Large Language Models in Drug Discovery: A Survey

Raghad AbuNasser

Computer Information Systems Department, Jordan University of Science and Technology,
Irbid, Jordan

Abstract

Drug Discovery is a very lengthy and resource-consuming process. However, a variety of advanced Artificial Intelligence (AI) and Deep Learning (DL) techniques are being utilized to accelerate and advance DD, such as Large Language Models (LLMs). This survey is in aim of discovering and comparing the currently available LLMs, their methodologies, used datasets, and the different tasks they are aiding in in the DD process, in particular; de novo drug design, drug-target interaction prediction, masked language models, variational auto encoders, binding affinity prediction, drug repurposing, molecular optimization, activity prediction, contrastive learning for drug-target interaction prediction, and other miscellaneous models. This survey gives insights into future directions and potential in this area.

Keywords

Large Language Model, Drug Discovery, Drug-Target Interaction, Contrastive Learning, High-throughput in-Silico Screening, Transformer, Computational Chemistry

Introduction

The process of Drug Discovery (DD) is very time and resource consuming; to come up with a single compound you will go through an extremely long and complex journey; starting with understanding the disease, then identifying your drug target, and last but not least searching through potential natural and synthetic compounds for the desired activity profile; this process might take billions of dollars and over 10 years before starting clinical testing or the actual production [1]. With the continuously evolving and advancing computational powers, many methods have been utilized and taken advantage of in the DD process; Artificial Intelligence (AI) and Deep Learning (DL) in particular have shown a profound impact in accelerating and advancing the DD process. Large Language Models (LLMs) are now becoming a great promise for researchers in the pharmaceutical industry, they are able to reduce the time, costs, and efforts required for DD [2].

LLMs are basically an architecture of multiple Neural Networks (NNs), harnessing their abilities in Natural Language Processing (NLP) tasks. The NN of a LLM is called a transformer; a transformer has an encoder that deals with input sentences as a sequence of tokens and learns out relations between them based on their positioning, and a decoder that generates an output text based on what has been learned in the training phase; this is what we call a self-attention mechanism [2]. Since the evolution of LLMs, many datasets were created; pre-training datasets, fine-tuning datasets, preference datasets, evaluation datasets, and traditional NLP datasets. In the case of utilizing LLMs for DD, our major drawback is the lack of comprehensive datasets for training [3].

The phases of LLM engineering are as follows: pre-training, fine-tuning, and transfer-learning, alignment, and evaluating [4]. Proper fine-tuning accompanied with prompt-engineering helps comprehend the insufficiency in data. Fine-tuning is used to tailor our LLM for a specific task of NLP [5], while prompt-engineering is to give the model certain examples to use as a guide in future predictions [6].

When providing suitable and comprehensive data we can utilize LLMs in many areas that advance the DD process, such as *de novo* drug design [7], Drug-Target Interaction (DTI) prediction [8], Masked Language Models (MLMs) [9], Variational Auto-Encoders (VAEs) [10], binding affinity prediction, drug repurposing [11], molecular optimization [12], compound activity prediction [13], contrastive learning for DTI prediction [14], and other miscellaneous models.

This survey is the first in literature reviewing the state of LLMs in DD. It consists of six main chapters; Introduction, Review of Literature, Comparative Analysis of Performance and Methodologies, Discussion, Conclusion, and References. In the introduction, we provided a general overview about the leveraging of LLMs in DD, our challenges, and the current state. In the Review of Literature, we are going to investigate the available literature on LLMs applied in DD, compare how similar models are designed, and how that affects their performance. In the Discussion chapter, we are going to show a clear overview of the current state-of-the-art, discuss the most valuable insights from literature, and to define potential areas for improvement in future work. This structure is represented in Figure 1.

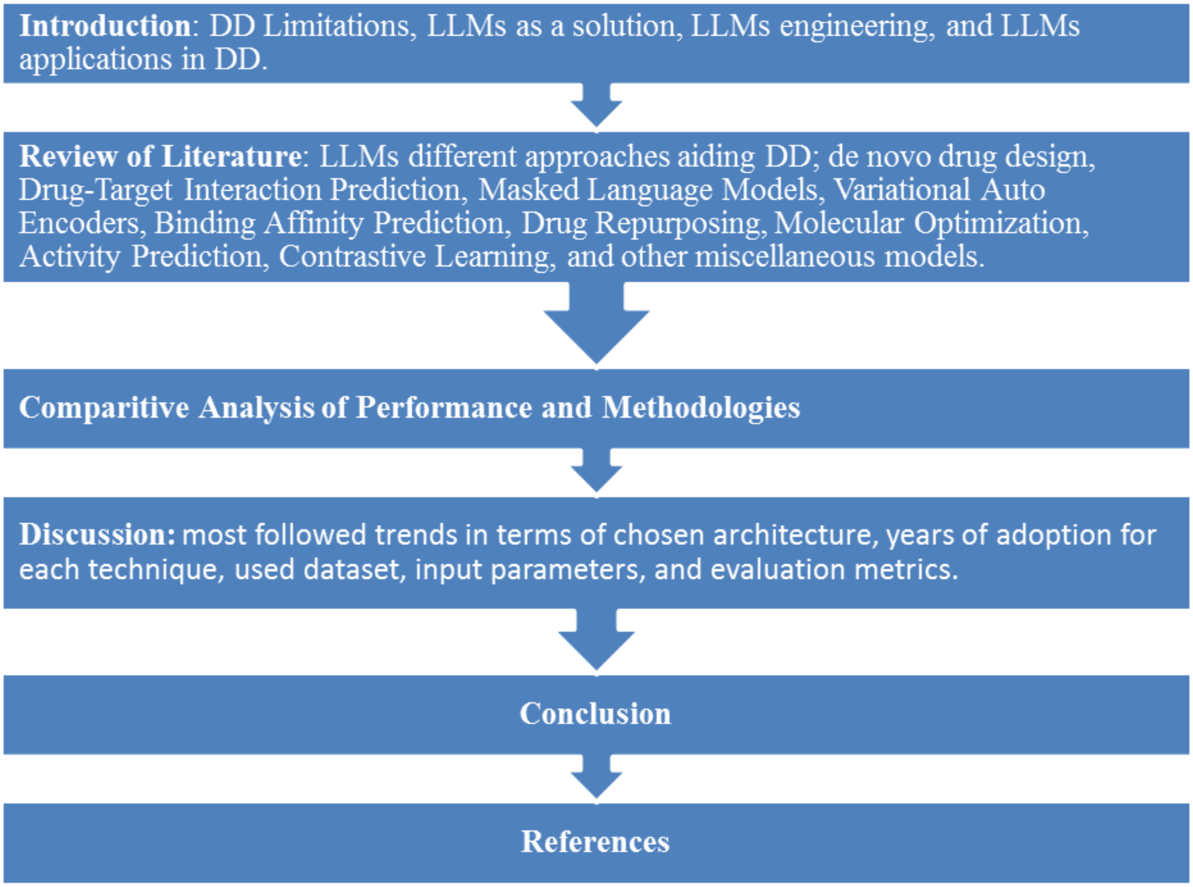

Introduction: DD Limitations, LLMs as a solution, LLMs engineering, and LLMs applications in DD.
Review of Literature: LLMs different approaches aiding DD; de novo drug design, Drug-Target Interaction Prediction, Masked Language Models, Variational Auto Encoders, Binding Affinity Prediction, Drug Repurposing, Molecular Optimization, Activity Prediction, Contrastive Learning, and other miscellaneous models.
Comparitive Analysis of Performance and Methodologies
Discussion: most followed trends in terms of chosen architecture, years of adoption for each technique, used dataset, input parameters, and evaluation metrics.
Conclusion
References

Figure 1: The survey's structure

Review of Literature

One of the first to address the possibilities LLMs can provide in the DD process was [15], in 2021, there was not a lot of focus on this topic at that time. His aim was to emphasize the applications of LLMs that can accelerate the discovery of a treatment for Corona Virus Disease 2019 (COVID-19). In his paper he focused on LLMs role in accelerating target identification, reshaping clinical trials, assisting the regulatory decision-making, and advancing post-marketing surveillance (pharmacovigilance).

LLMs for de novo Drug Design

LLMs are becoming of great advantage to the process of *de novo* drug design (in-silico design of drugs, usually proteins, from scratch). A transformer model proposed by [16] depended on protein sequences to predict binding affinities between generated molecules and their biological targets; avoiding all the obstacles that arise in structure-based *de novo* drug design that requires prior knowledge of the three-dimensional structure of proteins. This transformer model was composed of an encoder-decoder architecture, taking the protein sequence as an input to generate the corresponding Simplified Molecular-Input Line-Entry System (SMILES) string. The model is based only on self-attention mechanisms with no Convolutional Neural Networks (CNN) or Recurrent Neural Networks (RNN), with that it works with long-range inputs and is much faster than RNN-based models. Most of the generated SMILES strings were valid and unique, with more than 17% novel compound matching the ZINC15 database. Two proteins were selected; Insulin-like growth factor 1 receptor (IGF-1R) and Vascular endothelial growth factor receptor 2 (VEGFR2), both are from the receptor tyrosine kinases family and contribute to major diseases, such as cancer, arthritis, and diabetes. Most of the generated molecules had acceptable drug-like molecule boundaries with proper reproducibility of the property distribution of molecules as in the training set. However, this must be checked in experiment. 51% of generated molecules had a Tanimoto score lower than 0.5; suggesting major differences than molecules in the training dataset, this leads to a high variation of their functionalities.

TamGent (Target-aware molecule generator with Transformer) by [17] aimed to advance the structure-based drug design. This transformer model takes both the 3D structure and the protein sequence into consideration and is trained on 10 million compounds from PubChem; this provided the model with a huge generative power. Moreover, a variational auto-encoder (VAE) was used to handle the possibility of multiple compounds binding to the same target. TamGent outperformed benchmarks in both binding-affinity and drug-likeness properties, generated drugs that match drugs in the DrugBank, and generated novel compounds with better docking scores than other references. Two cases were studied; SARS-CoV-2 main protease (M pro) and the oncogenic mutant KRAS G12C. The generated compounds had both a good previously known inhibitors of M pro and potential novel compounds that bind to the enzymatic sites. Stronger interactions with G12C mutant were noted, in KRAS G12C, compared to wild type KRAS which means that the compounds are more specific on the mutated protein compared with normal cells.

One of the unique models that addressed the synthesizability of generated molecules is DeepLigBuilder+ [9], its framework addresses both 3D structure information and reaction-based pathways, supporting the retro-synthetic analysis. The model is a combination of a reinforcement learning method based on Monte Carlo tree search (MCTS), as a synthesizability constraint depending on purchasable building blocks, and an SE (3)-equivariant transformer conditioned on the shape and pharmacophore-based inputs. Also, a masking strategy was applied on each step; to guarantee the generation of synthesizable synthons. Two case studies were evaluated; inhibitor compounds targeting ATP-binding pocket of Bruton's tyrosine kinase (BTK) and the NAD+-binding pocket of human phosphoglycerate dehydrogenase (PHGDH), they have shown high predicted binding affinity and suitable binding modes within a proper synthesizability constraint. Furthermore, to reach the structure-based generation capability, they created a dataset of pharmacophore-ligand pairs using large-scale 3D alignment of molecules, and then use it to develop a novel SE (3)-equivariant transformer conditioned on 3D information, followed by MCTS as rolling out policy; resulting in a significantly increased search speed.

In a minireview by [7], he focused on how the advancements in NLP algorithms can be utilized in generating drugs with high-level properties; the better the semantic generation of bonded elements the better compounds we get, this technique is much more successful and easier than the previously utilized complex molecular graphs. In his study, many molecular representation entities were reviewed; the Simplified Molecular Input Line Entry Systems (SMILES) overcame Self-referencing embedded strings (SELFIES) and others in the ability of filtering out invalid molecules. However, each of the reviewed systems (SMILES, SELFIES, or DeepSMILES) can have a superiority depending on the application required, e.g., DeepSMILES predicts binding affinity.

ADA-T5, a novel model was proposed by [18]; to overcome the scarcity in available data, they depended on generating pseudo-data in a model teached with few-shot prompting, description on SMILES molecules was provided for the model to learn predicting correct molecules from the desired properties. A retrieval-based prompting strategy led to excellence in the model's performance and to a continuous increase in data size.

FSM-DDTR (End-to-end feedback strategy for multi-objective De Novo drug design using transformers) [19], a transformer model that uses SMILES as an input to generate molecules with proper physicochemical and pharmacological properties. The transformer's architecture is a basic encoder predictor and a decoder generator, along with a feedback loop to optimize generated compounds; top-scoring chemical compounds take place of least favorable inputs of the decoder. The model could generate valid SMILES strings with a great level of novelty, approximately 100% uniqueness. To estimate the unbiased decoder performance MOSES dataset was utilized, the decoder had superior performance in novelty rate (> 97%) and comparable internal diversity, uniqueness, and validity rate, with reduced chance of overfitting due to the novelty of generated compounds. A model trained on multiple features (pIC50, SAS, LogP, MW, and TPSA) achieved the overall best performance, with 99.36% of the generated chemical entities following Lipinski's rule of five; potentially orally available. Moreover, mostly they were valid and novel, and the average value of pIC50 toward the AA2AR receptor (6.81) outperformed the unbiased model result (5.81).

DGFN (Double Generative Flow Networks) [20], a concept introduced to advance the training stability and the exploration abilities of large state spaces of GFN (Generative Flow Networks). DGFN is developed with the influence of both reinforcement learning and generative models. The enhancement of the exploration capabilities serves as a great chance to improve the sampling of molecules for DD and small molecules generation purposes. DGFN was put into comparison with conventional GFN on two benchmark tasks: hypergrid and molecule generation; DGFN finds modes across the hypergrid faster than conventional GFN. The DD task was about generating small molecules, molecules with low binding energy to the soluble epoxide hydrolase (sEH) protein, three models were trained $\mathrm{DGFN}_{\mathrm{TB}}$ along with two baseline models, $\mathrm{GFN}_{\mathrm{TB}}$ and $\mathrm{GFN}_{\mathrm{SubTB.}}$ $\mathrm{GFN}_{\mathrm{TB}}$ exhibited more pronounced fluctuations in the training, and $\mathrm{DGFN}_{\mathrm{TB}}$ had lower variance. $\mathrm{DGFN}_{\mathrm{TB}}$ had also surpassed $\mathrm{GFN}_{\mathrm{SubTB}}$ in the discovery of modes with rewards > 0.9.

Structured State-Space Sequence Models (S4), engineered by [21], with an architecture combining a Long Short-Term Memory (LSTM) and a transformer. The uniqueness of this architecture is that it combines the fast-generation capabilities of the LSTM with its element-by-element strategy along with the all-at-once molecular processing by the transformer, avoiding the extensive computational power transformers require. This dual model has shown success when tested and benchmarked in its ability to learn the bioactivity of compounds, chemical space exploration, the design of natural products, and was preferred when long sequences were present. S4's ability in generating molecules with suitable Molecular Dynamic (MD) properties and potencies was validated by testing the Mitogen-Activated Protein Kinase 1 inhibitors (MAPK1-inhibitors) it produced.

Recently, [22] has revealed ProT-Diff, a novel modularized deep generative model, for *de novo* Anti-Microbial Peptides (AMPs) generation, AMPs are the current hope for overcoming the anti-microbial resistance world-wide problem. ProT-Diff is a sandwich model, embedding a continuous diffusion model between the encoder and the decoder of the transformer-based protein language model (PLM) ProtT5-XL-UniRef5023. This technique maximized the efficacy of the pre-trained PLM and minimized the needed computational power. In one experiment they trained the diffusion model with a dataset of AMPs and non-AMPs, pre-trained it on a peptide dataset, UniProtKB, to learn a general grammar of protein sequences, and then fine-tuned the pre-trained model on the specific AMP dataset to learn specific details from AMPs; this full training took less than forty hours. The model was evaluated in-silico, in-vitro, and in-vivo. The model has not only produced replicated sequences but has shown uniqueness in the produced AMPs; the sequences differed in a range of 20-100 %. Moreover, one of the produced AMPs, AMP_2, shown effectiveness against various anti-microbial-resistant bacteria, having low Minimum Inhibitory Concentration (MIC) value, and a safe toxicity profile (tested for hemolytic toxicity and cytotoxicity).

Those mentioned models on *de novo* Drug Design are listed in Table 1.

Table 1: Comparison between LLMs for *de novo* Drug Design

Model	**Architecture**	**Main advantage**	**Molecular entry system**	**Data source**	**Case studies**
[16]	Self-attention transformer	Works with long-range inputs and is much faster than RNN-based models.	SMILES	BindingDB, and datasets with proteins from human, bovine, rat, and mouse.	Most of the generated molecules targeting IGF-1R and VEGFR2 had acceptable drug-like molecule boundaries with proper reproducibility of the property distribution of molecules as the training set.

TamGent [17]	Transformer	Trained on a very large dataset which gave it huge computational power.	SMILES	PubChem	SARS-CoV-2 M pro and KRAS G12C inhibitors.
DeepLigBuilder + [9]	MCTS and SE (3)-equivariant transformer	Guaranteed synthesizability of compounds.	-------------	PDBBind, and synthon dataset.	PHGDH and BTK inhibitors.
ADA-T5 [18]	Transformer	Pseudo-data generation, continuous increase in data size.	SMILES	PubChem, DrugBank, ChEBI-20, PCdes, and pseudo-data (constructed PseudoMD-1M).	---------------------
FSM-DDTR [19]	Transformer and a feedback loop.	High novelty.	SMILES	ChEMBL, and MOSES.	A model trained on (pIC50, SAS, LogP, MW, and TPSA) achieved the overall best performance.
DGFN [20]	Same as the original GFN models.	The enhancement of exploration capabilities.	-------------	Dataset and proxy model provided by [23].	Two benchmarking tasks, on hypergrid and molecule generation.
S4 [21]	LSTM and a transformer.	Fast-generation capabilities of the LSTM with its element-by-element strategy along with the all-at-once molecular processing by the transformer, avoiding the extensive computational power transformers require.	SMILES	ChEMBL, LIT-PCBA, and COCONUT.	MAPK1-inhibitors with suitable MD properties and potencies.
ProT-Diff [22]	Transformer, with a	Maximized pre-trained PLM efficacy and	-------------	CAMPR4, ADAM, APD3,	AMP_2, effective against various anti-microbial-

	continuous diffusion model.	minimized needed computational power.		GRAMPA, and UniProtKB.	resistant bacteria, having low MIC value, and a safe toxicity profile (tested for hemolytic toxicity and cytotoxicity).

Drug-Target Interaction Prediction Models

Several LLMs were created with the task of Drug-Target Interaction (DTI) prediction, proposed models with their different approaches are listed in Table 2:

Table 2: LLMs with Drug-Target Interaction Prediction

Model	**Approach / Architecture**	**Main advantage**	**Molecular entry system**	**Data source**
MolTrans [24]	Augmented transformer with embedding module.	Better extraction of semantic relations from unlabeled data.	SMILES	UniProt, and ChEMBL.
IGT [25]	3-Way graph transformer (has a receptor graph, ligand graph and a complex graph in each network of the model).	Improved fitting and better generalizability.	------------------	DUD-E, LIT-PCBA, and PDBBind.
DACPGTN [26]	Graph transformer	Novel interactions predictions from integrated biomedical data.	------------------	Anatomical Therapeutic Chemical (ATC) benchmark [27], KEGG, and Drugbank.
DeepMGT-DTI [28]	Molecule Attention Transformer (MAT)	Integrated structural and sequential information; outperforming single module-models.	Targets sequences, and drugs' SMILES sequences.	Drugbank, KEGG, and PubChem.

MHTAN-DTI [29]	Metapath instance-level transformer, with single and multi-semantic attention layers.	Weakened noise influence, interpretable results, and better generalizability.	------------------	Dataset on interactions by [30], DrugBank, HPRD, and Comparative Toxicogenomics.
DrugormerDTI; molecule graph and Residual2vec [31]	Graph transformer	High-feature extraction and expression capabilities.	Encoded protein sequences, and drug graphs representing the drugs sequences.	C. elegans, Human, Davis, and GPCR (GPCR is particularly important for interactions learning).
Helix encoder [32]	Proteins sequences encoder	Specific for G protein-coupled receptors' (GPCRs) largest class, class A.	Sequences of transmembrane regions of class A.	GPCR dataset constructed from Compound-protein Interactions (CPI) in GLASS database.
MCL-DTI [33]	Transformer; encoder, decoder, feature fusion module, and a classifier.	Increasing drug representation and increasing multimodal features learning extensively; improving DTIs.	Multimodal drug features, and FASTA sequences.	Davis, C. elegans, Human, and Biosnap (for Drug-Drug Interactions (DDIs)).
FOTF-CPI [34]	Transformer	Fragmented-compounds understanding, and feature-fusion method improved affinity prediction and interpretability.	SMILES	BindingDB, Davis, Biosnap, and DUD-E.
DLM-DTI [35]	Dual-encoder transformer (the target encoder is a teacher-student model).	Integrated general-knowledge and target-knowledge; enhancing learning and predicting capabilities.	SMILES	Davis, BindingDB, and Biosnap.

ULDNA [8]	LSTM-attention network, and three unsupervised language models embeddings.	Higher DNA-Binding sites prediction accuracy than sequence-only models (Binding sites annotations were added).	Amino acids sequences.	PDNA-543, PDNA-41, PDNA-335, PDNA-52 and PDNA-316.
iNGNN-DTI [36]	interpretable Nested Graph Neural Network (iNGNN), with an attention-free transformer.	Shown consistent improvement and outperformed all baseline models.	Protein graphs; generated from SMILES using AlphaFold2.	KIBA, Davis, and Biosnap.

Masked Language Models in DD

In the case of DeepLigBuilder+ [9], stepwise masking with chemical constraints was applied to control the generated molecules to stay within a specific space of synthesizable and purchasable synthons. In this masking strategy, the action masks were based on what atoms and bonds are not favorable and would result in a molecule with unpurchasable building blocks.

Another Masked Language Model (MLM) by [37] was trained to serve as a random mutation operator, aiming to generate new optimized molecules. Two generation strategies were applied and compared; fixed and adaptive, the fixed strategy refers to the fixed pre-trained set of molecules while the adaptive strategy updates and learns from the newly generated mutated molecules in each iteration. Starting with a simpler task, these strategies were tested on molecular generation; the adaptive strategy produced mutated molecules with closer features of synthesizability and drug-likeness to the first generation while the fixed strategy was biased and produced mutations prevalent in the pre-trained set. However, there were not any significant differences in the generation time both strategies needed. In the case of molecule optimization, two opposite scenarios are possible; depending on the quality of the initial data. When had poor data, the adaptive strategy failed and generated less valid and less acceptable molecules due to continuously learning from molecules with poor scores. On the other hand, when we had high-score data, the adaptive strategy outperformed the fixed strategy and generated more valid and acceptable molecules in all aspects. After testing, the optimal strategy was to start with five fixed-generations followed by twenty adaptive-generations.

Variational Auto-Encoders

The model represented by [38], is an integrated Variational Auto-Encoder (VAE) with a Convolutional Neural Network (CNN) followed by an attention mechanism. The VAE aids in learning the Drug-Protein Interactions (DPIs) through probabilistic evaluations that also reduce redundancies, while the CNN is to extract local features of drugs and protein, and the attention mechanism figures out the key features that relates to DPI sites. Case studies have proved the ability of this model to generate proper results compared to base models.

SGVAE [39], a modified Grammar VAE (GVAE) by adding properties-information to the input data, resulting with a supervised environment. SGVAE covers two important small-molecules generation applications; properties prediction and custom novel molecules generation. Moreover, SGVAE has the ability to measure the properties' values of given molecules. SGVAE outperformed models that use SMILES for properties predicting tasks.

Protein Multimodal Network (PMN) model a novel model proposed by [10], with the superiority of being able to augment multiple protein-related informations (Multimodal), these informations include; primary structure sequences, and 3D structure residue-level graph and geometry. One of the successful cases, TargetVAE, a model that generates ligands that bind to specific proteins, with customized properties such as binding affinity and high synthesizability.

LLMs for Binding Affinity Prediction

DTITR [40], a concatenated-encoders approach, were two transformer-encoders are working in parallel to generate Drug-Target binding Affinity (DTA) predictions that are more reliable than simple interactions predictions; binding strength is evaluated based on structural and sequential information. One encoder takes protein sequences as an input, while the other is for drugs' SMILES sequences. Results from both encoders are concatenated through a cross-attention block and further go through a Fully Connected Feed Forward Network (FCNN). DTITR has outperformed or worked equally with tested benchmarks.

GSATDTA [41], a triple channel model that takes sequential and structural information to generate DTA predictions; a Graph Neural Network (GNN) learns drugs' topological features followed by a graph-sequence attention layer to catch important structural and sequential features, while the protein target is studied by a separate transformer, and at the end both outcomes are concatenated and go through a large number of NNs ending with a regression layer that gives a predicted DTA value. GSATDTA outperformed tested benchmarks on two different datasets.

TEFDTA [42], a combined encoder and fingerprint transformation model; aimed to study the covalent-bonds interactions rather than only focusing on non-covalent bonds. At first, FASTA sequences of the proteins are label encoded, go through an embedding layer, and then captured data go through several 1D-CNNs. While drugs' SMILES sequences are converted to MACCS fingerprints, to an embedding layer with position encoding employed, and finally features are extracted by the encoder. At the end, concatenated predictions are fed into an FCNN that generates predicted DTA values. TEFDTA was able to sensitively predict DTAs with minor structural variations.

LLMs in Drug Repurposing

RHGT [43], a novel model for Drug Repurposing; the process of utilizing available drugs for newer indications and diseases. The model is a Relation-aware Heterogeneous Graph Transformer, it learns relations between drugs, diseases, and genes; resulting with useful drug-disease associations. Although many Graph Neural Networks (GNNs) have been developed for Drug Repurposing tasks, RHGT model still outperformed them. The networking of the model can be divided into three consecutive embedding modules; subtype-level network, node-level network, and an edge-level network. The output of the first level is the input of the other two, and the output of the second level is the input the last one; level-by-level learning. Each level serves as an irreplaceable factor to the high-performance RHGT demonstrates.

A model for drug repurposing is proposed by [11], it was tested on Type 2 Diabetes Mellitus (T2DM). The model augments both structural information from StAR transformer model and semantic information of the name and description of drugs from HittER transformer model; this led to outperforming these single-embedding models. The model has successfully discovered five drugs that can be repurposed for T2DM; Triterpenes, Sho-saiko-to, LY294002, Clomiphene Citrate, and Mitogen-Activated Protein Kinase Inhibitors (MAPK Inhibitors).

WMAGT [44], an aggregated model of a graph convolutional network and graph transformer, it aims to find relation between diseases and drugs; discovering new indications for drugs and understanding drugs safety-profiles. The model was created by integrating three networks; drug–drug similarity, disease–disease similarity, and drug–disease association networks, followed by an end-to-end model to find unknown patterns and associations. WMAGT outperformed five state-of-the-art methods. To test the applicability of WMAGT, it was challenged to configure drugs for Parkinson’s disease; seven predicted drugs were relevant to Parkinson’s as shown in further literature analysis.

LLMs with Contrastive Learning for Drug-Target Interaction Prediction

ConPLex [45], a pre-trained Protein Language Model (PLex) with protein-anchored contrastive co-embeddings (Con). Based on genomic data (sequences) ConPLex generates predictions on Drug-Target Interactions (DTIs); it finds relations from the distances of learned entities. Therefore, ConPLex is able to work on huge genomic and compounds libraries. ConPLex embeddings are interpretable, with that, human cell-surface proteins functions’ can be characterized, enabling high-throughput in-silico screening, and it was able to detect compounds with sub-nanomolar activity. One of the main advantages of ConPLex, is that it overcame the DTIs decoy problem; this is owed to its co-embedding architecture. Moreover, with a simple modification on the last activation function, ConPLex can serve as a binding affinity predictor model.

CLAPE [14], a generalizable model developed for performing DNA-Binding sites predictions (CLAPE-DB) based on contrastive Learning and a pre-trained encoder. CLAPE was generalized as a binding site predictor model of DNA-binding sites (CLAPE-DB), protein-RNA binding sites (CLAPE-RB), and antibody-antigen binding sites (CLAPE-AB). In the case of DNA-binding sites, CLAPE-DB outperformed the second-best benchmark model. Moreover, it interestingly outperformed a structure-based model without being trained on any structural information. In the case of protein-RNA binding sites, CLAPE-RB outperformed pre-existing sequence-based models, and showed potential in protein structure prediction. CLAPE-AB also achieved relatively high Area Under the Curve (AUC) results, 0.92, accurately predicting antibody paratopes from sequence inputs.

LLMs for Activity Prediction

SYN-FUSION [46], a unique model that combines features from both Graph Neural Networks (GNNs) and transformers, in order to comprehend the global structure of the molecule along with individual characteristics of atoms. SYN-FUSION with its combined model architecture has outperformed the performance of its individual modules separated.

The following model [47], proposed a hybrid approach that leverages the importance of LLMs in protein sequence-analysis along with the 3D structure-information incorporation by the added numerical embeddings in Euclidean space for proteins; contact maps were generated, outperforming each of these single modules. The workflow ends with a concatenated features-result by the favor of these embeddings.

The next model, R-MAT [48], a relative molecule self-attention transformer with 3D structure representations and minimal inductive biases set, was designed in aim of overcoming the humble pre-existing pre-training methods; to enhance the prediction outcomes despite the small available datasets. The self-attention techniques employed led to the noticed improvement in predicting molecular properties.

LLMs for Molecular Optimization Tasks

GraphGPT [49], a conditioned-molecular generation model which depends on the scaffold information, incorporating both topological characteristics through graph structure information and enhanced molecular generation through a Generative Pre-trained Transformer (GPT) sequence-to-sequence method. GraphGPT proved its power in high-throughput screening for the aim of scaffold-based molecular generation.

DrugAssist [12], an interactive LLM designed for molecular optimization tasks. It was benchmarked with two models; a seq2seq model with attention architecture, and a transformer model. Properties tested were Blood-Brain-Barrier Permeability (BBBP) and solubility; their success rate, validity, and average similarity with the pre-optimized molecules. When DrugAssist was compared against traditional approaches; it scored the highest success rates in single-property and multi-property optimization and maintained high validity and similarity to the pre-optimized molecules. Benchmarks performed poorly; faced difficulty in comprehending the required work; the results offered a guide for users to websites that talk about molecule optimization instead of optimizing molecules. DrugAssist is of good transferability under zero-shot learning. DrugAssist was able to increase BBBP and QED by at least 0.1 simultaneously, resulting with a structurally similar molecule to the pre-optimized one. This means that DrugAssist is able to freely combine and learn individual properties, then to optimize them simultaneously. Moreover, when the model provides a molecule that does not satisfy the requirements, it can correct the error and compensate based on human-provided example, opening a potential to aid researchers in continuous optimization of molecules.

SGPT-RL [50], a model composed of a generative pre-trained transformer (a decoder-only version) augmented with the Reinforcement Learning (RL) strategy. SGPT-RL aimed to optimize binding affinities of molecules with their targets, the optimization was focused on two prerequisites; molecular docking and Quantitative Structure Activity Relationship (QSAR). The model has been evaluated on three measures; Moses distribution learning, and affinity towards Dopamine Receptor D2 (DRD2), and Angiotensin-Converting Enzyme 2 (ACE2). SGPT-RL generated notably valid and novel molecules in all tasks and outperformed the benchmark in molecular docking results. Moreover, SPGT-RL has learned conserved scaffold patterns on its own.

Another transformer model [51], with a primary task to generate molecules with optimized LogD and solubility. However, this time it is a multi-head self-attention encoder-decoder transformer with masking mechanisms. For the source and target molecules they were obtained from the Matched Molecular Pairs provided by ChEMBL, with SMILES representation. Moreover, two transformers were tested; a conditional transformer, and an unconditional transformer depending on source molecules only. At the end, both transformers generated target molecules with the properties desired; ten novel and valid molecules were generated, and the performance was comparable to other benchmarks. However, the conditional model generated a higher number of successful molecules compared to the unconditional transformer.

TSMMG [52], a novel teacher-student model that takes advantage of previously available models, by using their molecule-text/properties information as a teaching input to the student model, building its ability to generate custom novel molecules based on the entered prompt. TSMMG is of a quite simple transformer-based decoder architecture. Moreover, in zero-shot testing it led to generating molecules with novel augmentation of properties. This model overcame many important problems, such as data scarcity and low quality.

Miscellaneous LLMs for various DD Tasks

TransDTI [53], a model designed to advance the process of DTI prediction in a way that discriminates different types of interactions, to come up with novel interactions predictions; this was possible due to being trained on large DTI datasets. Three different categories were assigned to predicted drug-target pairs; inactive, intermediate, and active. TransDTI significantly outperformed traditional DTI benchmark models.

SELFormer [54], a model that outperformed all other benchmarks with SELFIES input, including graph transformers and SMILES-dependent models, it is capable of handling several prediction tasks; aqueous solubility, side effects, and discriminating structural differences. A model with smaller training size was generated from the default SELFormer, and named SELFormer-Lite.

In this study [55], meta-learning was studied and evaluated on the ability to enhance the performance and outcomes' quality of different models, specifically on potent compounds generation task. Particularly, meta-learning served as a compensation for the lack in comprehensive data; resulting with higher accuracy in Known-Target Compounds (KTCs) predictions compared to other benchmarks.

PrefixProt [56], a method that was designed to overcome the scarcity of data, such as the limitation within 20 amino acid sequences; hindering the ability of creating flexible control tags. PrefixProt's concept is to employ prefix-tuning on each property resulting with tokens that are further used in prompting the Protein Language Model (PLM). Results demonstrated the ability of PrefixProt to flexibly and controllably provide higher-quality molecular design suggestions.

FragAdd [57], a virtual screening model of target-binding small molecules and properties prediction, with a strategy of adding random fragments to entered molecules; improving the model's ability of learning high-quality features. Moreover, FragAdd works in a way that the original molecule is preserved unchanged, allowing for the augmentation with other learning protocols, such as masking. FragAdd demonstrated higher performance and accuracy than the average of other benchmarks.

Comparative Analysis of Performance and Methodologies

In this section, we are going to draw a comparison between the latest LLMs with similar tasks. The comparison will cover different aspects of models' performance and followed methodologies, such as architecture, used training datasets and their size, augmented parameters, performance evaluation metrics, and the availability of source code and data.

de novo Drug Design Models

A comparative analysis of performance and methodologies of *de novo d*rug design models is listed in Table 3.

Table 3: Comparative Analysis of Performance and Methodologies of *de novo* Drug Design Models

***de novo d*rug design model**	**Architecture**	**Used training datasets / Size**	**Performance evaluation metrics**	**Source code and data availability**
[16]	Self-attention transformer	BindingDB (238,147 records; 1,613 unique amino acid, and 154,924 unique SMILES).	ROC, AUC, Quantitative Estimate of Drug-likeness (QED), and Synthetic Accessibility (SA).	Available
TamGent [17]	Transformer	PubChem (random 10 million molecules).	Molecular Diversity (MD), QED, and SA.	Available
DeepLigBuilder+ [9]	MCTS and SE (3)-equivariant transformer.	PDBBind (12,456 pharmacophores and shapes), and constructed synthon	Maximum Mean Discrepancy (MMD), and	Not available

		DB (241,310 synthons from the global stock, 103,385 from the EU stock, and 783,195 synthons from the comprehensive catalog).	Root Mean Square Deviation (RMSD).	
ADA-T5 [18]	Transformer	PubChem, DrugBank, ChEBI-20, PCdes, and pseudo-data (constructed PseudoMD-1M).	Accuracy, validity, and Fingerprint Tani-moto Similarity (FTS).	Not available
FSM-DDTR [19]	Transformer and a feedback loop.	ChEMBL (1,046,964 compounds).	Accuracy, MSE, R^2, concordance correlation coefficient (CCC), Quantitative Estimate of Drug-likeness (QED), and percentage of molecules that strictly follow Lipinski's rule of five.	Available
DGFN [20]	Same as the original GFN models.	Dataset and proxy model provided by [23].	Mean and standard error.	Not available
S4 [21]	LSTM and a transformer.	ChEMBL, LIT-PCBA, and COCONUT.	Validity, uniqueness, and novelty.	Available
ProT-Diff [22]	Transformer, with a continuous diffusion model.	AMP DB (17,456 AMP-sequence): CAMPR4, ADAM, APD3, and GRAMPA. UniProtKB (567,834 AMP-sequence).	AUROC, and R^2.	Upon request

		Non-AMP (58,775 sequence)		

Drug-Target Interaction Prediction Models

A comparative analysis of performance and methodologies of Drug-Target Interaction (DTI) prediction models is listed in Table 4.

Table 4: Comparative Analysis of Performance and Methodologies of DTI Prediction Models

DTI model	Architecture	Used training datasets / Size	Performance evaluation metrics	Source code and data availability
MolTrans [24]	Augmented transformer with embedding module.	UniProt (560,823 unique proteins), and ChEMBL (1,870,461 SMILES).	AUROC, AUPRC, F1 score, sensitivity, and specificity.	Available
IGT [25]	3-Way graph transformer (has a receptor graph, ligand graph and a complex graph in each network of the model).	DUD-E (101 proteins), LIT-PCBA (9,780 active compounds, 407,839 compounds, and 15 targets), and PDBBind.	AUROC, LogAUC, AUPRC, Balanced accuracy, ROC enrichment, enrichment factor, and Matthews Correlation Coefficient (MCC).	Available
DACPGTN [26]	Graph transformer	Anatomical Therapeutic Chemical (ATC) benchmark [27], KEGG, and Drugbank. (1,749 diseases correlated between datasets, with data or target).	Aiming, coverage, accuracy, absolute true, and absolute false, proposed by [58].	Available
DeepMGT-DTI [28]	Molecule Attention Transformer (MAT)	[59] and Drugbank (12,496 molecular structures), KEGG (5,462 target sequences), and PubChem (21,158 DTIs).	AUC, AUPR, F1, sensitivity, accuracy, specificity, and precision.	Available

MHTAN-DTI [29]	Metapath instance-level transformer, with single and multi-semantic attention layers.	Dataset on interactions by [30], DrugBank, HPRD, and Comparative Toxicogenomics. (708 drugs, 1,512 proteins, 5,603 diseases, and 4,192 side effects).	AUROC, and AUPR.	Available
DrugormerDTI [31]	Graph transformer	C. elegans (2,504 proteins, 1,434 compounds, 6,728 DTIs), Human (852 proteins, 1,052 compounds, and 7,786 DTIs), Davis (379 proteins, 68 compounds, and 25,772 DTIs), and GPCR (356 proteins, and 5,359 compounds).	AUC, and AUPR.	Available
Helix encoder [32]	Proteins sequences encoder	GPCR dataset constructed from Compound-protein Interactions (CPI) in GLASS database (743,031 DTIs of 707 proteins and 316,814 compounds).	AUC, and ROC.	Available
MCL-DTI [33]	Transformer; encoder, decoder, feature fusion module, and a classifier.	Davis (64 drugs, and 379 targets), C. elegans, Human, and Biosnap (9,648 drugs, and 81,194 samples).	AUROC, and AUPRC.	Available
FOTF-CPI [34]	Transformer	BindingDB (10,665 drugs, 1,413 proteins), Davis (68 drugs, 379 proteins), Biosnap (4,510 drugs, 2,181 proteins), and DUD-E (22,886 drugs, 102 proteins).	AUC, PRC, sensitivity, specificity, F1, and cost.	Available
DLM-DTI [35]	Dual-encoder transformer (the target encoder is a teacher-student model).	Davis (68 drugs, 379 proteins, and 11,103 DTIs), BindingDB (10,665 drugs, 1,413 proteins, and 32,601 DTIs), and Biosnap (4,510 drugs, 2,181	AUROC, AUPRC, sensitivity, and specificity.	Available

		proteins, and 27,482 DTIs).		
ULDNA [8]	LSTM-attention network, and three unsupervised language models embeddings.	PDNA-543 (9549 binding, 134,995 non-binding), PDNA-41 (734 binding, 14,021 non-binding), PDNA-335 (6461 binding, 71,320 non-binding), PDNA-52 (973 binding, 16,225 non-binding) and PDNA-316 (5609 binding, 67,109 non-binding).	Sensitivity, specificity, accuracy, and MCC.	Available
iNGNN-DTI [36]	interpretable Nested Graph Neural Network (iNGNN), with an attention-free transformer.	KIBA (2,068 drugs, 229 proteins, and 118,254 DTIs), Davis (68 drugs, 442 proteins, and 30,056 DTIs), and Biosnap (4,510 drugs, 2,180 proteins, and 13,817 DTIs).	AUROC, AUPRC, sensitivity, and specificity.	Available

Masked Language Models

A comparative analysis of performance and methodologies of Masked Language Models (MLMs) is listed in Table 5.

Table 5: Comparative Analysis of Performance and Methodologies of Masked Language Models

Masked Language Model	**Architecture**	**Used training datasets / Size**	**Performance evaluation metrics**	**Source code and data availability**
DeepLigBuilder+ [9]	SE (3)-Equivariant transformer	Pharmacophore-ligand pairs synthesized from PDBBind ligands.	Mean and SD for MW, LogP, and QED, Wasserstein distance, MMD, RMSD value, Smina docking scores, distribution of shape and pharmacophore similarity, and reward value for the best molecule found.	Not available

[37]	Based on the Bidirectional Encoder Representations from Transformers (BERT)	Enamine REAL database augmented with a previously trained language model / 3.6 * 10^{10} molecules.	Mutation rate, drug-likeness, synthesizability, and number of generations.	Available

Variational Auto-Encoders

A comparative analysis of performance and methodologies of Variational Auto-Encoder (VAE) Models with DD tasks is listed in Table 6.

Table 6: Comparative Analysis of Performance and Methodologies of VAEs with DD tasks

Model	**Architecture**	**Used training datasets / Size**	**Augmented parameters**	**Performance evaluation metrics**	**Source code and data availability**
[38]	VAE and CNN with attention mechanisms.	Davis dataset (68 drug molecules and 442 proteins), BindingDB dataset (39,747 +ve samples and 31218 -ve samples), C.elegans and Human dataset (1,767 drug molecules, 1,876 protein, and 7,786 +ve and -ve samples), and KIBA dataset (2,111 drug molecules and 229 proteins).	Drugs molecules features and proteins sequences features.	ACC: (Davis: 0.85↓, KIBA: 0.841↑), AUC (Davis: 0.705↑, KIBA: 0.813↑). BindingDB (AUC: 0.913 (suboptimal), precision: 0.888↑, recall: 0.822↓, and F1: 0.854 (suboptimal)). C.elegans (AUC: 0.925 (suboptimal), precision: 0.927	Available

				(suboptimal), recall: 0.897 (suboptimal), and F1: 0.912 (suboptimal)). Human (AUC: 0.914↓, precision: 0.934↓, recall: 0.862↓, and F1: 0.897↓).	
SGVAE [39]	Modified Grammar VAE with supervised environment.	Modified Quantum-chemistry QM9 (~130,127 molecules), QM7-X, and PubChemQC PM6 (random 100,000 molecules of approximately 50-SMILES character).	--------------	Re-construction ACC (60.93%↑), prior validity (12.29%↑), novelty (72.66%↑), and uniqueness (93.06%↓).	Available
PMN, TargetVAE [10]	Novel Protein Multimodal Network (PMN)	PBDBind v2020.	Primary structure sequences, and 3D structure residue-level graph and geometry.	RMSE (0.035↓), MAE (0.032↓), pearson (0.01↑), spearman (0.016↑), R^2 (0.022↑), and CI (0.006↑).	Available

LLMs for Binding Affinity Prediction

A comparative analysis of performance and methodologies of LLMs for Drug-Target Affinity (DTA) prediction is listed in Table 7.

Table 7: Comparative Analysis of Performance and Methodologies of LLMs for Drug-Target

Affinity (DTA) Prediction

Model	**Architecture**	**Used training datasets / Size**	**Augmented parameters**	**Performance evaluation metrics**	**Source code and data availability**
DTITR [40]	End-to-End transformer with cross attention layers.	Davis (31,824 interactions of 72 kinase inhibitors and 442 kinase proteins), corresponding protein sequences were obtained from UniProt (of length 264-1,400 residue), and SMILES (38-72 long) from PubChem.	Protein sequences and SMILES strings.	MSE, RMSE, CI, R^2, and spearman.	Available
GSATDTA[41]	Novel triple-channel model; graph-sequence attention and a transformer.	Davis (68 drugs, 442 targets, and 30,056 DTIs, Affinity values: 5-10.8), and KIBA (2,111 drugs, 229 targets, and 118,254 DTIs, Affinity values: up to 17.2).	Graph information and sequences.	CI, MSE, and R^2.	Not available
TEFDTA [42]	Combined transformer-encoder and morgan fingerprint representation.	Non-covalent interactions databases; BinidingDB (80,324 drugs, 5,561 proteins, and 1,254,402	MACCS fingerprint (converted from SMILES), and FASTA sequences.	MSE, CI, and R^2.	Available

		DTI data), KIBA (2,111 drugs, 229 proteins, and 118,254 DTA values), and Davis (68 drugs, 442 proteins, and 30,056 DTA values), and fine-tuned on CovalentInDB (4,511 covalent inhibitor; 68 of them are approved drugs, and 57 reactive warheads of 280 protein).			

LLMs for Drug Repurposing

A comparative analysis of performance and methodologies of LLMs for Drug Repurposing is listed in Table 8.

Table 8: Comparative Analysis of Performance and Methodologies of LLMs for Drug Repurposing

Model	**Architecture**	**Used training datasets / Size**	**Augmented parameters**	**Performance evaluation metrics**	**Source code and data availability**
RHGT [43]	Relation-aware Heterogeneous Graph Transformer	CTD and TTD.	Drugs-genes-diseases consecutive embedding modules.	TTD: AUROC (0.7342), F1 Score (0.7611), and precision (0.7543). CTD: AUROC (0.7809), F1 score (0.7754), and	Available

				precision (0.7957).	
[11]	Two augmented transformers; StAR transformer model and HittER transformer model.	PubMed (4.5 billion words), and PMC full articles (13.5 billion words).	Structural information and semantic information of the name and description of drugs.	Test set compared to individual modules performance: Mean Rank (MR) (4,502.62↓), Mean Reciprocal Rank (MRR) (27.19↑), and Hits@1,3,10, and 100 (18.18↑, 33.23↑,43.17 ↑, and 50.11↑).	Not available
WMAGT [44]	Augmented graph convolutional network and graph transformer.	F dataset (313 diseases form OMIM, and 553 drugs from DrugBank), C dataset (663 drugs from DrugBank, and 409 diseases from OMIM), LRSSL dataset (763 drugs for DrugBank, 681 diseases from MeSH, and 3,051 validated drugs-diseases associations).	Drug–drug similarity, disease–disease similarity, and drug–disease associations.	ACC, Area Under the Precision-Recall Curve (AUPR), Area Under the Receiver Operating Characteristic Curve (AUC), F1 score, precision, and recall.	Available

LLMs with Contrastive Learning for Drug-Target Interaction Prediction

A comparative analysis of performance and methodologies of LLMs with contrastive learning for Drug-Target Interaction (DTI) prediction is listed in Table 9.

Table 9: Comparative Analysis of Performance and Methodologies of LLMs with Contrastive Learning for DTI prediction

Model	**Architecture**	**Used training datasets / Size**	**Performance evaluation metrics**	**Source code and**

				data availability
ConPlex [45]	Pre-trained Language Model (PLM), morgan fingerprint, and co-embedding layers.	BindingDB, DUD-E, STRING (15,816 proteins 50-800 amino acids long), and ChEMBL (1,533,652 compounds).	Average and standard deviation of AUPR, DUD-E evaluation sets, Pearson Correlation Coefficient (PCC), and in-vitro evaluation.	Available
CLAPE [14]	PLM (ProtBERT), 4-layers 1D-CNN, and a contrastive learning function.	Dataset1 (646 proteins with 15,636 DNA-Binding sites and 298503 non-binding sites) by [60], and Dataset2 (573 proteins with 14,479 DNA-Binding residues and 145,404 non-binding residues) by [61] extracted from [62].	Specificity, recall, precision, F1 score, MCC, AUC, AUPR, and amino acids composition statistical analysis.	Available

Activity Prediction Models

A comparative analysis of performance and methodologies of LLMs with contrastive learning for activity prediction models is listed in Table 10.

Table 10: Comparative Analysis of Performance and Methodologies of Activity Prediction Models

Model	**Architecture**	**Used training datasets / Size**	**Augmented Parameters**	**Performance evaluation metrics**	**Source code and data availability**
SYN-FUSION [46]	GNNs, and a transformer.	BBBP (2,039 molecules), Tox21 (7,831), ClinTox (1,476), HIV (41,127), BACE (1,513), SIDER (1,427), and	Global structure of molecules, and atoms' characteristics.	AUROC, RMSE, and MAE.	Not available

		MUV (93,087).			
[47]	LLM with 3D structure numerical embeddings.	Structural T-Cell Receptor Database from ProteinDB (325 human files, and 155 mouse files, sequences are 109-1,074.38 long, on average 5,415).	Structural and sequential information.	Avg ACC, precision, recall, AUROC, weighted F1, macro F1, and training runtime.	Available
R-MAT [48]	Novel graph transformer, with 3D representation and self-attention mechanisms.	Large hyper-parameter budget: ESOL, and FreeSolv. Small hyper-parameter budget: BBBP and Estrogen-B.	----------------	Only changing the learning rate, as in [63].	Available

Molecular Optimization Models

A comparative analysis of performance and methodologies of molecular optimization models is listed in Table 11.

Table 11: Comparative Analysis of Performance and Methodologies of Molecular Optimization Models

Model	**Architecture**	**Used training datasets / Size**	**Augmented parameters**	**Performance evaluation metrics**	**Source code and data availability**
GraphGPT [49]	Generative Pre-trained Transformer (GPT)	GuacaMol (1.6 million molecules abstracted from ChEMBL 24), and MOSES (1.9 million lead-like compounds	Molecular properties, and SMILES.	Synthetic Accessibility Score (SAS), Quantitative Estimation of Drug-likeness (QED), lipophilicity (logP),	Available

		abstracted from ZINC DB).		Topological Polar Surface Area (TPSA), Standard Deviation (SD), and Mean Absolute Deviation (MAD).	
DrugAssist [12]	Transformer	Constructed MolOpt-Instructions dataset (utilizing 1,000,000 molecules from ZINC DB).	SMILES	Success rates (solubility, blood brain barrier permeability (BBBP, and optimizing simultaneously)), validity, and average similarity before and after optimization.	Available
SGPT-RL [50]	GPT (decoder- only transformer) with Reinforcement Learning (RL).	ProteinDB 1R4L, ZINC (1.9 million lead-like molecules), ExCAPEDB (8,036 unique anti-DRD2 molecules, and 56 unique ACE2-inhibitors).	SMILES	Similarity to Nearest Neighbour (SNN), validity, uniqueness, and novelty.	Available
[51]	Transformer	Matched Molecular Pairs (MMPs) from ChEMBL (12,365 random molecules).	Properties information. and SMILES.	RMSE, and LogD.	Not available
TSMMG [52]	Teacher-Student LLM	Teacher models were depended on.	SMILES	Validity, novelty, diversity, and uniqueness.	Available

Discussion

This section discusses the findings that have been noted through this survey and focuses on the most followed trends through the ongoing publications on LLMs in DD.

LLMs for *de novo* Drug Design

The first LLMs aiding the DD process where on *de novo* drug design tasks, depending on protein sequences as input to transformers. However, incorporating 3D structure information was noticed to enhance the quality of generated molecules.

In Figure 2, some insights of noted trends in LLMs for *de novo* drug design are presented:

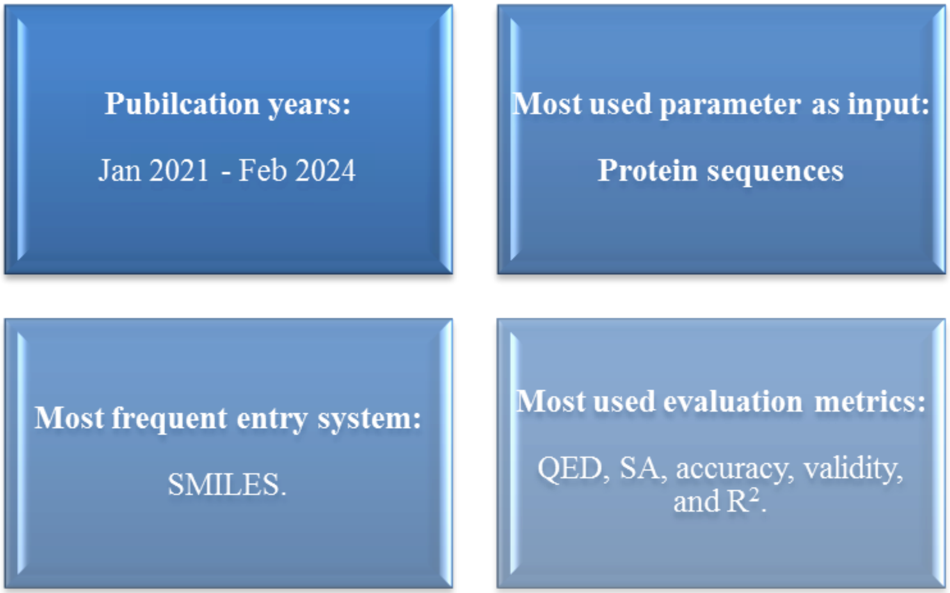

Pubilcation years:
Jan 2021 - Feb 2024
Most used parameter as input:
Protein sequences
Most frequent entry system:
SMILES.
Most used evaluation metrics:
QED, SA, accuracy, validity, and R^2.

Figure 2: LLMs for *de novo* DD trends

LLMs for Drug-Target Interaction Prediction

The most abundant publications of LLMs for DD were on DTI prediction. However, they varied in their adopted architecture; between encoder transformers, graph transformers, and LSTMs.

In Figure 3, most important details in the timeline of LLMs for DTI prediction are presented:

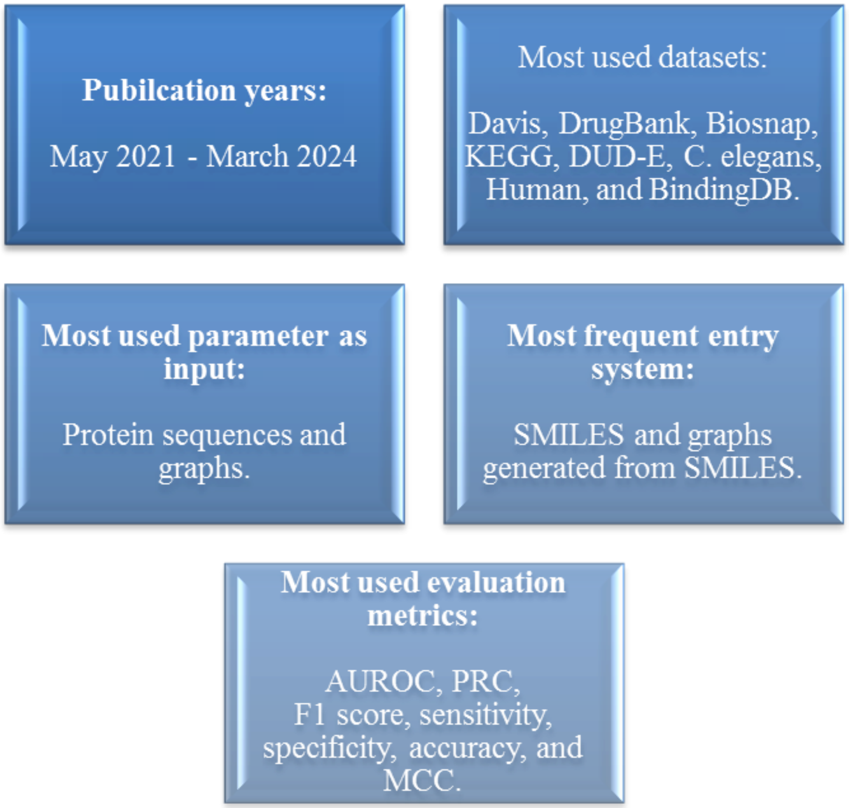

Pubilcation years:
May 2021 - March 2024
Most used datasets:
Davis, DrugBank, Biosnap, KEGG, DUD-E, C. elegans, Human, and BindingDB.
Most used parameter as input:
Protein sequences and graphs.
Most frequent entry system:
SMILES and graphs generated from SMILES.
Most used evaluation metrics:
AUROC, PRC, F1 score, sensitivity, specificity, accuracy, and MCC.

Figure 3: LLMs for DTI prediction trends

Masked Language Models in DD

There were two models that depended specifically on masking strategies; one performed stepwise masking, while the other was random. In both cases, they resulted with better drug-likeness and synthesizability scores.

Variational Auto-Encoders in DD

VAEs has been utilized in several models, were the best performance, producing proteins with customized features, and high-synthesizability was achieved when multiple information was provided; not depending solely on sequences.

In Figure 4, most important details in the timeline of VAEs in DD are presented:

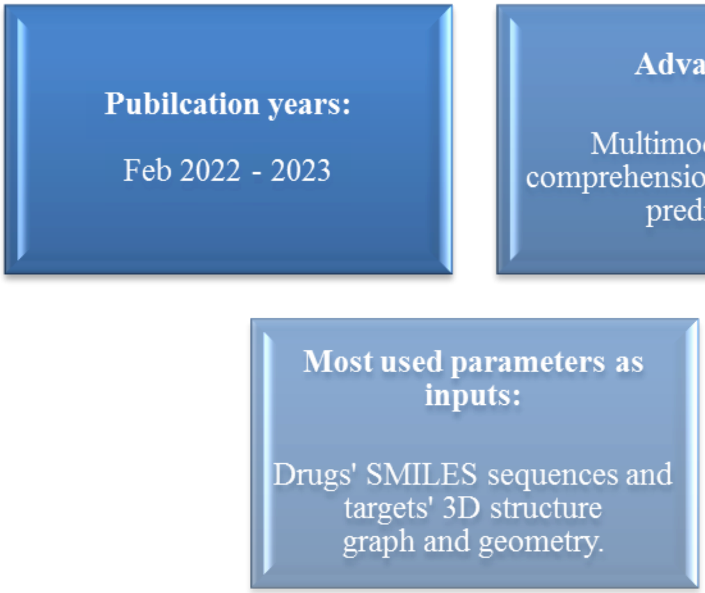

Pubilcation years:
Feb 2022 - 2023
Advantages:
Multimodal features comprehension; making better predictions.
Most used parameters as inputs:
Drugs' SMILES sequences and targets' 3D structure graph and geometry.

Figure 4: VAEs in DD trends

LLMs for Binding Affinity Prediction

LLMs with DTA prediction have a superior advantage over general DTI prediction models; they are able to differentiate types and strength of interactions, resulting with more profound predictions that are more likely to continue the process of DD successfully.

In Figure 5, details in the timeline of LLMs for DTA prediction are presented:

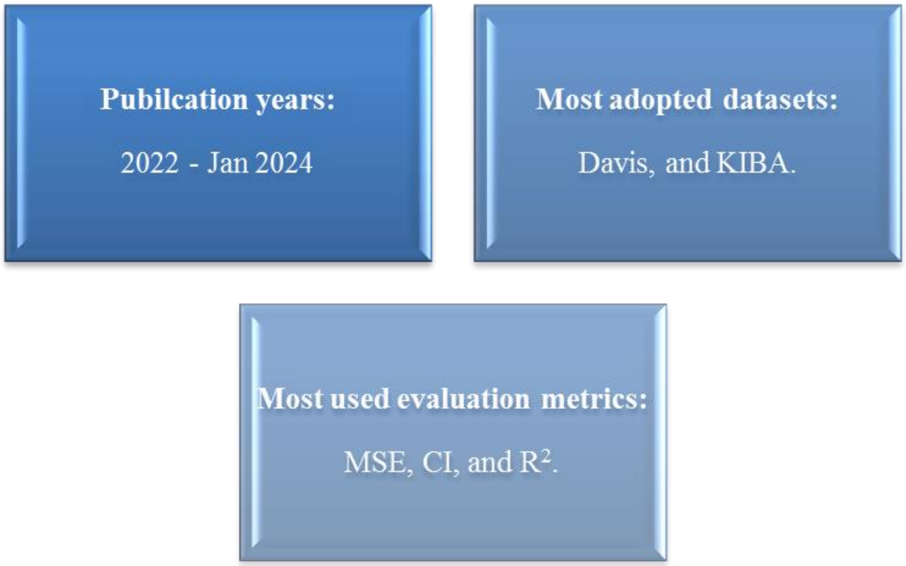


Figure 5: LLMs for DTA prediction trends

LLMs for Drug Repurposing

Drug repurposing is one of the most efficient approaches that accelerates finding cure while minimizing costs and efforts by utilizing pre-existing resources and finding new relations.

In Figure 6, noted details in the timeline of LLMs for drug repurposing are presented:

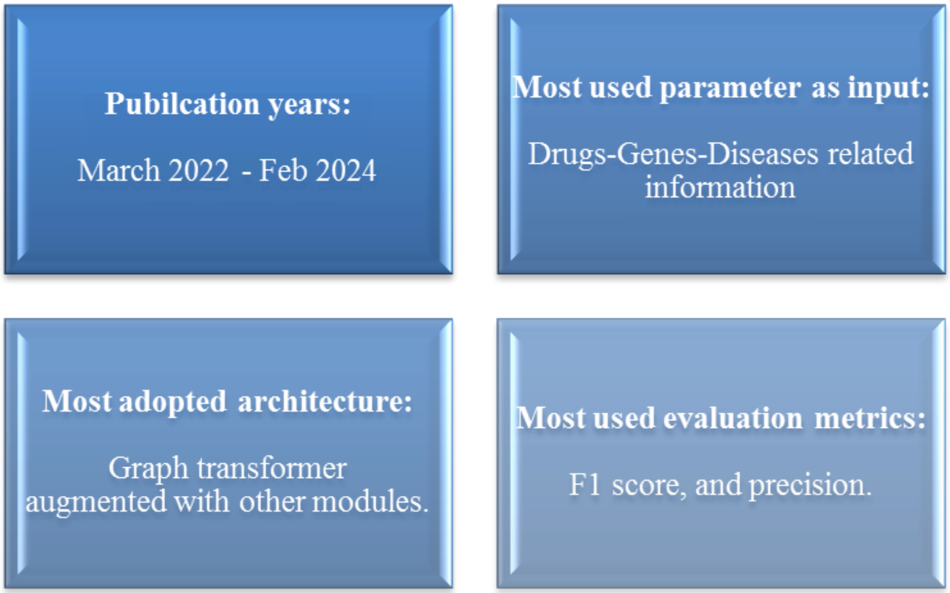
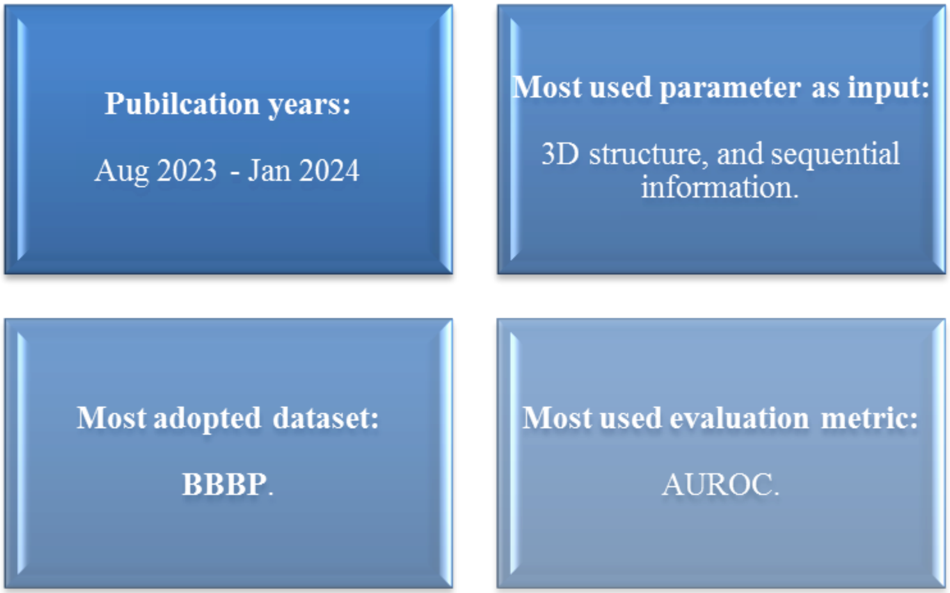

Pubilcation years:
March 2022 - Feb 2024
Most used parameter as input:
Drugs-Genes-Diseases related information
Most adopted architecture:
Graph transformer augmented with other modules.
Most used evaluation metrics:
F1 score, and precision.

Figure 6: LLMs for drug repurposing publications

LLMs for Activity Prediction

Models trained on augmented sequential and structural information yielded products of higher quality, even when datasets were of a smaller scale.

In Figure 7, noted trends in publications on LLMs for activity prediction are presented:



Pubilcation years:

Aug 2023 - Jan 2024

Most used parameter as input:

3D structure, and sequential information.

Most adopted dataset:

BBBP.

Most used evaluation metric:

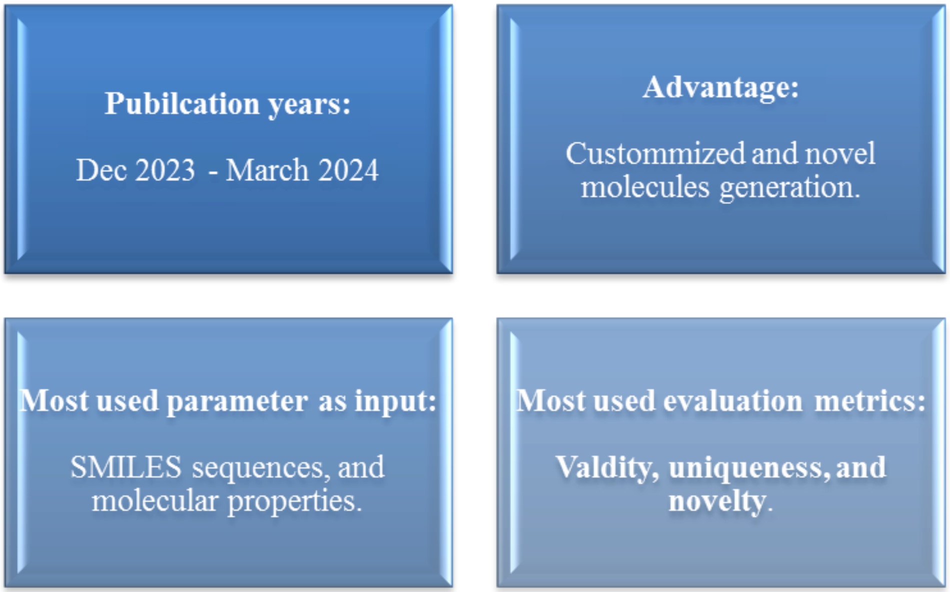
AUROC.

Figure 7: LLMs for activity prediction trends

LLMs for Molecular Optimization Tasks

Recently, a great milestone was achieved with the ability to generate customized and novel molecules with desired properties pre-defined.

In Figure 8, details of publications on LLMs for molecular optimization tasks are presented:



Pubilcation years:

Dec 2023 - March 2024

Advantage:

Custommized and novel molecules generation.

Most used parameter as input:

SMILES sequences, and molecular properties.

Most used evaluation metrics:

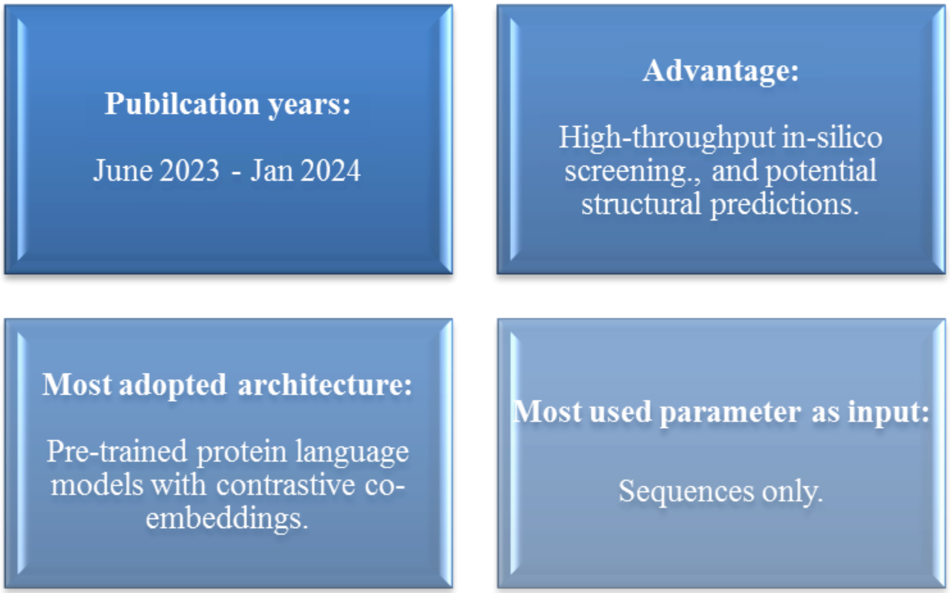
Valdity, uniqueness, and novelty.

Figure 8: LLMs for molecular optimization tasks trends

LLMs with Contrastive Learning for Drug-Target Interaction Prediction

The introduction of contrastive learning into DTI prediction models resulted with very efficient improvements while only depending on protein sequences for training the models.

In Figure 9, most important details in the timeline of LLMs with contrastive learning for DTI prediction are presented:



Pubilcation years:

June 2023 - Jan 2024

Advantage:

High-throughput in-silico screening., and potential structural predictions.

Most adopted architecture:

Pre-trained protein language models with contrastive co-embeddings.

Most used parameter as input:

Sequences only.

Figure 9: LLMs with contrastive learning for DTI prediction trends

Conclusion

This survey is one of the first in literature to present a comprehensive review on available Large Language Models (LLMs) in the Drug Discovery (DD) domain. It first differentiated the various tasks they serve and classified the models based on that. Then, it investigated their approaches, methodologies, performances, and the uniqueness of each model. Following that, there was a comparative analysis on their architecture, used training datasets and their size, augmented parameters, performance evaluation metrics, and the availability of source code and data. This survey is an evident that LLMs are of great potential to further advance the DD process and yield more efficient results. However, we must be aware that the scope is large and there are many areas of influence that can be contributed to in future work, with a superior advantage for models that succeed to integrate multiple parameters, resulting with more reliable predictions.

References

[1] E. T. Fokunang and C. N. Fokunang, "Overview of the Advancement in the Drug Discovery and Contribution in the Drug Development Process," *J Adv Med Pharm Sci*, pp. 10–32, Nov. 2022, doi: 10.9734/jamps/2022/v24i10580. [2] H. Naveed *et al.*, "A Comprehensive Overview of Large Language Models," Jul. 2023. [3] Y. Liu, J. Cao, C. Liu,

K. Ding, and L. Jin, “Datasets for Large Language Models: A Comprehensive Survey,” Feb. 2024. [4] R. Patil and V. Gudivada, “A Review of Current Trends, Techniques, and Challenges in Large Language Models (LLMs),” *Applied Sciences*, vol. 14, no. 5, p. 2074, Mar. 2024, doi: 10.3390/app14052074. [5]J. Liu, M. Yang, Y. Yu, H. Xu, K. Li, and X. Zhou, “Large language models in bioinformatics: applications and perspectives,” Jan. 2024. [6] G. Marvin, N. Hellen, D. Jjingo, and J. Nakatumba-Nabende, “Prompt Engineering in Large Language Models,” 2024, pp. 387–402. doi: 10.1007/978-981-99-7962-2_30. [7] F. Grisoni, “Chemical language models for de novo drug design: Challenges and opportunities,” *Curr Opin Struct Biol*, vol. 79, p. 102527, Apr. 2023, doi: 10.1016/j.sbi.2023.102527. [8] Y.-H. Zhu, Z. Liu, Y. Liu, Z. Ji, and D.-J. Yu, “ULDNA: integrating unsupervised multi-source language models with LSTM-attention network for high-accuracy protein–DNA binding site prediction,” *Brief Bioinform*, vol. 25, no. 2, Jan. 2024, doi: 10.1093/bib/bbae040. [9] Y. Li, J. Pei, and L. Lai, “Synthesis-driven design of 3D molecules for structure-based drug discovery using geometric transformers,” Dec. 2022. [10] N. K. Ngo and H.-Y. Son, “Multimodal Protein Representation Learning and Target-aware Variational Auto-encoders for Protein-binding Ligand Generation,” *Res Sq*, Nov. 2023. [11] S. Mam, D. Wichadakul, and P. Vateekul, “Drug Repurposing for Type 2 Diabetes Using Combined Textual and Structural Graph Representation Based on Transformer,” *IEEE Access*, vol. 11, pp. 65711–65724, 2023, doi: 10.1109/ACCESS.2023.3289863. [12] G. Ye *et al.*, “DrugAssist: A Large Language Model for Molecule Optimization,” Dec. 2023. [13] P. Seidl, A. Vall, S. Hochreiter, and G. Klambauer, “Enhancing Activity Prediction Models in Drug Discovery with the Ability to Understand Human Language,” Mar. 2023. [14] Y. Liu and B. Tian, “Protein–DNA binding sites prediction based on pre-trained protein language model and contrastive learning,” *Brief Bioinform*, vol. 25, no. 1, Jan. 2024, doi: 10.1093/bib/bbad488. [15] Z. Liu, R. A. Roberts, M. Lal-Nag, X. Chen, R. Huang, and W. Tong, “AI-based language models powering drug discovery and development,” *Drug Discov Today*, vol. 26, no. 11, pp. 2593–2607, Nov. 2021, doi: 10.1016/j.drudis.2021.06.009. [16] D. Grechishnikova, “Transformer neural network for protein-specific de novo drug generation as a machine translation problem,” *Sci Rep*, vol. 11, no. 1, p. 321, Jan. 2021, doi: 10.1038/s41598-020-79682-4. [17] K. Wu *et al.*, “Tailoring Molecules for Protein Pockets: a Transformer-based Generative Solution for Structured-based Drug Design,” Aug. 2022. [18] Y. Chen *et al.*, “From Artificially Real to Real: Leveraging Pseudo Data from Large Language Models for Low-Resource Molecule Discovery,” Sep. 2023. [19] N. R. C. Monteiro, T. O. Pereira, A. C. D. Machado, J. L. Oliveira, M. Abbasi, and J. P. Arrais, “FSM-DDTR: End-to-end feedback strategy for multi-objective De Novo drug design using transformers,” *Comput Biol Med*, vol. 164, p. 107285, Sep. 2023, doi: 10.1016/j.compbiomed.2023.107285. [20] E. Lau, N. Vemgal, D. Precup, and E. Bengio, “DGFN: Double Generative Flow Networks,” Oct.

2023. [21] R. Ozcelik, S. de Ruiter, E. Criscuolo, and F. Grisoni, “Chemical Language Modeling with Structured State Spaces,” 2024. [22] X.-F. Wang *et al.*, “ProT-Diff: A Modularized and Efficient Approach to De Novo Generation of Antimicrobial Peptide Sequences through Integration of Protein Language Model and Diffusion Model,” *Cold Spring Hrabor Laboratory*, Feb. 2024. [23] E. Bengio, M. Jain, M. Korablyov, D. Precup, and Y. Bengio, “Flow Network based Generative Models for Non-Iterative Diverse Candidate Generation,” Jun. 2021. [24]K. Huang, C. Xiao, L. M. Glass, and J. Sun, “MolTrans: Molecular Interaction Transformer for drug–target interaction prediction,” *Bioinformatics*, vol. 37, no. 6, pp. 830–836, May 2021, doi: 10.1093/bioinformatics/btaa880. [25] S. Liu *et al.*, “Improved drug–target interaction prediction with intermolecular graph transformer,” *Brief Bioinform*, vol. 23, no. 5, Sep. 2022, doi: 10.1093/bib/bbac162. [26] C. Yan, Z. Suo, J. Wang, G. Zhang, and H. Luo, “DACPGTN: Drug ATC Code Prediction Method Based on Graph Transformer Network for Drug Discovery,” *Front Pharmacol*, vol. 13, Jun. 2022, doi: 10.3389/fphar.2022.907676. [27] L. Chen, W.-M. Zeng, Y.-D. Cai, K.-Y. Feng, and K.-C. Chou, “Predicting Anatomical Therapeutic Chemical (ATC) Classification of Drugs by Integrating Chemical-Chemical Interactions and Similarities,” *PLoS One*, vol. 7, no. 4, p. e35254, Apr. 2012, doi: 10.1371/journal.pone.0035254. [28] P. Zhang, Z. Wei, C. Che, and B. Jin, “DeepMGT-DTI: Transformer network incorporating multilayer graph information for Drug–Target interaction prediction,” *Comput Biol Med*, vol. 142, p. 105214, Mar. 2022, doi: 10.1016/j.compbiomed.2022.105214. [29] R. Zhang, Z. Wang, X. Wang, Z. Meng, and W. Cui, “MHTAN-DTI: Metapath-based hierarchical transformer and attention network for drug–target interaction prediction,” *Brief Bioinform*, vol. 24, no. 2, Mar. 2023, doi: 10.1093/bib/bbad079. [30] Y. Luo *et al.*, “A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information,” *Nat Commun*, vol. 8, no. 1, p. 573, Sep. 2017, doi: 10.1038/s41467-017-00680-8. [31] J. Hu *et al.*, “DrugormerDTI: Drug Graphormer for drug–target interaction prediction,” *Comput Biol Med*, vol. 161, p. 106946, Jul. 2023, doi: 10.1016/j.compbiomed.2023.106946. [32] H. Yamane and T. Ishida, “Helix encoder: a compound-protein interaction prediction model specifically designed for class A GPCRs,” *Frontiers in Bioinformatics*, vol. 3, May 2023, doi: 10.3389/fbinf.2023.1193025. [33] Y. Qian, X. Li, J. Wu, and Q. Zhang, “MCL-DTI: using drug multimodal information and bi-directional cross-attention learning method for predicting drug–target interaction,” *BMC Bioinformatics*, vol. 24, no. 1, p. 323, Aug. 2023, doi: 10.1186/s12859-023-05447-1. [34]Z. Yin, Y. Chen, Y. Hao, S. Pandiyan, J. Shao, and L. Wang, “FOTF-CPI: A compound-protein interaction prediction transformer based on the fusion of optimal transport fragments,” *iScience*, vol. 27, no. 1, p. 108756, Jan. 2024, doi: 10.1016/j.isci.2023.108756. [35] J. Lee, D. W. Jun, I. Song, and Y. Kim, “DLM-DTI: a dual language model for the prediction of drug-target interaction with hint-based learning,” *J*

Cheminform, vol. 16, no. 1, p. 14, Feb. 2024, doi: 10.1186/s13321-024-00808-1. [36]Y. Sun, Y. Y. Li, C. K. Leung, and P. Hu, "iNGNN-DTI: prediction of drug–target interaction with interpretable nested graph neural network and pretrained molecule models," *Bioinformatics*, vol. 40, no. 3, Mar. 2024, doi: 10.1093/bioinformatics/btae135. [37] A. E. Blanchard *et al.*, "Adaptive language model training for molecular design," *J Cheminform*, vol. 15, no. 1, p. 59, Jun. 2023, doi: 10.1186/s13321-023-00719-7. [38]Y. Zhang, Y. Hu, H. Li, and X. Liu, "Drug-protein interaction prediction via variational autoencoders and attention mechanisms," *Front Genet*, vol. 13, Oct. 2022, doi: 10.3389/fgene.2022.1032779. [39] A. F. Oliveira, J. L. F. Da Silva, and M. G. Quiles, "Molecular Property Prediction and Molecular Design Using a Supervised Grammar Variational Autoencoder," *J Chem Inf Model*, vol. 62, no. 4, pp. 817–828, Feb. 2022, doi: 10.1021/acs.jcim.1c01573. [40] N. R. C. Monteiro, J. L. Oliveira, and J. P. Arrais, "DTITR: End-to-end drug–target binding affinity prediction with transformers," *Comput Biol Med*, vol. 147, p. 105772, Aug. 2022, doi: 10.1016/j.compbiomed.2022.105772. [41] X. Yan and Y. Liu, "Graph–sequence attention and transformer for predicting drug–target affinity," *RSC Adv*, vol. 12, no. 45, pp. 29525–29534, 2022, doi: 10.1039/D2RA05566J. [42] Z. Li, P. Ren, H. Yang, J. Zheng, and F. Bai, "TEFDTA: a transformer encoder and fingerprint representation combined prediction method for bonded and non-bonded drug–target affinities," *Bioinformatics*, vol. 40, no. 1, Jan. 2024, doi: 10.1093/bioinformatics/btad778. [43] X. Mei, X. Cai, L. Yang, and N. Wang, "Relation-aware Heterogeneous Graph Transformer based drug repurposing," *Expert Syst Appl*, vol. 190, p. 116165, Mar. 2022, doi: 10.1016/j.eswa.2021.116165. [44] S. He, L. Yun, and H. Yi, "Fusing graph transformer with multi-aggregate GCN for enhanced drug–disease associations prediction," *BMC Bioinformatics*, vol. 25, no. 1, p. 79, Feb. 2024, doi: 10.1186/s12859-024-05705-w. [45]R. Singh, S. Sledzieski, B. Bryson, L. Cowen, and B. Berger, "Contrastive learning in protein language space predicts interactions between drugs and protein targets," *Proceedings of the National Academy of Sciences*, vol. 120, no. 24, Jun. 2023, doi: 10.1073/pnas.2220778120. [46] M. V. S. Prakash, N. S. Reddy, G. Parab, V. Varun, V. Vaddina, and S. Gopalakrishnan, "Synergistic Fusion of Graph and Transformer Features for Enhanced Molecular Property Prediction," *Cold Spring Hrabor Laboratory*, Aug. 2023. [47] S. Ali, P. Chourasia, and M. Patterson, "When Protein Structure Embedding Meets Large Language Models," *Genes (Basel)*, vol. 15, no. 1, p. 25, Dec. 2023, doi: 10.3390/genes15010025. [48] Ł. Maziarka *et al.*, "Relative molecule self-attention transformer," *J Cheminform*, vol. 16, no. 1, p. 3, Jan. 2024, doi: 10.1186/s13321-023-00789-7. [49] H. Lu, Z. Wei, X. Wang, K. Zhang, and H. Liu, "GraphGPT: A Graph Enhanced Generative Pretrained Transformer for Conditioned Molecular Generation," *Int J Mol Sci*, vol. 24, no. 23, p. 16761, Nov. 2023, doi: 10.3390/ijms242316761. [50]X. Xu *et al.*, "Optimization of binding affinities in chemical space with generative pre-trained transformer and deep reinforcement learning," *F1000Res*, vol. 12, p. 757, Feb. 2024, doi: 10.12688/f1000research.130936.2. [51] Z. Xu, X. Lei, M. Ma, and Y. Pan, "Molecular Generation and Optimization of Molecular Properties Using a Transformer Model," *Big Data Mining and Analytics*, vol. 7, no. 1, pp. 142–155, Mar. 2024, doi: 10.26599/BDMA.2023.9020009. [52] P. Zhou *et al.*, "Instruction Multi-Constraint Molecular Generation Using a Teacher-Student Large Language Model," Mar. 2024. [53]

Y. Kalakoti, S. Yadav, and D. Sundar, "TransDTI: Transformer-Based Language Models for Estimating DTIs and Building a Drug Recommendation Workflow," *ACS Omega*, vol. 7, no. 3, pp. 2706–2717, Jan. 2022, doi: 10.1021/acsomega.1c05203. [54] A. Yüksel, E. Ulusoy, A. Ünlü, and T. Doğan, "SELFormer: molecular representation learning via SELFIES language models," *Mach Learn Sci Technol*, vol. 4, no. 2, p. 025035, Jun. 2023, doi: 10.1088/2632-2153/acdb30. [55] H. Chen and J. Bajorath, "Meta-learning for transformer-based prediction of potent compounds," *Sci Rep*, vol. 13, no. 1, p. 16145, Sep. 2023, doi: 10.1038/s41598-023-43046-5. [56] J. Luo, X. Liu, J. Li, Q. Chen, and J. Chen, "Flexible and Controllable Protein Design by Prefix-tuning Large-Scale Protein Language Models," Dec. 2023. [57] Z. Meng, C. Chen, X. Zhang, W. Zhao, and X. Cui, "Exploring fragment adding strategies to enhance molecule pretraining in AI-driven drug discovery," *Big Data Mining and Analytics*, pp. 1–12, 2024, doi: 10.26599/BDMA.2024.9020003. [58]

K.-C. Chou, "Some remarks on predicting multi-label attributes in molecular biosystems," *Mol Biosyst*, vol. 9, no. 6, p. 1092, 2013, doi: 10.1039/c3mb25555g. [59]

C. Tang, C. Zhong, D. Chen, and J. Wang, "Drug-target interactions prediction using marginalized denoising model on heterogeneous networks," *BMC Bioinformatics*, vol. 21, no. 1, p. 330, Dec. 2020, doi: 10.1186/s12859-020-03662-8. [60] S. Patiyal, A. Dhall, and G. P. S. Raghava, "A deep learning-based method for the prediction of DNA interacting residues in a protein," *Brief Bioinform*, vol. 23, no. 5, Sep. 2022, doi: 10.1093/bib/bbac322. [61] Y. Xia, C.-Q. Xia, X. Pan, and H.-B. Shen, "GraphBind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues," *Nucleic Acids Res*, vol. 49, no. 9, pp. e51–e51, May 2021, doi: 10.1093/nar/gkab044. [62] J. Yang, A. Roy, and Y. Zhang, "BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions," *Nucleic Acids Res*, vol. 41, no. D1, pp. D1096–D1103, Oct. 2012, doi: 10.1093/nar/gks966. [63] Ł. Maziarka, T. Danel, S. Mucha, K. Rataj, J. Tabor, and S. Jastrzębski, "Molecule Attention Transformer," Feb. 2020.