid's Gaussian-like density approach provides a more accurate and scalable representation of molecular structures. Furthermore, Libmolgrid is optimized for GPU architectures, significantly accelerating the training process and enabling the model to handle larger datasets effectively. By using Gaussian-like densities and multi-channel grids, Libmolgrid improves the 3D grid-based VAE model's capacity to learn and analyze molecular structures with precision. Although this model has shown outstanding performance, it still requires validation in real-world drug discovery experiments. Nevertheless, other studies have explored Generative AI models for creating molecules based on 3D spatial data [64–66] providing further advancements in this field.

Recently, researchers have increasingly recognized the advantages of the VAE disentangled representation approach. The main goal of this model is to guarantee that every latent variable within the latent vector represents a unique feature or property of the input data [67]. For molecular data, specific latent variables correspond to input data features such as molecular size, shape, charge distribution, or functional groups. This demonstrates that the VAE mechanism is capable of generating molecules effectively. In summary, although no standardized framework exists for generative models, there are strong motivations behind the growing adoption of VAEs. Compared to GANs, VAEs offer significant advantages in drug discovery, where it is critical to produce diverse outputs with desired properties. VAE models offer greater stability and are simpler to train compared to GANs. This approach also helps prevent the mode collapse problem [68–70]. Moreover, GANs require large datasets for training, whereas VAEs can perform effectively even with smaller datasets. Numerous studies have further refined and optimized the VAE model, demonstrating promising performance. As a result, VAEs hold substantial potential for future applications in drug discovery [50–70].

3.3.3. *Flow-based models*

Flow-based models are a class of generative models that learn invertible mappings between simple and complex distributions, enabling both efficient sampling and precise control over the data generation process. Unlike other generative models that rely on approximate sampling methods, flow-based models use exact likelihood evaluation, making them especially powerful for tasks requiring high-quality data generation [71]. In drug discovery, flow-based models can be applied to generate novel molecules with specific biological and chemical properties by directly modeling the underlying data distribution. These models excel in producing molecules with high validity by learning complex distributions over chemical space and ensuring the generated molecules adhere to the desired criteria.

One of the key advantages of flow-based models in drug discovery is their ability to generate diverse and high-quality molecular structures. By transforming simple distributions into more complex ones, these models are able to produce molecules that are not only valid but also novel, offering significant potential in identifying drug candidates with unique properties. Furthermore, the invertible nature of flow-based models allows for easy interpolation between molecules, enabling the exploration of a wide variety of chemical space [72]. This flexibility is beneficial for drug discovery, where the goal is often to generate molecules with specific desired attributes, such as bioactivity, solubility, or binding affinity.

Representative works in flow-based models include NICE (Non-linear Independent Component Estimation) [73], RealNVP (Real-valued Non-Volume Preserving) [74], and the Glow model [63]. NICE [73], introduced tractable calculations for reversible transformations, which are implemented using affine coupling layers. The basic idea behind flow-based models is to learn an invertible mapping between complex distributions and simpler prior distributions. By exploiting exact and tractable likelihood estimation for training, flow models enable efficient one-shot inference and 100 % reconstruction of the training data, making them highly suitable for tasks where precise data generation is essential.

In drug discovery, flow-based models have been applied to molecular graph generation. One notable example is GraphNVP. GraphNVP is the first flow-based model for generating molecular graphs and decomposes the graph generation process into two steps: generating an adjacency tensor and generating node attributes [75]. This approach allows for exact likelihood maximization on the graph using two reversible flows. GraphNVP is capable of generating valid molecules with minimal duplicates, and the learned latent space can be further exploited to generate molecules with desired properties.

Another significant advancement in flow-based models is GraphAF, an autoregressive flow-based model. Unlike previous models that generate graphs in a single-shot manner, GraphAF adopts an iterative sampling process that incorporates chemical domain knowledge, such as valency checking, at each step [76]. This integration of chemical rules ensures that GraphAF generates molecules with 100 % validity, outperforming models like GCPN [77] in terms of training speed. Furthermore, GraphAF can be fine-tuned with reinforcement learning (RL) to optimize molecular properties, demonstrating improved performance compared to other models like JT-VAE [186].

MoFlow, applies a validity correction to the generated graph, which enables efficient molecular graph generation in a single-shot manner while also guaranteeing chemical validity. By learning a continuous latent space through encoding molecular graphs, MoFlow can generate novel and optimized molecules during the decoding process, enhancing the precision of molecular property optimization [71]. Additionally, GraphDF, aims to learn a discrete latent representation of the molecular graphs without introducing real-valued noise. By sequentially sampling discrete latent variables, GraphDF can generate new nodes and edges via invertible transforms, circumventing the computational costs associated with continuous latent spaces. GraphDF has shown state-of-the-art performance in random molecule generation, property optimization, and constrained optimization tasks [78].

Flow-based models' prominent feature is their ability to exactly reconstruct all input data without duplicates, due to the precise likelihood maximization. This exactness is particularly important in drug discovery, where molecular properties can be highly sensitive to minor structural changes, such as activity cliffs. For instance, a flow model could replace a specific atom (node) in a molecule, allowing for more precise modifications of existing molecular structures [71]. This feature makes flow-based models a powerful tool for generating molecules with highly specific and optimized properties for drug discovery.

Overall, flow-based models are becoming increasingly important in drug discovery, offering significant advantages in terms of data reconstruction accuracy, molecule validity, and the ability to generate molecules with desired properties. With the continued development of these models and their integration with other techniques like reinforcement learning, flow-based models are poised to make a lasting impact on the field of drug discovery.

3.3.4. *Transformer*

Vaswani [79] initially presented the Transformer model in the groundbreaking article "Attention is All You Need." This model, while similar to RNNs in its ability to analyze sequential data, offers a more advanced mechanism. RNNs process and analyze sequential data one step at a time, whereas the Transformer model processes and analyzes entire sequences (such as text, audio, video, or time series data)

simultaneously. Transformers are specifically designed to handle full sequences by employing a self-attention mechanism. This mechanism enables parallelization and efficiently captures long-range dependencies, even when elements are far apart within the sequence.

Transformers have found extensive applications, such as in search engines (e.g., Google) and as foundational models for systems like ChatGPT. The encoder and the decoder are the two main parts of the Transformer architecture, just like in VAEs. Both parts depend on the self-attention mechanism. By using self-attention, the Transformer identifies the most significant parts of a sequence to understand its overall context. For instance, in the sentence, *"The child laughed with joy because he was happy to meet his friend,"* the Transformer can discern the relationship between the pronouns *"he"* and *"his"* and the noun *"child,"* despite these words being separated by others.

After applying self-attention, the Transformer utilizes a feed-forward neural network to refine the sequence further. The identification of complex relationships between element is made possible by this feed-forward neural network. Multi-headed attention, which gives the model more flexibility and the ability to focus on multiple elements of the data at once, is frequently used in parallel to carry out the attention process, as shown in Fig. 8. This combination of self-attention, parallelization, and flexibility has established the Transformer as a state-of-the-art model in modern AI.

Before the advent of Transformer models, researchers were required to train neural networks on large labeled datasets, a process that was both time-consuming and costly. The introduction of the Transformer revolutionized this approach by enabling the efficient analysis and identification of relationships within data across billions of records on the Internet. Moreover, its parallel processing mechanism also significantly accelerates execution. For example, SuperGLUE serves as a benchmark for language processing systems that leverage the Transformer model [80].

Transformers, which can analyze sequences both forward and backward, are frequently used to handle text data since it is the most accessible and prevalent type of data. By capturing complex word associations, this bidirectional strategy improves the model's comprehension of sentence meaning. One notable example is the Bidirectional Encoder Representations from Transformers (BERT) model, which is incorporated into Google's search engine algorithm. BERT excels at understanding the context of words within a sentence due to its training on a vast corpus of text data, including books, Wikipedia, and other Internet sources. During training, BERT employs a masked language modeling technique, where certain words are hidden, and the model predicts these hidden words based on the surrounding context. This approach compels BERT to learn the relationships between words more effectively [81].

Following the implementation of BERT in Google's systems, numerous researchers have fine-tuned the model to cater to various languages and applications. One significant application of the BERT model is sentiment analysis [82], which demonstrates its ability to understand relationships within sequences. The Transformer model has become a prominent tool in modern AI due to its effectiveness in analyzing large and complex sequence data. It is also paving the way for advancements in Generative AI, including applications in drug discovery.

Building on this foundation, the use of Transformers in the pharmaceutical domain highlights their versatility in handling sequence data. In this context, molecular structures, such as SMILES strings, represent a unique form of sequential data. A noteworthy example is AlphaFold2 from DeepMind, which is transforming structural biology by using the Transformer model to predict 3D molecule structures from amino acid sequences [22]. Similarly, NVIDIA has developed MegaMolBART, an enhanced Transformer model specifically designed for drug discovery [83]. MegaMolBART builds upon the capabilities of the MolBART model, which processes molecular data in SMILES format. It draws inspiration from NLP models, where SMILES strings are treated as sentences and decomposed into individual characters or groups of characters through a tokenization strategy. The MolBART model has been trained on large chemical structure datasets, enabling it to generalize molecular representations effectively. Beyond molecular

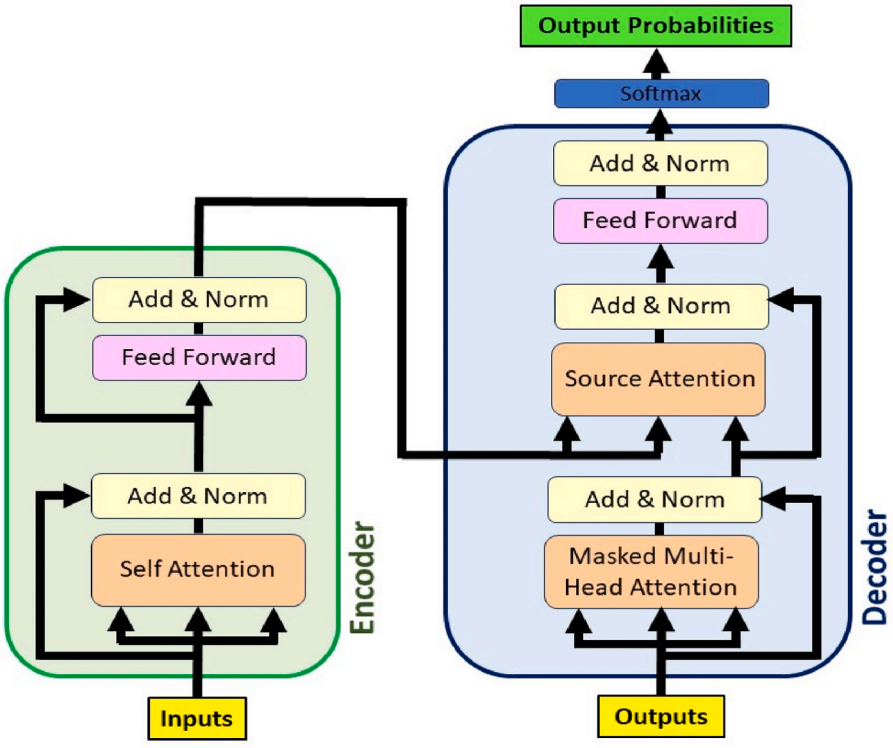

Fig. 8. The architecture of the Transformer-based Model.

generation, MolBART can also optimize SMILES strings and predict molecular properties like solubility and bioactivity, based on desired criteria.

MegaMolBART, an advanced variant of the MolBART model, incorporates NVIDIA's Megatron framework to enhance scalability and performance. This framework facilitates training on extensive molecular data libraries, enabling MegaMolBART to learn complex relationships between molecular structures. As a result, MegaMolBART is well-suited for advanced drug discovery applications, such as predicting molecular interactions, surpassing the general-purpose capabilities of MolBART. With its ability to handle large-scale datasets and deliver high-performance results, MegaMolBART has become a valuable tool for the pharmaceutical industry.

Recent advancements have further demonstrated the potential of Transformers in addressing pressing challenges in the pharmaceutical field, such as identifying bioactive molecules for cancer treatment. For example, the DeepTraSynergy approach leverages the Transformer model to predict drug combination synergy [84]. Thus, indirectly enhancing drug efficacy in cancer treatment. By integrating data such as protein-protein interactions, drug-target interactions, and cell-target interactions, DeepTraSynergy offers a comprehensive framework for studying drug efficacy. The model predicts three key outputs: the toxic effects of drugs, drug-receptor interactions, and drug combination synergy. To optimize its predictions, DeepTraSynergy employs three functional loss functions: toxicity loss, synergy loss, and drug-protein interaction loss. When tested on datasets from DrugComboDB [85] and Oncology-Screen [86], the model achieved accuracy scores of 0.77 and 0.81, respectively, demonstrating its robust performance. The study highlights the critical role of protein-protein interaction data to enhance the model's capacity in predicting drug combination synergy. It emphasizes the importance of incorporating such data for enhanced performance.

3.4. Performance evaluation metrics in drug discovery

Evaluating the performance of Generative AI models in drug discovery involves applying robust metrics to quantify prediction accuracy and the quality of generated molecules. Computational modelling traditionally employs metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) to capture predictive deviations between predicted and actual molecular properties. In contrast, evaluating generative outcomes relies on biological and chemical performance indicators, such as diversity, novelty, validity, and drug-likeness. These evaluation metrics provide insight into the structural and pharmacological suitability of generated compounds. Table 3 lists commonly used performance evaluation metrics in drug discovery research.

However, while Table 3 presents a comprehensive list of commonly used metrics for AI-based drug discovery, current evaluation strategies often fall short in capturing the broader applicability of Generative AI in drug discovery, as discussed in Section 5.5 Standardization of Evaluation Metrics.

3.5. Usability and accessibility of generative AI-based drug discovery tools

The usability and accessibility of Generative AI tools are critical factors in enabling their adoption by researchers or users without computational expertise. This is especially important given the complexity of the datasets, molecular representations, and models discussed in the previous subsections. While Section 3.1 highlighted the widely used datasets like ZINC and ChEMBL for training Generative AI models, and Section 3.2 emphasized the role of molecular representations such as SMILES, graph-based structures, and 3D configurations and the practical application of these tools depends heavily on their ease of use and accessibility. Many platforms now offer user-friendly interfaces that simplify complex programming tasks. This allows researchers to

Table 3
List of performance evaluation metrics for AI-Based model in drug discovery.

Evaluation Metrics	Decriptions in Drug Discovery Scenario	Indicators
Computing Metrics		
Root Mean Square Deviation (RMSD)	Quantifies the average discrepancy between predicted molecular properties and observed values.	Lower
Mean Absolute Error (MAE)	Evaluates the average absolute error in predicted molecular properties or biological activities.	Lower
Mean Squared Error (MSE)	Penalizes larger prediction errors in molecular property estimation	Lower
Cross Entropy Loss	Used in classification tasks like predicting molecule activity classes	Lower
Log Loss	Evaluates the confidence of probabilistic predictions for molecule classifications	Lower
Standard Deviation (SD)	Reflects the variability in predicted molecular properties	Lower for error, higher for diversity
Mean	Calculate the average value of predicted molecular properties	Context-specific
Precision	Assesses the ratio of accurately identified active molecules to all molecules predicted as active.	Higher
Recall	Determines the model's capacity to detect all active molecules present in the dataset.	Higher
F1 Score	Integrates precision and recall evaluating the model's accuracy in detecting active molecules.	Higher
AUC (Area Under the Curve)	Indicates how well the model differentiates active molecules from inactive ones.	Higher
Accuracy	Measures the proportion of correctly classified molecules.	Higher
AUPR (Area Under Precision-Recall Curve)	Particularly useful for imbalanced datasets where active molecules are rare.	Higher
AUROC	Evaluates a model's capability to discriminate between distinct classes.	
Concordance index (CI)		
MCC (Matthews Correlation Coefficient)	Used to measure model performance	Higher
Reconstruction	Evaluates the model's ability to recreate molecules after encoding/decoding.	Higher
KL Divergence	Measures how well the generated molecular property distributions match the target distributions.	Lower
Biological Metrics		
Validity	Evaluates the percentage of generated molecules that are chemically valid.	Higher
Uniqueness	Evaluates the distinctiveness of generated molecules to avoid duplicates.	Higher
Novelty	Assesses how different the generated molecules are from known drugs or training data.	Higher
Recovery	Tests the model's ability to regenerate molecules from the training set, indicating learning efficacy.	Higher
Instability	Measures the stability of generated molecules under different conditions	Lower

(continued on next page)

Table 3 (*continued*)

Evaluation Metrics	Decriptions in Drug Discovery Scenario	Indicators
Similarity	Compares generated molecules to known drugs.	Higher when targeting similarity, lower when aiming for novelty
Scaffold Similarity	Evaluates the similarity of the molecular scaffolds to those in known reference datasets.	
Tanimoto Coefficient	Calculates the similarity of molecular fingerprints to reference molecules.	
Antimicrobial/ Anticancer	Assesses the potential activity of generated molecules against specific targets or diseases.	Higher
Docking	Evaluates how well molecules bind to a protein target.	Higher
QED (Quantitative Estimate of Drug-likeness)	Quantifies drug-likeness based on desirable chemical properties	Higher

focus on the biological and chemical aspects of drug discovery. For instance, PandaOmics and Chemistry42, the drug discovery platforms developed by Insilico Medicine, provides a graphical user interface (GUI). This allows users to perform tasks such as target identification, hit generation, and lead optimization without interacting with the underlying code.

Similarly, AlphaFold3, developed by DeepMind, offers a web-based interface that enables to predict 3D protein structures by providing the sequences of proteins, nucleic acids, and other molecules as input [24a,b]. This advanced tool models the interactions between proteins, DNA, RNA, small molecules, and even chemical modifications. This tool will generate a joint 3D structure of the entire complex and illustrating how these molecules fit together [125]. This makes the tool accessible to a broad user. These GUIs simplify data input, model execution, and result interpretation. This is particularly beneficial for users who may not have a background in computer science.

However, not all Generative AI tools are equally accessible. As discussed in Section 3.3, models like MegaMolBART and MolGAN often require basic knowledge of Python for customization. This could involve modifying scripts, adjusting hyperparameters, or integrating the tool into a larger computational workflow. While this flexibility is advantageous for researchers with programming skills, it may pose a barrier for those without such expertise. To address this, many tools provide extensive documentation, tutorials, and pre-built scripts to help users get started. For example, AlphaFold's pre-trained models are available on GitHub, along with detailed instructions for deployment. This enables researchers to predict protein structures with minimal setup. GitHub repositories often include step-by-step guides, installation instructions, and example workflows. This makes it easier for researchers or users to implement these tools in their work.

GitHub (https://github.com/) has become a central hub for sharing and accessing Generative AI tools, particularly for researchers in academia and industry. Many Generative AI models, such as AlphaFold, MegaMolBART, and MolBERT, are hosted on GitHub, where developers provide comprehensive documentation and tutorials to guide users through the setup and usage process. For example, the AlphaFold GitHub repository includes detailed instructions for installing dependencies, running predictions, and interpreting results. It also provides example scripts and Jupyter notebooks that demonstrate how to use the tool for specific tasks, such as predicting protein structures from amino acid sequences. This level of detail is invaluable for researchers who may not have prior experience with AI or deep learning. Despite the complexity of Generative AI tools, advancements in user-friendly interfaces, comprehensive documentation, and cloud-based solutions enable researchers without computational expertise to effectively utilize these technologies in drug discovery.

Table 4
Generative AI models used for drug discovery.

Name	Algorithm	Purposes	References
Protein-Protein Interactions			
DeepHomo2.0	Transformer	Predict PPIs sites	[25]
HN-PPISP	Transformer + MLP	Predict PPIs sites	[26]
AGAT-PPIS	Transformer + Graph	Predict PPIs sites	[28]
MaTPIP	Transformer + CNN	Predict PPIs	[27]
ProtInteract	Autoencoder	Predict PPIs	[29]
GACT-PPIS	Transformer + Graph	Predict PPIs sites	[87]
SPIDER	Transformer + Graph	Predict PPIs	[88]
RPI-GGCN	Co-VAE + GNN	Predict PPIs	[89]
MPRL	Autoencoders	Predict PPIs	[90]
Drug-Target Interactions			
DeepDTnet	Autoencoders	Predict DTIs	[91]
TransformerCPI	Transformer	Predict DTIs	[92]
MolTrans	Transformer	Predict DTIs binding site	[93]
HyperAttentionDTI	Transformer + CNN	Predict DTIs	[94]
QuoteTarget	Transformer	Predict DTIs binding site	[95]
TransformerCPI 2.0	Transformer	Predict DTIs binding site	[96]
AttentionSiteDTI	Transformer	Predict DTIs binding site	[97]
TransVAE-DTA	Transformer + VAE	Predict DTIs binding affinity	[98]
AttentionMGT-DTA	Transformer	Predict DTIs binding affinity	[99]
TopoFormer	Transformer	Predict DTIs	[100]
CmhAttCPI	Transformer	Predict DTIs	[101]
GraphormerDTI	Transformer + Graph	Predict DTIs	[102]
CAT-DTI	Transformer	Predict DTIs	[103]
FragXsiteDTI	Transformer + Graph	Predict DTIs	[104]
De novo drug design			
DruGAN	AAE	De novo drug design	[105]
OrGAN	GAN + RL	De novo drug design	[44]
ORGANIC	GAN + RL	De novo drug design	[45]
CVAE	VAE + RNN	De novo drug design	[106]
JTVAE	VAE + RNN	De novo drug design	[51]
LatentGAN	GAN + Latent Vector	De novo drug design	[107]
PaccMann	VAE + RL	De novo drug design	[108]
AlphaDrug	Transformer	De novo drug design	[109]
Pocket2MOL	GNN	De novo drug design	[110]
DrugGPT	Transformer	De novo drug design	[111]
UnCorrupt SMILES	Transformer	De novo drug design	[112]
PETrans	Transformer + TL	De novo drug design	[113]
FSM-DDTR	Transformer	De novo drug design	[114]
DNMG	GAN + TL	De novo drug design	[115]
cMolGPT	Transformer	De novo drug design	[116]
DragonNET	Transformer + VAE	De novo drug design	[117]
MedGAN	WGAN + GCN	De novo drug design	[48]
GxVAEs	VAE	De novo drug design	[54]
Protein Structure Prediction			
trRosetta	Transformer	3D protein structure	[118]
AlphaFold2	Transformer	3D protein structure	[22]
AlphaFold3	Diffusion model	3D protein structure and interactions	[24a,b]
ProteinBERT	Transformer	2D protein structure	[119]
EigenFold	Diffusion	3D protein structure	[120]
ESMFold	Transformer	3D protein structure	[121]
RoseTTAFold All-Atom	MSA Transformer	3D protein structure	[122]
ProteinVAE	VAE	Protein structure	[123]
LIME	Transformer + XAI	Protein secondary structure	[124]

4. Current trends and directions of drug discovery and beyond

Recent advancements in Generative AI-based drug discovery increasingly emphasize the use of transformer-based models. Protein-protein interactions (PPIs), drug-target interactions (DTIs), and de novo drug design represent key areas where transformers are driving innovation, as shown in Table 4. Transformers have become more and more popular because of their exceptional capacity to capture contextual relationships in sequence data, which was sparked by the success of applications such as ChatGPT in text processing and analysis. Unlike graph-based or 3D structural data, sequence data is more readily available, making it a practical and dominant choice within the drug discovery community.

A notable shift in research focus now prioritizes the exploration of PPIs and DTIs. This shift addresses limitations of earlier approaches that primarily generated novel molecules without adequately considering their interaction properties. This omission often led to high failure rates in clinical trials, as generated molecules lacked necessary biological activity or binding specificity. PPIs and DTIs are now pivotal in advancing the hit and lead identification stages of drug discovery as they offering critical insights into molecular interactions. During the hit identification stage, PPIs help uncover interactions between proteins involved in disease pathways. This guides the identification of targetable proteins for therapeutic intervention. Conversely, DTIs make it easier to forecast how small molecules or compounds will interact with certain biological targets. This makes it possible to find early hits that show promise. Transformer-based models significantly enhance this process by leveraging sequence data to predict interactions with high throughput and accuracy, filtering viable candidates from large chemical libraries efficiently.

At the lead identification stage, PPI and DTI data are further utilized to improve the hits that were previously found into leads with better drug-like properties. PPIs are used to evaluate how specific protein interfaces may be disrupted or stabilized by a compound, while DTI analysis assesses the binding affinity, specificity, and selectivity of drug candidates. By incorporating these interaction insights, researchers can prioritize molecules that exhibit both strong binding to the target and minimal off-target effects, reducing the likelihood of failure in subsequent stages of drug development. Transformers are essential in this refinement process due to their capability to extract deeper contextual information from sequence data. This capability allows for the selection of candidates with higher clinical relevance.

By integrating PPIs and DTIs into Generative AI models, researchers aim to improve the biological relevance and success rates of AI-generated compounds, marking a significant evolution in computational strategies for drug discovery. These advancements underscore the growing impact of transformer-based architectures in reshaping medicine design and therapeutic development. By addressing key challenges like binding specificity and biological activity, these models offer a pathway to reduce failure rates in clinical trials. Additionally, they accelerate the drug discovery timeline and enable the creation of more effective therapeutic agents.

4.1. Protein-Protein Interactions (PPIs)

Several studies have explored PPIs using Transformer models, as these interactions are essential in the drug discovery process. Understanding PPIs allows researchers to study the biology of diseases and develop drugs that specifically target diseased proteins. For example [126], utilized DeepHomo2.0 to analyze the interactions of homodimeric proteins. This study sought to address the limitations of the AlphaFold2 model, which, while achieving significant success in predicting protein monomer structures, falls short in accurately modeling interactions between two identical polypeptide chains. DeepHomo2.0 combines Direct-Coupling Analysis (DCA) with the Transformer model, using datasets of homodimeric protein structures with C2 symmetry, sequence identity data from the Protein Data Bank (PDB), and datasets from Critical Assessment of Predicted Interactions (CAPRI). The study showed that DeepHomo2.0 outperformed eight other benchmark models, achieving a precision value of over 70 % when using observed monomer structures and over 60 % when using anticipated monomer structures.

Similarly [127], investigated PPIs using their structure-based model, GACT-PPIS, which integrates an Enhanced Graph Attention Network (EGAT), a Transformer, and a Graph Convolutional Network (GCN). The Transformer component consists of 10 neural network layers and was fine-tuned over 35 epochs to achieve optimal performance. This study utilized three datasets derived from GraphPPIS and reported that GACT-PPIS achieved an accuracy exceeding 80 %. Furthermore, multiple performance metrics demonstrated that GACT-PPIS outperformed several other comparative models. These investigations highlight how Transformer-based topologies can improve protein-protein interaction prediction. This advancement advances the drug development process and aids in the creation of personalized medicines.

4.2. Drug-Target Interactions (DTIs)

In addition to its application in identifying PPIs, Transformer models have also been increasingly used to predict DTIs. DTIs refer to the interactions between drugs and target proteins. Target proteins can be enzymes, receptors, or ion channels. It is critical to determine whether a drug can activate, inhibit, or disrupt the target protein, depending on the intended therapeutic purpose. However, a major limitation of existing methods is their poor representation of drugs, as they often rely solely on SMILES sequences or molecular graphs. These representations, while useful, fail to capture all the critical features of a drug, resulting in suboptimal model performance. To address this issue, researchers have begun combining Transformer models with other AI algorithms and incorporating molecular graph data as one of the input data types.

For instance 128, combined a Transformer model with multilayer graph information to identify the structural features of drugs. This approach indirectly improved the study of DTIs and enhanced the understanding of their importance in drug discovery. This study utilized three input data types: SMILES sequence strings of the drug, target protein sequences, and drug-target relationship data. To overcome the problem of limited drug representation, the input SMILES sequences were transformed into molecular structure maps obtained from PubChem. Meanwhile, target protein sequence data were sourced from the KEGG database, and drug-target relationship data were extracted from the DrugBank database. The study achieved remarkable results, with an area under the curve (AUC) of 90.24 %, an area under the precision-recall curve (AUPR) of 77.11 %, an F1-score of 79.31 %, and an accuracy rate of 85.15 %. These results indicate that the DeepMGT-DTI model outperformed previously used DTI models, such as Transformer-CPI.

Given the promising potential demonstrated in DTIs when incorporating molecular graphs of drugs, researchers began to adopt molecular graphs as a primary input. This approach ensures that the model receives comprehensive and detailed information about the desired drug. Thus, enable it to better capture critical features and interactions necessary for accurately identifying DTIs. Additionally, recent studies have used molecular graphs and sequences as input data. One study employed a Transformer-based model consisting of 12 Graph Transformer layers and 3 stacked 1D-Convolutional Neural Networks (1D-CNNs) layers, which were optimized to achieve high performance. One such model, GraphormerDTI, combines a Graph Transformer Neural Network with 1D-CNNs to effectively model molecular structures [25]. The performance of GraphormerDTI was evaluated by comparing molecular graphs and amino acid sequences from three datasets: DrugBank, Davis [77], and KIBA [129]. The molecular graph data of the drug provides essential information, including functional groups, charge distribution, hydrophobicity, and molecular flexibility. These factors influence how the

drug interacts with the target protein. Meanwhile, the sequence data of the protein's amino acids captures critical details, such as binding or catalytic residues. These details help predict how the protein will respond to modulation by the drug. Both data types are crucial for accurately predicting drug-target interactions. They collectively provide the structural and biochemical context necessary for understanding the interaction mechanisms. The study results showed that the GraphormerDTI model delivered the best performance across all three datasets, highlighting its effectiveness in predicting drug-target interactions (DTIs).

In addition, the study [127] utilized two types of drug-related data, namely SMILES sequence data and molecular graphs of the drug, as well as one dataset of target protein sequences. These datasets were extracted from Davis and KIBA. The proposed model, GSATDTA, builds on the Transformer architectures' self-attention mechanism. The aim is to predict the binding affinity between drugs and target proteins. The model employed three layers of Bi-directional Gated Recurrent Units (BiGRU) to analyze and store contextual information from the SMILES sequences of drugs, while GNNs were utilized to analyze and learn the topological structures of the molecular graph structures of the drugs. Subsequently, information from the two kinds of drug-related data was then integrated using a graph-sequence attention method. For the amino acid sequence input data, an efficient Transformer was used to analyze and capture long-range relationships within the amino acid sequences. This study evaluated performance on the Davis and KIBA datasets, and the results demonstrated that GSATDTA outperformed other models. GSATDTA achieved mean squared error (MSE) values of 0.200 and 0.126, concordance index (CI) values of 0.906 and 0.902, and rm^2 values of 0.732 and 0.790 on the Davis and KIBA datasets. These results demonstrate the notable progress made in these tasks and demonstrate how well Transformer-based models predict DTIs and drug-target affinities (DTAs).

4.3. De novo drug design

De novo drug design is an in-silico approach used to create new drug molecules from scratch rather than modifying existing molecules. This method creates new chemical molecules capable of interacting with the target protein's binding pocket efficiently by using the structural insights of the target protein. Two main strategies are employed in de novo drug design: structure-based design and ligand-based design. In structure-based design, details of the target protein are utilized to create compounds that precisely fit the active site of the target protein. Conversely, ligand-based design focuses on creating molecules that share similarities with known active compounds. Both approaches benefit from information derived from PPIs and DTIs. Combining PPIs and DTIs data improves the development of molecules with enhanced binding affinities and functional characteristics.

Most Generative AI models used in de novo drug design prioritize virtual screening or the generation of molecules with optimized pharmacological and physicochemical properties. However, during the development process, these models often neglect the functional information of the target protein. In contrast, the Transformer architecture can incorporate protein-specific functional data to generate molecules tailored to distinct protein targets. As example, AlphaDrug. It exemplifies a de novo drug design method that harnesses Transformers, Monte Carlo Tree Search (MCTS), and docking scores to design new drug candidates or molecules for specific protein targets. The Transformer model in AlphaDrug improves efficiency in learning protein sequence information [109]. Input data in the form of target protein sequences from BindingDB is used to generate ligand molecules with strong binding affinities. A key innovation in AlphaDrug's Transformer architecture is the hierarchical skip connections between the protein encoder and the drug decoder. Thus, it can enhance information flow and enable better target-specific molecule generation. The performance of AlphaDrug surpasses other Generative AI-based models, such as LiGANN, across various evaluation metrics, including docking score, uniqueness, octanol-water partition coefficient (logP), quantitative estimate of drug-likeness (QED), synthetic accessibility (SA), and natural product-likeness (NP-likeness). Remarkably, even with 1D sequence-based input data, AlphaDrug outperforms models designed to utilize 3D molecular structure representations. Despite its success, AlphaDrug still faces challenges inherent to the MCTS method, necessitating further refinement.

Similar to AlphaDrug, CMolGPT employs the Transformer model to generate target-specific molecules (Wang et al., 2023). However, CMolGPT uniquely uses SMILES notation, a 1D sequence-based representation of chemical structures, as input data from the MOSES dataset. Unlike AlphaDrug, CMolGPT is trained without explicit target protein information, instead incorporating randomness into the sampling process to promote molecular diversity and uniqueness. The architecture of CMolGPT draws inspiration from NLP models, particularly GPT. In NLP, models like GPT are trained on vast text corpus to predict subsequent words based on context, using unsupervised learning. Similarly, CMolGPT applies a GPT-like framework to learn chemical structures. The model consists of a molecular sequence encoder and a decoder designed to predict chemical tokens sequentially. The encoder-decoder mechanism captures both small-scale and large-scale chemical features. Thus, enables the generation of syntactically valid and chemically diverse molecules. Training CMolGPT involves optimizing a likelihood-based objective function that balances chemical validity, uniqueness, and diversity. By incorporating randomness during sampling, CMolGPT can explore unexplored chemical space. This makes it a powerful model for generating innovative molecular candidates across various datasets, including EGFR, HTR1A, and S1PR1.

5. Challenges and opportunities: generative AI in drug discovery

Failures in the pharmaceutical field are often seen as inevitable due to the critical nature of the domain, where even the smallest issue can have significant repercussions. This field demands precise analysis to identify suitable molecules, as evidenced by setbacks in Generative AI-driven drug discovery. For instance, EXS-21546, a cancer drug candidate discovered by the UK-based Exscientia, was discontinued during Phase 1/2 trials. Similarly, Recursion Pharmaceuticals faced clinical setbacks despite no failures during clinical trials. Such occurrences raise questions about the capability and efficiency of Generative AI in accelerating drug discovery while reducing associated costs. However, it is crucial to recognize that the limitations observed in Generative AI models may not always stem from the models themselves but rather from the quality of datasets used during training.

5.1. Applicability domain issues: Balancing Synthesizability and Biological Activity

One of the most critical challenges in Generative AI-driven drug discovery is the applicability domain of the generated molecules, particularly the trade-off between synthesizability and biological activity. The applicability domain refers to the range of conditions under which a Generative AI model can reliably produce molecules that are both chemically viable and biologically relevant. This concept is crucial because the success of drug discovery depends on the ability to generate molecules that not only exhibit strong binding affinity to target proteins but are also feasible to synthesize in a laboratory setting.

Datasets like ZINC focus on commercially available and synthetically accessible compounds, making them ideal for generating molecules with high synthesizability [130]. However, these molecules may lack the necessary biological activity to effectively interact with target proteins, limiting their therapeutic potential. Conversely, datasets like ChEMBL prioritize biologically active compounds, which are often more complex and challenging to synthesize [131]. This trade-off highlights a fundamental limitation in the applicability domain of Generative AI models.

For example, models trained on ZINC may produce molecules that are easy to synthesize but lack biological activity [130], while those trained on ChEMBL may generate biologically active molecules that are difficult or expensive to produce [131]. This separation of chemical and biological datasets can lead to failures in clinical trials, as the generated molecules may only possess half of the required properties.

To address this challenge, researchers have developed several strategies to balance synthesizability and biological activity within the applicability domain of Generative AI models. One of the key approaches involves transfer learning [132–134], a technique that enables models to leverage knowledge gained from one domain to improve performance in another. In the context of drug discovery, transfer learning typically involves fine-tuning a model pre-trained on large, synthetically focused datasets (such as ZINC) using smaller, more specialized biological datasets (such as ChEMBL). This process helps the model retain its ability to generate molecules that are synthesizable while improving its capacity to produce biologically active compounds [133]. By transferring knowledge from the broad chemical domain to the more biologically oriented domain, the model can generate molecules that are not only feasible to synthesize but also exhibit stronger binding affinities and relevant biological activity.

Several recent studies have explored the use of transfer learning in drug discovery. For instance Ref. [132], demonstrated the application of transfer learning to fine-tune a model trained on synthetic datasets for the generation of molecules with desired biological activities. This approach was shown to significantly enhance the model's ability to generate compounds with both synthetic accessibility and bioactivity. Similarly [135], employed transfer learning to adapt models trained on chemical structure data to predict drug-target interactions, showing its potential in producing more effective drug candidates. These studies highlight how transfer learning can effectively bridge the gap between synthesizability and biological activity in drug discovery.

An extension of transfer learning that has been particularly useful in generative models is teacher forcing. Teacher forcing is a method often used during the training of sequence-to-sequence models, such as those applied in molecular generation. In teacher forcing, the model is trained on real data points from the target distribution, rather than relying solely on its own predictions during training. This ensures that the model is guided towards producing realistic, biologically relevant outputs. In the context of drug discovery, teacher forcing can be employed to refine the generation of molecules by directly feeding the model biological data during training [54], which helps steer the model toward generating molecules with desirable biological characteristics. By integrating these feedback mechanisms, the model is able to more effectively balance the complexity of biological activity with the practical constraints of chemical synthesis.

In addition to transfer learning and teacher forcing, other techniques, such as domain adaptation [136] and meta-learning [137], have shown promise in drug discovery. Domain adaptation involves modifying a model to perform well across different domains [136], such as chemical datasets and biological datasets, by adapting the model's parameters to the specific requirements of each domain. This technique can be particularly useful in drug discovery, where the task is to adapt chemical models to biological or clinical data. Another related technique is meta-learning, where models are trained to learn how to adapt quickly to new tasks with limited data. This approach could be beneficial in drug discovery, where new therapeutic targets or molecular properties may not have abundant data available for training [137]. Both techniques allow models to generalize across different drug discovery tasks, improving their adaptability and performance in a real-world setting.

Another promising strategy is the integration of multiomics data (such as genomics, proteomics, and metabolomics) into Generative AI models. By incorporating multiomics data, models can better capture the complex relationships between molecular structure, biological function, and disease pathways. For example, models like DeepTraSynergy and GraphormerDTI leverage protein-protein interaction (PPI) and drug-target interaction (DTI) data to generate molecules with enhanced binding specificity and biological relevance. These models not only predict the interactions between drugs and target proteins but also optimize the generated molecules for synthesizability, ensuring that they are both biologically active and chemically viable.

However, beyond generation, a crucial step in validating the real-world potential of these molecules is their in-silico evaluation [138]. While generative models can suggest molecules with promising theoretical properties, computational validation tools provide an essential bridge between design and experimental verification [139]. Among the most widely used tools is AutoDock Vina, a molecular docking software that simulates the binding of small molecules to target proteins, providing quantitative estimates of binding affinity and interaction strength [140,141]. Alongside docking, molecular dynamics simulations [142] and ADMET predictions [143] are also employed to assess the stability, toxicity, and pharmacokinetic behavior of candidate compounds. These computational validation approaches offer scalable, cost-effective means to prioritize molecules with favorable profiles, significantly reducing the risk of failure in subsequent experimental and clinical stages. By integrating these validation methods into the generative workflow, researchers can better assess the practical viability of AI-designed compounds. This ensures that the selected candidates are not only biologically and chemically meaningful within the model's applicability domain but also possess the structural and functional attributes required to proceed through the drug development pipeline.

5.2. Interpretability of generated molecules

Secondly, another significant challenge is the interpretability of the newly generated drug molecules. Generative AI analyses numerous variables at once while operating in high-dimensional spaces. This complexity makes it challenging to identify which features influence molecule generation and the reasons behind the production of particular molecules. Due to this lack of interpretability, scientists are unable to reliably assess the pharmacokinetic profiles and biological activity of the compounds they make, which slows down the drug discovery process.

To overcome this challenge, researchers can utilize attention mechanisms and explainable AI (XAI) methods. Attention mechanisms, such as those used in Transformer models, can highlight important molecular features and interactions. It will provide insights into why certain molecules are generated. For example, in Transformer-based models, the self-attention mechanism can identify which atoms or functional groups contribute most to a molecule's binding affinity or synthesizability. This allows researchers to understand the model's decision-making process and validate the generated molecules.

Besides, XAI tools can further enhance interpretability by visualizing the decision-making process of Generative AI models. For instance, feature attribution methods like SHAP (SHapley Additive exPlanations) can quantify the contribution of each input feature (e.g., molecular fragments or protein-ligand interactions) to the model's predictions. By providing a clear explanation of how the model generates molecules, XAI techniques enable researchers to trust and refine the outputs of Generative AI models. Additionally, validation through real-world experiments is crucial for ensuring efficacy and safety. For example, molecular docking simulations and in vitro assays can be used to verify the binding affinity and biological activity of generated molecules. Thus, providing a bridge between computational predictions and experimental validation.

5.3. Data scarcity and novelty

Third, data scarcity and the issue of novelty are closely related challenges in Generative AI-driven drug discovery. Generative AI models require extensive, high-quality data for training, but such data is often scarce, particularly for rare diseases. While public repositories like PubChem and ChEMBL offer substantial biological and chemical data,

suitable and high-quality data are often inaccessible. Insufficient data during training may lead to overfitting, where models generate molecules nearly identical to existing drugs instead of exploring new chemical spaces. This lack of diversity limits the potential for discovering truly novel and innovative compounds, as the models become too specialized in replicating the training data rather than generating novel molecules.

To address these challenges, researchers can employ a combination of innovative techniques. Data augmentation generates synthetic or modified versions of existing data, helping to expand limited datasets and reduce overfitting. For example, SMILES enumeration and molecular graph augmentation can create diverse representations of the same molecule, increasing the variability of the training data. Transfer learning leverages knowledge from abundant datasets to improve model performance on smaller datasets. For instance, a model pre-trained on a large chemical dataset like ZINC can be fine-tuned on a smaller, disease-specific dataset to generate molecules with targeted biological activity. One-shot learning enables models to learn effectively from minimal data by leveraging prior knowledge and analogies to known molecules.

To enhance novelty, diversity-promoting techniques can be implemented. This will encourage the exploration of underrepresented chemical spaces. For example, reinforcement learning (RL) can be used to reward the generation of molecules that are dissimilar to existing compounds in the training data. Enforcing dissimilarity constraints during molecule generation ensures the production of unique and innovative compounds. Furthermore, graph-based or three-dimensional molecule representations can capture structural complexities while preserving interpretability. Techniques like Transformers' attention mechanisms can highlight important molecular features, ensuring that the generated molecules possess the necessary characteristics for drug discovery. By addressing data scarcity and novelty together, these approaches enable Generative AI models to explore broader chemical spaces and produce more effective drug candidates.

5.4. Scalability and computational resources

Fourth, scalability and the need for high computational resources are significant challenges in Generative AI-driven drug discovery. Training and deploying Generative AI models require substantial computational power, particularly for large-scale datasets and complex architectures like Transformers, VAEs, and GANs. These models often demand high-performance GPUs and extensive memory, which can be a barrier for researchers in academic or small-scale industrial settings. Additionally, the computational demands increase when working with complex molecular representations, such as 3D structures, which are essential for capturing detailed chemical and biological properties but are resource-intensive to process.

To mitigate these challenges, researchers can leverage cloud-based solutions and model optimization techniques. Platforms like Google Cloud and Amazon Web Services (AWS) provide scalable infrastructure for running AI models, enabling researchers to access high-performance computing resources without significant upfront investments. For example, cloud-based platforms allow researchers to train large-scale models on distributed computing clusters, reducing the time and cost associated with model development. Model optimization techniques, such as pruning and quantization, can reduce the computational demands of Generative AI models. Pruning removes redundant parameters from the model, while quantization reduces the precision of the model's calculations, making it more efficient and accessible.

Furthermore, federated learning allows collaborative model training across multiple institutions without sharing raw data, addressing both scalability and data privacy concerns. In federated learning, models are trained locally on decentralized datasets, and only the model updates (rather than the raw data) are shared with a central server. This approach enables researchers to leverage diverse datasets from multiple sources, improving the robustness and generalizability of Generative AI models. By combining these approaches, researchers can overcome the computational barriers and scale Generative AI models effectively for drug discovery.

5.5. Standardization of Evaluation Metrics

The fifth challenges in Generative AI models for drug discovery is the absence of standardized evaluation procedures. Current performance metrics primarily focus on structural validity, chemical diversity, and synthetic accessibility, yet these metrics fail to capture the broader applicability and biological relevance necessary for real-world drug development. As Generative AI progresses in areas such as protein–-protein interaction modeling, multi-target drug design, and disease modeling, these traditional metrics fall short in assessing the full potential of the models. This gap emphasizes the need for a more comprehensive evaluation framework that incorporates biological functionality, clinical translatability, and task-specific considerations.

One critical issue in the evaluation of Generative AI models is model hallucination. Hallucinated molecules are those that appear chemically plausible but are biologically irrelevant or synthetically impractical. These outputs can mislead downstream analyses, wasting valuable resources and time on compounds that are unlikely to succeed in experimental validation or clinical settings. Traditional evaluation metrics, which often prioritize structural novelty or diversity, may fail to identify such hallucinations, leading to misdirected research efforts.

In addition to hallucinations, current evaluation methods neglect essential pharmacological considerations such as drug-target binding affinity, metabolic stability, off-target effects, and toxicity. These factors are crucial for evaluating a compound's real-world efficacy and safety. Without including these dimensions, models may generate molecules that are synthetically feasible but fail to meet the biological and pharmacological criteria needed for therapeutic success. For example, a compound may pass chemical validity tests and be structurally diverse, yet it could have poor binding affinity or unacceptable toxicity, rendering it unsuitable for further development.

Moreover, the lack of standardized benchmarks and ground truth datasets limits the ability to compare different generative models and evaluate their performance across diverse disease targets and biological systems. The absence of universally accepted benchmarks impedes cross-model validation, making it challenging to determine the true strengths and weaknesses of different approaches.

A further limitation of current evaluation practices is the failure to incorporate human-in-the-loop feedback, which integrates expert knowledge from chemists, biologists, and pharmacologists. This lack of expert curation in the evaluation process can result in AI-generated candidates that are unverified or misprioritized, thus diminishing the reliability of the generative model. Incorporating human feedback, along with wet-lab validation, is essential to refine AI models and ensure that the generated molecules align with biological realities and therapeutic needs.

Finally, many generative models are optimized for a single objective, such as structural diversity or synthetic accessibility, but drug discovery requires a multi-objective optimization approach. In real-world drug development, models need to balance multiple competing objectives, such as drug-likeness, biological activity, and synthetic feasibility. Current models that focus on a single objective may generate molecules that excel in one area but fail to meet other critical criteria, thus limiting their utility in drug development pipelines. A more holistic evaluation framework is necessary to account for these multiple, interconnected objectives and to ensure that generative models produce molecules that are not only synthetically feasible but also biologically relevant.

To overcome these challenges, a more integrated and standardized evaluation framework is needed. This framework should incorporate biologically relevant metrics, such as binding affinity, toxicity predictions, and metabolic stability, alongside traditional metrics like chemical validity and synthetic accessibility. Furthermore,

incorporating multi-modal data from genomics, proteomics, and patient-specific biomarkers will provide a more comprehensive understanding of the biological context in which generative models operate. Collaboration between AI researchers and domain experts in chemistry, biology, and pharmacology will be essential to refine the evaluation process and align AI-generated molecules with the practical requirements of drug discovery.

In conclusion, a more robust and standardized evaluation framework is crucial to advancing the role of Generative AI in drug discovery. By integrating a broader range of metrics, incorporating expert feedback, and emphasizing multi-objective optimization, we can enhance the reliability and applicability of AI-generated molecules, paving the way for more effective and innovative therapeutics.

6. Future directions of generative AI-based drug discovery

Looking ahead, several promising avenues hold the potential to further enhance the applicability domain and overall effectiveness of Generative AI in drug discovery. One key direction is the development of integrated datasets that combine chemical, biological, and multiomics data [144]. This strategy will enable models to simultaneously optimize for synthesizability, biological activity, and other drug-like properties. By creating datasets that bridge the gap between chemical and biological domains, researchers can train models to generate molecules that are both biologically relevant and chemically viable. Collaborative efforts between academic institutions, pharmaceutical companies, and data providers will be essential to curate and share such integrated datasets, fostering innovation in the field.

Next, the incorporation of quantum chemistry [145] and molecular dynamics simulations [142] into Generative AI models represents another exciting frontier. Quantum chemistry calculations can provide a more accurate representation of molecular energy landscapes, ensuring that generated molecules are stable and synthetically feasible. Similarly, molecular dynamics simulations can assess the binding affinity and conformational stability of generated molecules. Thus, offering insights into their biological activity and interaction with target proteins. By integrating these computational techniques, Generative AI models can generate molecules with a higher likelihood of success in preclinical and clinical trials. Hence, will bridge the gap between computational predictions and experimental validation.

Furthermore, the exploration of novel generative architectures, particularly diffusion models [146,147], could revolutionize the field of Generative AI for drug discovery. Diffusion models have emerged as the current state-of-the-art (SOTA) approach due to their ability to generate structurally complex and chemically valid molecules with high fidelity and diversity. These models operate by learning to reverse a stochastic noise process, where they begin by gradually adding noise to a molecular representation and then learn to denoise it step-by-step, effectively generating new molecules from pure noise [147]. This iterative process allows diffusion models to capture intricate molecular patterns and dependencies that are often lost in simpler generative frameworks. A prominent example is AlphaFold3 [24,125], developed by DeepMind, which combines diffusion modeling with structural prediction to model multi-molecular complexes, including protein-ligand and protein-RNA interactions, at atomic resolution. Another impactful tool is DiffDock [148], a generative docking method that uses a denoising diffusion model to predict ligand binding poses within protein pockets, outperforming classical docking algorithms in both speed and accuracy.

Additionally, hybrid architectures such as DM-VAE combine the representational strength of VAEs with the sample quality of diffusion models, supporting tasks like scaffold hopping and lead optimization [149]. However, diffusion models also present challenges. They are computationally intensive, often requiring thousands of iterative steps during inference, and can be difficult to condition on specific molecular properties without complex architectural modifications. Despite these limitations, diffusion models are particularly well-suited for structure-based drug design, especially when modeling 3D molecular interactions or generating ligands with specific spatial constraints. Their integration with other methods, such as quantum simulations or reinforcement learning, promises to further enhance their utility in drug discovery.

Finally, the adoption of federated learning offers a promising solution to address data privacy concerns while enabling collaborative model training across multiple institutions. In federated learning, models are trained locally on decentralized datasets, and only the model updates are shared with a central server [150]. This approach allows researchers to leverage diverse datasets from multiple sources while improving the robustness and generalizability of Generative AI models. By fostering collaboration and data sharing [151] without compromising privacy, federated learning could accelerate the development of more effective Generative AI models for drug discovery.

In conclusion, drug discovery can overcome existing challenges by fostering interdisciplinary collaboration, integrating diverse algorithms, and promoting data access. By addressing the trade-offs between synthesizability and biological activity, improving model interpretability, and leveraging advanced computational techniques, Generative AI models can generate molecules with a higher likelihood of success in preclinical and clinical trials. These advancements could pave the way for groundbreaking innovations in addressing both existing and emerging diseases.

7. Conclusions

Generative AI has the potential to create new treatments and medicines for both existing and emerging diseases while also advancing personalized medicine [152]. It leverages three fundamental models: GANs, VAEs, and Transformers. These models have been extensively explored to improve Generative AI capabilities, enabling faster and more efficient drug discovery. By accelerating this process, Generative AI can also help reduce the significant financial investments typically associated with drug development.

The digitalization initiative within the healthcare and pharmaceutical industries is a key driver behind the progress of Generative AI in drug discovery. This advancement is largely due to the vast amounts of high-quality digital data necessary for processing and analyzing information to generate novel drugs. While digital data is predominantly available in text form, image data remains limited. The success of applications like ChatGPT, which generates coherent and grammatically correct sentences, showcases the capabilities of Transformer models in natural language processing. Similarly, AlphaFold 2.0 demonstrates its potential in predicting the 3D protein structures based on the amino acid sequences. Building upon the limitations of AlphaFold2, the recent release of AlphaFold3 marks a significant advancement in the field. AlphaFold3 extends its predictive capabilities beyond individual protein structures to encompass complex biomolecular assemblies. This enhancement allows for more accurate modeling of biomolecular interactions and offers richer structural insights. The development of AlphaFold3 underscores the rapid progression of technology in this domain, highlighting the continuous efforts to overcome previous limitations and improve predictive accuracy in drug discovery.

The ease of accessing digital data in text form, coupled with the achievements of ChatGPT, has driven the increasing adoption of the Transformer model within the drug discovery community. Current research trends emphasize the generation of drugs with favorable pharmacokinetic and pharmacodynamic properties, ensuring that newly developed molecules can successfully progress through the clinical trial phase. Pharmacokinetics, often associated with PPIs, and pharmacodynamics, related to DTIs, are crucial factors in this process.

However, Generative AI still faces significant challenges, including insufficient data quantity and quality, as well as a lack of interpretability in both data and models. To address these issues, strategies such as explainable AI are being implemented to improve model transparency

and reliability. Further collaboration among researchers and industry stakeholders is also essential to maximize collective knowledge and establish standardized protocols for AI-driven drug design. With ongoing advancements, Generative AI will continue to empower the scientific community to develop more precise and effective medications targeting a wide range of diseases. Overall, Generative AI shows remarkable promise in drug discovery, and further research in this area may result in groundbreaking innovations soon.

CRediT authorship contribution statement

Ainin Sofia Jusoh: Writing – review & editing, Writing – original draft, Visualization, Methodology, Data curation, Conceptualization. **Muhammad Akmal Remli:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition. **Mohd Saberi Mohamad:** Writing – review & editing. **Tristan Cazenave:** Writing – review & editing. **Chin Siok Fong:** Writing – review & editing.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author, Ainin Sofia Jusoh, used ChatGPT to improve sentence conciseness and readability. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Muhammad Akmal bin Remli reports financial support was provided by Sultan Mizan Antarctic Research Foundation. Muhammad Akmal bin Remli reports financial support was provided by France-Malaysia Collaboration Programme for Joint Research. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was financially supported by the Sultan Mizan Antarctic Research Foundation-Yayasan Penyelidikan Antartika Sultan Mizan (YPASM) under grant number R/YPASM/B1500/01850A/003/2024/01332, and the France-Malaysia Collaboration Programme for Joint Research (MyTIGER 2023).

Data availability

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

References

[1] P. Mittal, H. Chopra, K.P. Kaur, R.K. Gautam, New Drug Discovery Pipeline, Elsevier eBooks, 2023, pp. 197–222, https://doi.org/10.1016/b978-0-323-99137-7.00003-4.

[2] F.W. Pun, I.V. Ozerov, A. Zhavoronkov, AI-powered therapeutic target discovery, Trends Pharmacol. Sci. 44 (9) (2023) 561–572, https://doi.org/10.1016/j.tips.2023.06.010.

[3] A. Lavecchia, Transform drug discovery and development with generative artificial intelligence, Gener. Artif. Intell.Biomed. Smart Health Inform. (2025) 489–537, https://doi.org/10.1002/9781394280735.ch25.

[4] F. Ren, A. Aliper, J. Chen, H. Zhao, S. Rao, C. Kuppe, I.V. Ozerov, M. Zhang, K. Witte, C. Kruse, V. Aladinskiy, Y. Ivanenkov, D. Polykovskiy, Y. Fu, E. Babin, J. Qiao, X. Liang, Z. Mou, H. Wang, A. Zhavoronkov, A small-molecule TNIK inhibitor targets fibrosis in preclinical and clinical models, Nat. Biotechnol. (2024), https://doi.org/10.1038/s41587-024-02143-0.

[5] X. Zeng, F. Wang, Y. Luo, S. Kang, J. Tang, F.C. Lightstone, E.F. Fang, W. Cornell, R. Nussinov, F. Cheng, Deep generative molecular design reshapes drug discovery, Cell Rep. Med. 3 (12) (2022) 100794, https://doi.org/10.1016/j.xcrm.2022.100794.

[6] M. Bordukova, N. Makarov, R. Rodriguez-Esteban, F. Schmich, M.P. Menden, Generative artificial intelligence empowers digital twins in drug discovery and clinical trials, Expet Opin. Drug Discov. 19 (1) (2023) 33–42, https://doi.org/10.1080/17460441.2023.2273839.

[7] N. Rane, ChatGPT and similar generative Artificial Intelligence (AI) for smart industry: role, challenges and opportunities for industry 4.0, industry 5.0 and society 5.0, SSRN Electron. J. (2023), https://doi.org/10.2139/ssrn.4603234.

[8] Z. Xianyu, C. Correia, C.Y. Ung, S. Zhu, D.D. Billadeau, H. Li, The rise of hypothesis-driven artificial intelligence in oncology, Cancers 16 (4) (2024) 822, https://doi.org/10.3390/cancers16040822.

[9] W. Ying, D. Wang, X. Hu, J. Qiu, J. Park, Y. Fu, Revolutionizing biomarker discovery: leveraging generative AI for bio-knowledge-embedded continuous space exploration. CIKM '24: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, 2024, pp. 5046–5053, https://doi.org/10.1145/3627673.3680041.

[10] Y. Bian, X. Xie, Generative chemistry: drug discovery with deep learning generative models, J. Mol. Model. 27 (3) (2021), https://doi.org/10.1007/s00894-021-04674-8.

[11] V. Romanelli, C. Cerchia, A. Lavecchia, Unlocking the Potential of Generative Artificial Intelligence in Drug Discovery, Springer eBooks, 2024, pp. 37–63, https://doi.org/10.1007/978-3-031-46238-2_3.

[12] K. Zhang, X. Yang, Y. Wang, Y. Yu, N. Huang, G. Li, X. Li, J.C. Wu, S. Yang, Artificial intelligence in drug development, Nat. Med. 31 (1) (2025) 45–59, https://doi.org/10.1038/s41591-024-03434-4.

[13] H. Öztürk, E. Ozkirimli, A. Özgür, A comparative study of SMILES-Based compound similarity functions for drug-target interaction prediction, BMC Bioinf. 17 (1) (2016), https://doi.org/10.1186/s12859-016-0977-x.

[14] N. Aksamit, A. Tchagang, Y. Li, B. Ombuki-Berman, Hybrid fragment-SMILES tokenization for ADMET prediction in drug discovery, BMC Bioinf. 25 (1) (2024), https://doi.org/10.1186/s12859-024-05861-z.

[15] B. Xu, Y. Lu, C. Li, L. Yue, X. Wang, N. Hao, T. Fu, J. Chen, SMILES-Mamba: Chemical mamba foundation models for drug ADMET prediction. http://arxiv.org/abs/2408.05696, 2024.

[16] Y. Chen, Z. Li, Z. Wan, H. Yu, X. Wei, CTAGE: curvature-based topology-aware graph embedding for learning molecular representations. http://arxiv.org/abs/2307.13275, 2023.

[17] H. Huang, L. Sun, B. Du, W. Lv, Learning joint 2-D and 3-D graph diffusion models for complete molecule generation, IEEE Transact. Neural Networks Learn. Syst. 35 (9) (2024) 11857–11871, https://doi.org/10.1109/TNNLS.2024.3416328.

[18] Z. Huang, J. Yu, W. He, J. Yu, S. Deng, C. Yang, W. Zhu, X. Shao, AI-enhanced chemical paradigm: from molecular graphs to accurate prediction and mechanism, J. Hazard Mater. 465 (2024) 133355.

[19] A. Alakhdar, B. Poczos, N. Washburn, Diffusion Models in De Novo Drug Design. https://doi.org/10.1021/acs.jcim.4c01107, 2024.

[20] A. Zholus, M. Kuznetsov, R. Schutski, R. Shayakhmetov, D. Polykovskiy, S. Chandar, A. Zhavoronkov, BindGPT: a scalable framework for 3D molecular design via language modeling and reinforcement learning. http://arxiv.org/abs/2406.03686, 2024.

[21] A.U. Mazlan, N.A. Sahabudin, M.A. Remli, N.S.N. Ismail, M.S. Mohamad, H. W. Nies, N.B. Abd Warif, A review on recent progress in machine learning and deep learning methods for cancer classification on gene expression data, Processes 9 (8) (2021) 1466, https://doi.org/10.3390/pr9081466.

[22] P. Bryant, G. Pozzati, A. Elofsson, Improved prediction of protein-protein interactions using AlphaFold2, Nat. Commun. 13 (1) (2022), https://doi.org/10.1038/s41467-022-28865-w.

[23] Z. Yang, X. Zeng, Y. Zhao, R. Chen, AlphaFold2 and its applications in the fields of biology and medicine. In signal transduction and targeted therapy, Springer Nat. 8 (1) (2023), https://doi.org/10.1038/s41392-023-01381-z.

[24] (a) J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A.J. Ballard, J. Bambrick, S.W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, J.M. Jumper, Accurate structure prediction of biomolecular interactions with AlphaFold 3, Nature 630 (8016) (2024) 493–500. https://doi.org/10.1038/s41586-024-07487-w;
(b) S. Bond-Taylor, A. Leach, Y. Long, C.G. Willcocks, Deep generative modelling: a comparative review of VAEs, GANs, normalizing flows, energy-based and autoregressive models, IEEE Trans. Pattern Anal. Mach. Intell. 44 (11) (2022) 7327–7347, https://doi.org/10.1109/TPAMI.2021.3116668.

[25] P. Lin, Y. Yan, S. Huang, DeepHomo2.0: improved protein–protein contact prediction of homodimers by transformer-enhanced deep learning, Briefings Bioinf. 24 (1) (2022), https://doi.org/10.1093/bib/bbac499.

[26] Y. Kang, Y. Xu, X. Wang, B. Pu, X. Yang, Y. Rao, J. Chen, HN-PPISP: a hybrid network based on MLP-mixer for protein–protein interaction site prediction, Briefings Bioinf. 24 (1) (2022), https://doi.org/10.1093/bib/bbac480.

[27] S. Ghosh, P. Mitra, MaTPIP: a deep-learning architecture with eXplainable AI for sequence-driven, feature mixed protein-protein interaction prediction, Comput. Methods Progr. Biomed. 244 (2023) 107955, https://doi.org/10.1016/j.cmpb.2023.107955.

[28] Y. Zhou, Y. Jiang, Y. Yang, AGAT-PPIS: a novel protein–protein interaction site predictor based on augmented graph attention network with initial residual and identity mapping, Briefings Bioinf. 24 (3) (2023), https://doi.org/10.1093/bib/bbad122.

[29] F. Soleymani, E. Paquet, H.L. Viktor, W. Michalowski, D. Spinello, ProtInteract: a deep learning framework for predicting protein–protein interactions, Comput.

Struct. Biotechnol. J. 21 (2023) 1324–1348, https://doi.org/10.1016/j.csbj.2023.01.028.
[30] W.-H. Shin, X. Zhu, M. Bures, D. Kihara, Three-dimensional compound comparison methods and their application in drug discovery, Molecules 20 (7) (2015) 12841–12862, https://doi.org/10.3390/molecules200712841.
[31] H. He, G. Chen, C. Yu-Chian Chen, 3DGT-DDI: 3D graph and text based neural network for drug–drug interaction prediction, Briefings Bioinf. 23 (3) (2022), https://doi.org/10.1093/bib/bbac134.
[32] Johannes Gasteiger, Janek Groß, Stephan Günnemann, Directional message passing for molecular graphs, Mach. Learn. (2022), https://doi.org/10.48550/arXiv.2003.03123.
[33] Y. Xiao, L. Wu, J. Guo, J. Li, M. Zhang, T. Qin, T.-Y. Liu, A survey on non-autoregressive generation for neural machine translation and beyond, IEEE Trans. Pattern Anal. Mach. Intell. 45 (10) (2023) 11407–11427, https://doi.org/10.1109/TPAMI.2023.3277122.
[34] Yang Song, P. Kingma Diederik, How to train your energy-based models, Mach. Learn. (2021), https://doi.org/10.48550/arXiv.2101.03288.
[35] Chelsea Finn, Christiano Paul, Pieter Abbeel, Levine Sergey, A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models, Mach. Learn. (2016), https://doi.org/10.48550/arXiv.1611.03852.
[36] A. Katharopoulos, A. Vyas, N. Pappas, F. Fleuret, Transformers Are RNNs: Fast Autoregressive Transformers with Linear Attention, 2020.
[37] X. Zhang, T. Ling, Z. Jin, S. Xu, Z. Gao, B. Sun, Z. Qiu, J. Wei, N. Dong, G. Wang, G. Wang, L. Li, M. Abdul-Mageed, L.V.S. Lakshmanan, F. He, W. Ouyang, C. Chang, S. Sun, π-PrimeNovo: an accurate and efficient non-autoregressive deep learning model for de novo peptide sequencing, Nat. Commun. 16 (1) (2025), https://doi.org/10.1038/s41467-024-55021-3.
[38] H. Li, S. Yu, J. Principe, Causal recurrent variational autoencoder for medical time series generation, Proc. AAAI Conf. Artif. Intell. 37 (7) (2023) 8562–8570, https://doi.org/10.1609/aaai.v37i7.26031.
[39] Q. Bai, J. Ma, T. Xu, AI Deep Learning Generative Models for Drug Discovery, Springer eBooks, 2024, pp. 461–475, https://doi.org/10.1007/978-3-031-46238-2_23.
[40] H. Liao, J. He, K. Shu, Generative model with dynamic linear flow, IEEE Access 7 (2019) 150175–150183, https://doi.org/10.1109/access.2019.2947567.
[41] M. Sofi, A. Dhanpratap Singh, T. Ahmed Teli, De novo Molecular Generation augmentation for drug discovery using Deep Learning Approaches: a Comparative Study of variational Autoencoders, J. Angiotherapy 8 (10) (2024) 1–13, https://doi.org/10.25163/angiotherapy.8109996.
[42] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, GAN(Generative adversarial nets), J. Japan Soc. Fuzzy Theor. Intell. Inf. 29 (5) (2017) 177, https://doi.org/10.3156/jsoft.29.5_177_2.
[43] H. Thanh-Tung, T. Tran, Catastrophic forgetting and mode collapse in GANs. 2022 International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1–10, https://doi.org/10.1109/ijcnn48605.2020.9207181.
[44] G.L. Guimaraes, B. Sánchez-Lengeling, P.L.C. Farias, A. Aspuru-Guzik, Objective-reinforced generative adversarial networks (ORGAN) for sequence generation models, arXiv (Cornell University) (2017), https://doi.org/10.48550/arxiv.1705.10843.
[45] B. Sanchez-Lengeling, C. Outeiral, G.L. Guimaraes, A. Aspuru-Guzik, Optimizing distributions over molecular space. An Objective-Reinforced Generative Adversarial Network for Inverse-design Chemistry (ORGANIC), Theor. Comput. Chem. (2017), https://doi.org/10.26434/chemrxiv.5309668.v3.
[46] D.C. Nicola, T. Kipf, MolGAN: an implicit generative model for small molecular graphs, arXiv (2018), https://doi.org/10.48550/arxiv.1805.11973. Cornell University.
[47] T. Song, Y. Ren, S. Wang, P. Han, L. Wang, X. Li, A. Rodriguez-Patón, DNMG: deep molecular generative model by fusion of 3D information for de novo drug design, Methods 211 (2023) 10–22, https://doi.org/10.1016/j.ymeth.2023.02.001.
[48] B. Macedo, I.R. Vaz, T.T. Gomes, MedGAN: optimized generative adversarial network with graph convolutional networks for novel molecule design, Sci. Rep. 14 (1) (2024), https://doi.org/10.1038/s41598-023-50834-6.
[49] D. Manu, J. Yao, W. Liu, X. Sun, GRAphGANFED: a federated generative framework for graph-structured molecules towards efficient drug discovery, IEEE ACM Trans. Comput. Biol. Bioinf 21 (2) (2024) 240–253, https://doi.org/10.1109/tcbb.2024.3349990.
[50] D.B. Kell, S. Samanta, N. Swainston, Deep learning and generative methods in cheminformatics and chemical biology: navigating small molecule space intelligently, Biochem. J. 477 (23) (2020) 4559–4580, https://doi.org/10.1042/bcj20200781.
[51] W. Jin, R. Barzilay, T. Jaakkola, Junction tree variational autoencoder for molecular graph generation, arXiv (2018), https://doi.org/10.48550/arxiv.1802.04364. Cornell University.
[52] J. Lim, S. Ryu, J.W. Kim, W.Y. Kim, Molecular generative model based on conditional variational autoencoder for de novo molecular design, J. Cheminf. 10 (1) (2018), https://doi.org/10.1186/s13321-018-0286-7.
[53] J. Cadow, J. Born, M. Manica, A. Oskooei, M.R. Martínez, PaccMann: a web service for interpretable anticancer compound sensitivity prediction, Nucleic Acids Res. 48 (W1) (2020) W502–W508, https://doi.org/10.1093/nar/gkaa327.
[54] C. Li, Y. Yamanishi, GxVAEs: two joint VAEs generate hit molecules from gene expression profiles, Proc. AAAI Conf. Artif. Intell. 38 (12) (2024) 13455–13463, https://doi.org/10.1609/aaai.v38i12.29248.
[55] Y. Wang, P. Ding, C. Wang, S. He, X. Gao, B. Yu, RPI-GGCN: prediction of RNA–protein interaction based on interpretability gated graph convolution neural network and Co-Regularized variational autoencoders, IEEE Transact. Neural Networks Learn. Syst. (2024) 1–15, https://doi.org/10.1109/tnnls.2024.3390935.
[56] C. Li, Y. Matsukiyo, Y. Yamanishi, Gx2Mol: de Novo Generation of Hit-like Molecules from Gene Expression Profiles via Deep Learning, arXiv (2024), https://doi.org/10.48550/arxiv.2412.19422. Cornell University.
[57] Z. Alperstein, A. Cherkasov, J.T. Rolfe, All SMILES variational autoencoder for molecular property prediction and optimization, in: Challenges and Advances in Computational Chemistry and Physics, 2023, pp. 85–115, https://doi.org/10.1007/978-3-031-28401-4_4.
[58] W. Jin, R. Barzilay, T.S. Jaakkola, Junction tree variational autoencoder for molecular graph generation, Int. Conf. Mach. Learn. (2018) 2323–2332, in: http://proceedings.mlr.press/v80/jin18a/jin18a.pdf.
[59] J. Lim, S. Hwang, S. Moon, S. Kim, W.Y. Kim, Scaffold-based molecular design with a graph generative model, Chem. Sci. 11 (4) (2019) 1153–1164, https://doi.org/10.1039/c9sc04503a.
[60] S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2012) 221–231, https://doi.org/10.1109/tpami.2012.59.
[61] A. Lavecchia, Navigating the frontier of drug-like chemical space with cutting-edge generative AI models, Drug Discov. Today 29 (9) (2024) 104133, https://doi.org/10.1016/j.drudis.2024.104133.
[62] T.I. Adelusi, A.K. Oyedele, I.D. Boyenle, A.T. Ogunlana, R.O. Adeyemi, C. D. Ukachi, M.O. Idris, O.T. Olaoba, I.O. Adedotun, O.E. Kolawole, Y. Xiaoxing, M. Abdul-Hammed, Molecular modeling in drug discovery, Inform. Med. Unlocked 29 (2022) 100880, https://doi.org/10.1016/j.imu.2022.100880.
[63] J. Sunseri, D.R. Koes, Libmolgrid: graphics processing unit accelerated molecular gridding for deep learning applications, J. Chem. Inf. Model. 60 (3) (2020) 1079–1084, https://doi.org/10.1021/acs.jcim.9b01145.
[64] H. Wu, X. Ye, J. Yan, QVAE-Mole: the quantum VAE with spherical latent variable learning for 3-D molecule generation, Adv. Neural Inf. Process. Syst. 37 (2025) 22745–22771.
[65] Y. Xiang, G. Huang, X. Shi, G. Hao, G. Yang, 3D molecular generation models expand chemical space exploration in drug design, Drug Discov. Today (2024) 104282, https://doi.org/10.1016/j.drudis.2024.104282.
[66] C. Xu, L. Zheng, Q. Fan, Y. Liu, C. Zeng, X. Ning, H. Liu, K. Du, T. Lu, Y. Chen, Y. Zhang, Progress in the application of artificial intelligence in molecular generation models based on protein structure, Eur. J. Med. Chem. 277 (2024) 116735, https://doi.org/10.1016/j.ejmech.2024.116735.
[67] X. Wang, H. Chen, S. Tang, Z. Wu, W. Zhu, Disentangled representation learning, IEEE Trans. Pattern Anal. Mach. Intell. 46 (12) (2024) 9677–9696, https://doi.org/10.1109/tpami.2024.3420937.
[68] G. Baykal, M. Kandemir, G. Unal, EdVAE: mitigating codebook collapse with evidential discrete variational autoencoders, Pattern Recogn. 156 (2024) 110792, https://doi.org/10.1016/j.patcog.2024.110792.
[69] Y. Ichikawa, K. Hukushima, Learning dynamics in linear VAE: posterior collapse threshold, superfluous latent space pitfalls, Speedup KL Annealing 238 (2024).
[70] T. Song, J. Sun, X. Liu, W. Peng, Scale-VAE: Preventing Posterior Collapse in Variational Autoencoder, 2024.
[71] C. Zang, F. Wang, MoFlow: an invertible flow model for generating molecular graphs. KDD '20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, https://doi.org/10.1145/3394486.3403104.
[72] C. Bilodeau, W. Jin, T. Jaakkola, R. Barzilay, K.F. Jensen, Generative models for molecular discovery: recent advances and challenges, Wiley Interdiscip. Rev. Comput. Mol. Sci. 12 (5) (2022), https://doi.org/10.1002/wcms.1608.
[73] L. Dinh, D. Krueger, Y. Bengio, NICE: non-Linear independent components estimation, arXiv (2014), https://doi.org/10.48550/arxiv.1410.8516. Cornell University.
[74] L. Dinh, J. Sohl-Dickstein, S. Bengio, Density estimation using real NVP, arXiv (Cornell University) (2016), https://doi.org/10.48550/arxiv.1605.08803.
[75] K. Madhawa, K. Ishiguro, K. Nakago, M. Abe, GraphNVP: an invertible flow model for generating molecular graphs, arXiv (Cornell University) (2019), https://doi.org/10.48550/arxiv.1905.11600.
[76] C. Shi, M. Xu, Z. Zhu, W. Zhang, M. Zhang, J. Tang, GraphAF: a flow-based autoregressive model for molecular graph generation, arXiv (Cornell University) (2020), https://doi.org/10.48550/arxiv.2001.09382.
[77] M.I. Davis, J.P. Hunt, S. Herrgard, P. Ciceri, L.M. Wodicka, G. Pallares, M. Hocker, D.K. Treiber, P.P. Zarrinkar, Comprehensive analysis of kinase inhibitor selectivity, Nat. Biotechnol. 29 (11) (2011) 1046–1051, https://doi.org/10.1038/nbt.1990.
[78] Y. Luo, K. Yan, S. Ji, GRAPHDF: a discrete flow model for molecular graph generation, arXiv (Cornell University) (2021), https://doi.org/10.48550/arxiv.2102.01189.
[79] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, arXiv (Cornell University) 30 (2017) 5998–6008. https://arxiv.org/pdf/1706.03762v5.
[80] P.-E. Sarlin, D. Detone, T. Malisiewicz, A. Rabinovich, E. Zurich, Superglue: Learning Feature Matching with Graph Neural Networks, 2020.
[81] J. Devlin, M.-W. Chang, K. Lee, K.T. Google, A.I. Language, BERT: pre-training of deep bidirectional transformers for language understanding. https://github.com/tensorflow/tensor2tensor, 2018.
[82] H. Xu, B. Liu, L. Shu, P.S. Yu, BERT post-training for review reading comprehension and aspect-based sentiment analysis, arXiv (Cornell University) (2019), https://doi.org/10.48550/arxiv.1904.02232.

[83] D. Reidenbach, M. Livne, R.K. Ilango, M. Gill, J. Israeli, Improving small molecule generation using mutual information machine. http://arxiv.org/abs/2208.09016, 2023.
[84] F. Rafiei, H. Zeraati, K. Abbasi, J.B. Ghasemi, M. Parsaeian, A. Masoudi-Nejad, DeepTraSynergy: drug combinations using multimodal deep learning with transformers, Bioinformatics 39 (8) (2023), https://doi.org/10.1093/bioinformatics/btad438.
[85] H. Liu, W. Zhang, B. Zou, J. Wang, Y. Deng, L. Deng, DrugCombDB: a comprehensive database of drug combinations toward the discovery of combinatorial therapy, Nucleic Acids Res. (2019), https://doi.org/10.1093/nar/gkz1007.
[86] J. O'Neil, Y. Benita, I. Feldman, M. Chenard, B. Roberts, Y. Liu, J. Li, A. Kral, S. Lejnine, A. Loboda, W. Arthur, R. Cristescu, B.B. Haines, C. Winter, T. Zhang, A. Bloecher, S.D. Shumway, An unbiased oncology compound screen to identify novel combination strategies, Mol. Cancer Therapeut. 15 (6) (2016) 1155–1162, https://doi.org/10.1158/1535-7163.MCT-15-0843.
[87] L. Meng, H. Zhang, GACT-PPIS: prediction of protein-protein interaction sites based on graph structure and transformer network, Int. J. Biol. Macromol. 283 (2024) 137272, https://doi.org/10.1016/j.ijbiomac.2024.137272.
[88] Y. Kupershmidt, S. Kasif, R. Sharan, SPIDER: constructing cell-type-specific protein-protein interaction networks, Bioinf. Adv. 4 (1) (2024), https://doi.org/10.1093/bioadv/vbae130.
[89] Y. Wang, P. Ding, C. Wang, S. He, X. Gao, B. Yu, RPI-GGCN: prediction of RNA–protein interaction based on interpretability gated graph convolution neural network and Co-Regularized variational autoencoders, IEEE Transact. Neural Networks Learn. Syst. (2024) 1–15, https://doi.org/10.1109/tnnls.2024.3390935.
[90] V.T.D. Nguyen, T.S. Hy, Multimodal pretraining for unsupervised protein representation learning, Biology Methods Protoc. 9 (1) (2024), https://doi.org/10.1093/biomethods/bpae043.
[91] X. Zeng, S. Zhu, W. Lu, Z. Liu, J. Huang, Y. Zhou, J. Fang, Y. Huang, H. Guo, L. Li, B.D. Trapp, R. Nussinov, C. Eng, J. Loscalzo, F. Cheng, Target identification among known drugs by deep learning from heterogeneous networks, Chem. Sci. 11 (7) (2020) 1775–1797, https://doi.org/10.1039/c9sc04336e.
[92] L. Chen, X. Tan, D. Wang, F. Zhong, X. Liu, T. Yang, X. Luo, K. Chen, H. Jiang, M. Zheng, TransformerCPI: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments, Bioinformatics 36 (16) (2020) 4406–4414, https://doi.org/10.1093/bioinformatics/btaa524.
[93] K. Huang, C. Xiao, L.M. Glass, J. Sun, MolTrans: Molecular interaction transformer for drug–target interaction prediction, Bioinformatics 37 (6) (2020) 830–836, https://doi.org/10.1093/bioinformatics/btaa880.
[94] Q. Zhao, H. Zhao, K. Zheng, J. Wang, HyperAttentionDTI: improving drug–protein interaction prediction by sequence-based deep learning with attention mechanism, Bioinformatics 38 (3) (2021) 655–662, https://doi.org/10.1093/bioinformatics/btab715.
[95] J. Chen, Z. Gu, Y. Xu, M. Deng, L. Lai, J. Pei, QuoteTarget: a sequence-based transformer protein language model to identify potentially druggable protein targets, Protein Sci. 32 (2) (2022), https://doi.org/10.1002/pro.4555.
[96] L. Chen, Z. Fan, J. Chang, R. Yang, H. Guo, Y. Zhang, T. Yang, C. Zhou, Z. Chen, C. Zheng, X. Hao, K. Zhang, R. Cui, Y. Ding, N. Zhang, X. Luo, H. Jiang, S. Zhang, M. Zheng, Drug design and repurposing with a sequence-to-drug paradigm, bioRxiv (2022), https://doi.org/10.1101/2022.03.26.485909.
[97] M. Yazdani-Jahromi, N. Yousefi, A. Tayebi, E. Kolanthai, C.J. Neal, S. Seal, O. O. Garibay, AttentionSiteDTI: an interpretable graph-based model for drug-target interaction prediction using NLP sentence-level relation classification, Briefings Bioinf. 23 (4) (2022), https://doi.org/10.1093/bib/bbac272.
[98] C. Zhou, Z. Li, J. Song, W. Xiang, TransVAE-DTA: transformer and variational autoencoder network for drug-target binding affinity prediction, Comput. Methods Progr. Biomed. 244 (2023) 108003, https://doi.org/10.1016/j.cmpb.2023.108003.
[99] H. Wu, J. Liu, T. Jiang, Q. Zou, S. Qi, Z. Cui, P. Tiwari, Y. Ding, AttentionMGT-DTA: a multi-modal drug-target affinity prediction using graph transformer and attention mechanism, Neural Netw. 169 (2023) 623–636, https://doi.org/10.1016/j.neunet.2023.11.018.
[100] G. Wei, D. Chen, J. Liu, TopoFormer: multiscale topology-enabled structure-to-sequence transformer for protein-ligand interaction predictions, Research Square (Research Square) (2024), https://doi.org/10.21203/rs.3.rs-3640878/v1.
[101] M. Wang, J. Wang, Z. Rong, L. Wang, Z. Xu, L. Zhang, J. He, S. Li, L. Cao, Y. Hou, K. Li, A bidirectional interpretable compound-protein interaction prediction framework based on cross attention, Comput. Biol. Med. 172 (2024) 108239, https://doi.org/10.1016/j.compbiomed.2024.108239.
[102] M. Gao, D. Zhang, Y. Chen, Y. Zhang, Z. Wang, X. Wang, S. Li, Y. Guo, G.I. Webb, A.T. Nguyen, L. May, J. Song, GraphormerDTI: a graph transformer-based approach for drug-target interaction prediction, Comput. Biol. Med. 173 (2024) 108339, https://doi.org/10.1016/j.compbiomed.2024.108339.
[103] X. Zeng, W. Chen, B. Lei, CAT-DTI: cross-attention and transformer network with domain adaptation for drug-target interaction prediction, BMC Bioinf. 25 (1) (2024), https://doi.org/10.1186/s12859-024-05753-2.
[104] A.K. Yalabadi, M. Yazdani-Jahromi, N. Yousefi, A. Tayebi, S. Abdidizaji, O. O. Garibay, FragXsiteDTI: revealing responsible segments in drug-target interaction with transformer-driven interpretation, in: Lecture Notes in Computer Science, 2024, pp. 68–85, https://doi.org/10.1007/978-1-0716-3989-4_5.
[105] A. Kadurin, S. Nikolenko, K. Khrabrov, A. Aliper, A. Zhavoronkov, druGAN: an Advanced Generative Adversarial Autoencoder Model for de Novo Generation of New Molecules with Desired Molecular Properties in Silico, Mol. Pharm. 14 (9) (2017) 3098–3104, https://doi.org/10.1021/acs.molpharmaceut.7b00346.
[106] 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 Cent. Sci. 4 (2) (2018) 268–276, https://doi.org/10.1021/acscentsci.7b00572.
[107] O. Prykhodko, S.V. Johansson, P. Kotsias, J. Arús-Pous, E.J. Bjerrum, O. Engkvist, H. Chen, A de novo molecular generation method using latent vector based generative adversarial network, J. Cheminf. 11 (1) (2019), https://doi.org/10.1186/s13321-019-0397-9.
[108] J. Born, M. Manica, A. Oskooei, J. Cadow, G. Markert, M.R. Martínez, PaccMannRL: de novo generation of hit-like anticancer molecules from transcriptomic data via reinforcement learning, iScience 24 (4) (2021) 102269, https://doi.org/10.1016/j.isci.2021.102269.
[109] H. Qian, C. Lin, D. Zhao, S. Tu, L. Xu, AlphaDrug: protein target specific de novo molecular generation, PNAS Nexus 1 (4) (2022), https://doi.org/10.1093/pnasnexus/pgac227.
[110] W. Feng, L. Wang, Z. Lin, Y. Zhu, H. Wang, J. Dong, R. Bai, H. Wang, J. Zhou, W. Peng, B. Huang, W. Zhou, Generation of 3D molecules in pockets via a language model, Nat. Mach. Intell. 6 (1) (2024) 62–73, https://doi.org/10.1038/s42256-023-00775-6.
[111] Y. Li, C. Gao, X. Song, X. Wang, Y. Xu, S. Han, DrugGPT: a GPT-Based strategy for designing potential ligands targeting specific proteins, bioRxiv (2023), https://doi.org/10.1101/2023.06.29.543848 *(Cold Spring Harbor Laboratory)*.
[112] L. Schoenmaker, O.J.M. Béquignon, W. Jespers, G.J.P. Van Westen, UnCorrupt SMILES: a novel approach to de novo design, J. Cheminf. 15 (1) (2023), https://doi.org/10.1186/s13321-023-00696-x.
[113] X. Wang, C. Gao, P. Han, X. Li, W. Chen, A.R. Patón, S. Wang, P. Zheng, PETrans: De Novo Drug Design with Protein-Specific Encoding Based on Transfer Learning, Int. J. Mol. Sci. 24 (2) (2023) 1146, https://doi.org/10.3390/ijms24021146.
[114] N.R. Monteiro, T.O. Pereira, A.C.D. Machado, J.L. Oliveira, M. Abbasi, J.P. Arrais, FSM-DDTR: end-to-end feedback strategy for multi-objective De Novo drug design using transformers, Comput. Biol. Med. 164 (2023) 107285, https://doi.org/10.1016/j.compbiomed.2023.107285.
[115] T. Song, Y. Ren, S. Wang, P. Han, L. Wang, X. Li, A. Rodriguez-Patón, DNMG: deep molecular generative model by fusion of 3D information for de novo drug design, Methods 211 (2023) 10–22, https://doi.org/10.1016/j.ymeth.2023.02.001.
[116] Y. Wang, H. Zhao, S. Sciabola, W. Wang, cMolGPT: a Conditional Generative Pre-Trained Transformer for Target-Specific De Novo Molecular Generation, Molecules 28 (11) (2023) 4430, https://doi.org/10.3390/molecules28114430.
[117] C. Yamanaka, S. Uki, K. Kaitoh, M. Iwata, Y. Yamanishi, De novo drug design based on patient gene expression profiles via deep learning, Mol. Inform. 42 (8–9) (2023), https://doi.org/10.1002/minf.202300064.
[118] Z. Du, H. Su, W. Wang, L. Ye, H. Wei, Z. Peng, I. Anishchenko, D. Baker, J. Yang, The trRosetta server for fast and accurate protein structure prediction, Nat. Protoc. 16 (12) (2021) 5634–5651, https://doi.org/10.1038/s41596-021-00628-9.
[119] N. Brandes, D. Ofer, Y. Peleg, N. Rappoport, M. Linial, ProteinBERT: a universal deep-learning model of protein sequence and function, Bioinformatics 38 (8) (2022) 2102–2110, https://doi.org/10.1093/bioinformatics/btac020.
[120] B. Jing, E. Erives, P. Pao-Huang, G. Corso, B. Berger, T. Jaakkola, EigenFold: generative protein structure prediction with diffusion models, arXiv (2023), https://doi.org/10.48550/arxiv.2304.02198. Cornell University.
[121] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A.D.S. Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, A. Rives, Evolutionary-scale prediction of atomic-level protein structure with a language model, Science 379 (6637) (2023) 1123–1130, https://doi.org/10.1126/science.ade2574.
[122] R. Krishna, J. Wang, W. Ahern, P. Sturmfels, P. Venkatesh, I. Kalvet, G.R. Lee, F. S. Morey-Burrows, I. Anishchenko, I.R. Humphreys, R. McHugh, D. Vafeados, X. Li, G.A. Sutherland, A. Hitchcock, C.N. Hunter, A. Kang, E. Brackenbrough, A. K. Bera, D. Baker, Generalized biomolecular modeling and design with RoseTTAFold all-atom, Science 384 (6693) (2024), https://doi.org/10.1126/science.adl2528.
[123] S. Lyu, S. Sowlati-Hashjin, M. Garton, ProteinVAE: variational AutoEncoder for translational protein design, bioRxiv (2023), https://doi.org/10.1101/2023.03.04.531110 *(Cold Spring Harbor Laboratory)*.
[124] U. Vignesh, R. Parvathi, K.G. Ram, Ensemble deep learning model for protein secondary structure prediction using NLP metrics and explainable AI, Results Eng. (2024) 103435, https://doi.org/10.1016/j.rineng.2024.103435.
[125] D. Desai, S. v Kantliwala, J. Vybhavi, R. Ravi, H. Patel, J. Patel, Review of AlphaFold 3: transformative advances in drug design and therapeutics, Cureus (2024), https://doi.org/10.7759/cureus.63646.
[126] M. Gao, D. Zhang, Y. Chen, Y. Zhang, Z. Wang, X. Wang, S. Li, Y. Guo, G.I. Webb, A.T. Nguyen, L. May, J. Song, GraphormerDTI: a graph transformer-based approach for drug-target interaction prediction, Comput. Biol. Med. 173 (2024) 108339, https://doi.org/10.1016/j.compbiomed.2024.108339.
[127] X. Yan, Y. Liu, Graph–sequence attention and transformer for predicting drug–target affinity, RSC Adv. 12 (45) (2022) 29525–29534, https://doi.org/10.1039/d2ra05566j.
[128] P. Zhang, Z. Wei, C. Che, B. Jin, DeepMGT-DTI: transformer network incorporating multilayer graph information for drug–target interaction prediction, Comput. Biol. Med. 142 (2022) 105214, https://doi.org/10.1016/j.compbiomed.2022.105214.

[129] J. Tang, A. Szwajda, S. Shakyawar, T. Xu, P. Hintsanen, K. Wennerberg, T. Aittokallio, Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis, J. Chem. Inf. Model. 54 (3) (2014) 735–743, https://doi.org/10.1021/ci400709d.
[130] J.J. Irwin, T. Sterling, M.M. Mysinger, E.S. Bolstad, R.G. Coleman, ZINC: a free tool to discover chemistry for biology, J. Chem. Inf. Model. 52 (7) (2012) 1757–1768, https://doi.org/10.1021/ci3001277.
[131] A.P. Bento, A. Gaulton, A. Hersey, L.J. Bellis, J. Chambers, M. Davies, F.A. Krüger, Y. Light, L. Mak, S. McGlinchey, M. Nowotka, G. Papadatos, R. Santos, J. P. Overington, The ChEMBL bioactivity database: an update, Nucleic Acids Res. 42 (D1) (2014) D1083–D1090, https://doi.org/10.1093/nar/gkt1031.
[132] C. Cai, S. Wang, Y. Xu, W. Zhang, K. Tang, Q. Ouyang, L. Lai, J. Pei, Transfer learning for drug discovery, J. Med. Chem. 63 (16) (2020) 8683–8694, https://doi.org/10.1021/acs.jmedchem.9b02147.
[133] M. Elbadawi, S. Gaisford, A.W. Basit, Advanced machine-learning techniques in drug discovery, Drug Discov. Today 26 (3) (2021) 769–777, https://doi.org/10.1016/j.drudis.2020.12.003.
[134] D. Dana, S. Gadhiya, L. st Surin, D. Li, F. Naaz, Q. Ali, L. Paka, M. Yamin, M. Narayan, I. Goldberg, P. Narayan, Deep learning in drug discovery and medicine; scratching the surface, Molecules 23 (9) (2018) 2384, https://doi.org/10.3390/molecules23092384.
[135] M.R. Prabhu, P. Nancy, R.A.A. Rosaline, A.P. Pandian, A. Devipriya, B. Arunagiri, Exploring machine learning applications and future prospects in drug discovery. 2024 10th International Conference on Communication and Signal Processing (ICCSP), 2024, pp. 708–713, https://doi.org/10.1109/ICCSP60870.2024.10543411.
[136] X. Zhang, Z. Chen, H. Ren, Y. Tian, Knowledge and task-driven multimodal adaptive transfer through LLMs with limited data, IEEE Int. Conf. Bioinform. Biomed. (BIBM) (2024) 5343–5348, https://doi.org/10.1109/BIBM62325.2024.10822808.
[137] G. George, S. Juliet, A comparative study of metric-based meta-learning methods for improving few-shot learning in drug discovery with limited data. 2024 International Conference on Cognitive Robotics and Intelligent Systems (ICC - ROBINS), 2024, pp. 601–606, https://doi.org/10.1109/ICC-ROBINS60238.2024.10533948.
[138] M. Viceconti, F. Pappalardo, B. Rodriguez, M. Horner, J. Bischoff, F. Musuamba Tshinanu, In silico trials: verification, validation and uncertainty quantification of predictive models used in the regulatory evaluation of biomedical products, Methods 185 (2021) 120–127, https://doi.org/10.1016/j.ymeth.2020.01.011.
[139] Y. Hu, Q. Ren, X. Liu, L. Gao, L. Xiao, W. Yu, *In silico* prediction of human organ toxicity via artificial intelligence methods, Chem. Res. Toxicol. 36 (7) (2023) 1044–1054, https://doi.org/10.1021/acs.chemrestox.2c00411.
[140] O. Trott, A.J. Olson, AutoDock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading, J. Comput. Chem. 31 (2) (2010) 455–461, https://doi.org/10.1002/jcc.21334.
[141] J. Eberhardt, D. Santos-Martins, A.F. Tillack, S. Forli, AutoDock vina 1.2.0: new docking methods, expanded force field, and python bindings, J. Chem. Inf. Model. 61 (8) (2021) 3891–3898, https://doi.org/10.1021/acs.jcim.1c00203.
[142] Q. Bai, S. Liu, Y. Tian, T. Xu, A.J. Banegas-Luna, H. Pérez-Sánchez, J. Huang, H. Liu, X. Yao, Application advances of deep learning methods for de novo drug design and molecular dynamics simulation, WIREs Comput. Mol. Sci. 12 (3) (2022), https://doi.org/10.1002/wcms.1581.
[143] P. Sucharitha, K. Ramesh Reddy, S.V. Satyanarayana, T. Garg, Absorption, distribution, metabolism, excretion, and toxicity assessment of drugs using computational tools, in: Computational Approaches for Novel Therapeutic and Diagnostic Designing to Mitigate SARS-CoV-2 Infection, Elsevier, 2022, pp. 335–355, https://doi.org/10.1016/B978-0-323-91172-6.00012-1.
[144] W. Jiang, W. Ye, X. Tan, Y.-J. Bao, Network-based multi-omics integrative analysis methods in drug discovery: a systematic review, BioData Min. 18 (1) (2025) 27, https://doi.org/10.1186/s13040-025-00442-z.
[145] N.S. Blunt, J. Camps, O. Crawford, R. Izsák, S. Leontica, A. Mirani, A.E. Moylett, S.A. Scivier, C. Sünderhauf, P. Schopf, J.M. Taylor, N. Holzmann, Perspective on the current state-of-the-art of quantum computing for drug discovery applications, J. Chem. Theor. Comput. 18 (12) (2022) 7001–7023, https://doi.org/10.1021/acs.jctc.2c00574.
[146] N.S. Blunt, J. Camps, O. Crawford, R. Izsák, S. Leontica, A. Mirani, A.E. Moylett, S.A. Scivier, C. Sünderhauf, P. Schopf, J.M. Taylor, N. Holzmann, Perspective on the current state-of-the-art of quantum computing for drug discovery applications, J. Chem. Theor. Comput. 18 (12) (2022) 7001–7023, https://doi.org/10.1021/acs.jctc.2c00574.
[147] A. Alakhdar, B. Poczos, N. Washburn, Diffusion Models in De Novo Drug Design, J. Chem. Inf. Model. 64 (19) (2024) 7238–7256, https://doi.org/10.1021/acs.jcim.4c01107.
[148] Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, Jaakkola Tommi, DiffDock: diffusion steps, twists, and turns for molecular docking, Biomolecules (2022), https://doi.org/10.48550/arXiv.2210.01776.
[149] M. Lee, V. Pavlovic, Private-shared disentangled multimodal VAE for learning of hybrid latent representations, arXiv (Cornell University) (2020), https://doi.org/10.48550/arxiv.2012.13024.
[150] R. Saidi, T. Moulahi, S. Aladhadh, S. Zidi, Advancing federated learning: optimizing model accuracy through privacy-conscious data sharing. 2024 IEEE 25th International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2024, pp. 64–69, https://doi.org/10.1109/wowmom60985.2024.00022.
[151] N.H. Hazman, R.M. Zawawi, A.S. Jusoh, M.A. Remli, M.C. Leong, M.S. Mohamad, S. Harun, PolarBytes: advancing polar research with a centralized open-source data sharing platform, Environ. Model. Software 185 (2025) 106325, https://doi.org/10.1016/j.envsoft.2025.106325.
[152] Y.W. Choon, Y.F. Choon, N.A. Nasarudin, F. al Jasmi, M.A. Remli, M.H. Alkayali, M.S. Mohamad, Artificial intelligence and database for NGS-based diagnosis in rare disease, Front. Genet. 14 (2024), https://doi.org/10.3389/fgene.2023.1258083.