An overview of Pocket-based drug design

Pablo Varas Pardo

March 2024

1 Introduction

In recent years, advances in artificial intelligence (AI) techniques have greatly impacted the field of drug discovery. Researchers have tried to estimate the size of the chemical space, which refers to the number of possible compounds that could be synthesized. However, there is no scientific consensus on the exact number, with some estimating it to be between 10^{30} to 10^{60} (1). The discrepancy in the estimations is due to different criteria applied in the studies, such as the maximum size of the molecules, the types of atoms that compose them, and the presence of physicochemical restrictions like the Lipinski rules (2, 3, 4).

There are various techniques available to examine the vast chemical space and develop new potential drug candidates. Two popular methods to create new molecules are ligand-based and pocket-based. Ligand-based methods use a set of high-affinity molecules to a target protein as input to create new molecules (5, 6). This method is useful when the target protein structure is not available or when the research focuses on a narrow chemical space. The algorithm generates diverse R groups at a specific site on a molecule by referencing the input molecules.

On the other hand, pocket-based methods are used to generate appropriate ligands for protein pockets by utilizing detailed structural information of the target binding site (7). These algorithms are capable of generating suitable ligands for protein pockets based on the three-dimensional structure of the binding site. For this reason, pocket-based algorithms are highly effective when the three-dimensional structure of the protein is available, allowing for precise molecular docking and virtual screening to design molecules that fit within the structural constraints of the binding site. Such approaches can accurately model interactions between the ligand and the protein, facilitating the design of molecules that are optimally configured to the target site

This article will focus on the current methods of generating molecules based on the pocket structure of proteins. This is because they can generate completely novel molecules given a protein conformational structure, which makes them especially useful for the creation of new drugs that inhibit certain proteins involved in the appearance of diseases. Protein overexpression may play a crucial role in the development and progression of several diseases (8). Abnormal regulation of protein expression is associated with multiple disorders, including cancers, neurodegenerative diseases, autoimmune disorders, and cardiovascular diseases, among others (9).

In summary, the main contributions of this article pretend to be:

- Creation of a base framework of current methods based on protein pockets, focusing mainly on those that are open source.

- Carry out experiments on them for evaluation of which ones seem to be most promising in the field of drug design for protein structures, creating a reference framework for the evaluation of the algorithms.
- Develop an open-source approach for filtering optimal compounds given a target protein.

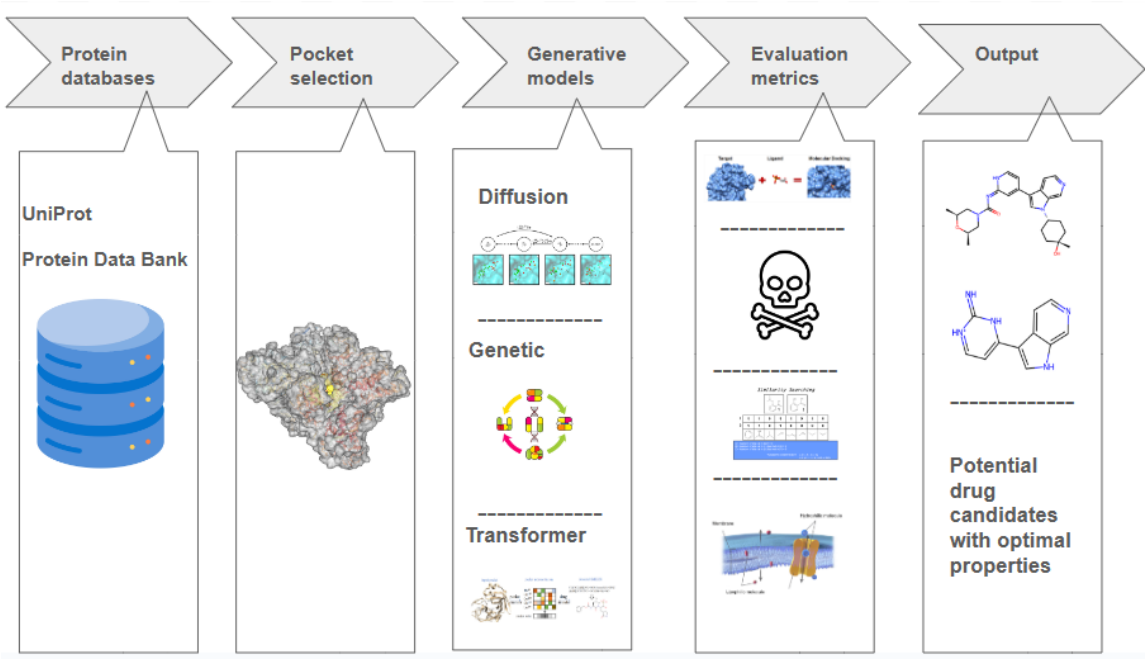

Figure 1: Average Atomic Distance Comparison by Model

2 Pocket-based algorithms

Protein pocket-based models are currently being used to generate new ligands and explore a vast chemical space. In this section, the technical features of each model will be examined. Additionally, Table 1 lists the databases and the number of compounds used to train these algorithms.

2.1 DiffSBDD

Diffusion models have gained a special interest in the last few years due to their capacity to generate novel compounds executing several denoising steps. In the field of pocket-based drug design, this denoising procedure is conditioned on the protein pocket for which the model will create new potential drug candidates. DiffSBDD (10) has proved the capability to generate novel ligands with high predicted binding affinities to given protein pockets, highlighting its potential as a tool for candidate molecule refinement in structure-based drug design.

2.2 DrugGPT

DrugGPT (11) presents a strategy for ligand design using the autoregressive model GPT, focusing on exploring chemical space and discovering ligands for specific proteins. Leveraging deep learning language models, which have shown significant potential across various domains including protein design and biomedical text analysis, DrugGPT employs the GPT-2 model, optimized and trained with a redefined tokenizer using the BPE algorithm for better adaptation to drug design requirements. This improvement enables DrugGPT to capture and comprehend the structural information and chemical rules of drug molecules accurately, enhancing its ability to understand binding information between proteins and ligands, thereby generating potentially active drug candidate molecules.

By redesigning the tokenizer and training the model from scratch, DrugGPT showcases its ability to efficiently generate ligand designs that not only reflect theoretical advantages but also reveal potential applications in the drug development process.

2.3 Lingo3DMol

Transformer-based algorithms are becoming increasingly popular in the field of drug discovery. This is because these algorithms are capable of capturing complex relationships between training data. Lingo3DMol (7) proposes a new representation for SMILES called FSMILES, which fragments the SMILES into molecular groups according to specific rules. The protein pocket is then encoded through the encoder, and starting from the initially calculated growth point, the decoder builds the molecule by iteratively joining the molecular groups produced by the decoder.

2.4 Pocket2Mol

Pocket2Mol (12) introduces an E(3)-equivariant generative network aiming to efficiently sample molecular structures based on the 3D structure of protein pockets. It addresses the limitations of previous approaches that either operate in graph space without considering detailed chemical structures like bond types and functional groups or fail to incorporate the 3D coordinates effectively. The innovation lies in its dual-module design: a novel graph neural network capturing spatial and bonding relationships, and an efficient algorithm for conditional molecular sampling from a tractable distribution without resorting to MCMC. The model significantly outperforms in generating molecules with higher docking values, indicating better fitness as drug candidates. It showcases an advanced ability to optimize the docking value by starting with a selection of drug candidates and improving them through mutation and crossover operations guided by reinforcement learning.

2.5 RGA

Genetic Algorithms (GA) are heuristic algorithms that are inspired by natural evolutionary processes. They use mutation and/or crossover to explore the chemical space. To generate molecules that are based on the protein pocket, a genetic algorithm was used.

RGA proposed (13) a technique that relied on reinforcement learning to choose mutation and crossover operations. The goal was to maximize the docking value, which served as the fitness function to optimize. The process began with the selection of approximately 100 drug candidates

from the initial database. These compounds were then subjected to mutation and crossover to create new ones that would improve the docking score.

As demonstrated in Table 1, a variety of computational models have been trained using distinct datasets, each with its unique size and composition of training compounds. The DrugGPT model utilizes a substantial dataset from 'jglaser/binding affinity' and 'ZINC20', totaling 1.9 million and 20 billion compounds, respectively, underscoring the model's capacity to handle vast chemical spaces. In contrast, DiffSBDD and Pocket2Mol are trained exclusively on the CrossDocked dataset, with DiffSBDD further supplemented by Binding MOAD, incorporating both a large number of compounds (22.5 million) and additional specific data (41,409 compounds). This diverse training regimen enhances the models' ability to generalize across different molecular environments. Lingo3DMol, in its fine-tuning phase, uses a smaller subset of CrossDocked and DUD-E, focusing on precise adjustments with 11.8 thousand and 6.5 thousand compounds, respectively. The use of varied datasets across these models illustrates the breadth of approaches in current computational drug design, aiming to optimize model performance across different types of chemical interactions.

Model	**Databases**	**Training Compounds**
DiffSBDD	CrossDocked, Binding MOAD	22.5 M, 41409
DrugGPT	jglaser/binding affinity, ZINC20	1.9 M, 20 B
Lingo3DMol (fine-tuning phase)	CrossDocked, DUD-E	11.8 K, 6.5 K
Pocket2Mol	CrossDocked	22.5 M
RGA	Zink20	-

Table 1: Summary of models and databases used for training.

3 Evaluation metrics

To create a benchmark for evaluating the different models the metrics that will be taken into account are:

- Virtual Docking: predicts the interaction between a drug molecule and a target protein or enzyme. It simulates the binding affinity and mode of a molecule to a specific biological target. This helps identify potential drug candidates by assessing how well they fit into the target site.

  The use of tools like Smina, provided by PyScreener (14), enhances this process by offering a streamlined, Python-based approach to handling the complex computations required for docking simulations. As outlined in the provided material, PyScreener acts as a wrapper that simplifies integration with various backend docking engines, including Smina, allowing for flexible and efficient computation of docking scores. This functionality is crucial for large-scale screenings and the evaluation of numerous compounds against a set of protein targets, making it an invaluable component of the drug development pipeline

- Pharmacological activity: measures the biological effects of a drug molecule on the body or specific cells or tissues. It evaluates the efficacy and mechanism of action of a compound in producing a therapeutic effect, such as inhibiting a disease-related enzyme or activating a receptor.

The continuous variable pChEMBL, defined as the -$\log_{10}$ of the molar concentration (IC_{50}, XC_{50}, EC_{50}, AC_{50}, Ki, Kd, or Potency), is associated with the pharmacological activity of a compound.

An algorithm (15) has been developed to evaluate pChEMBL based on an initial database of compounds. For preprocessing, Morgan fingerprints of each molecule will be calculated. The user can choose from available models such as Gaussian process, Neural Networks, XGBoost, and Random Forest. For accessing databases of compounds related to a specific protein, please visit ChEMBL (16).

- Quantitative estimation of drug likeliness (QED): measures the likelihood of a chemical compound to be a successful drug candidate based on its physicochemical properties. It quantifies drug-like properties in a single score, considering factors such as solubility, permeability, and molecular weight. These factors are indicative of a compound's ability to be orally active in humans.

- Lipophycity (LogP): measures the tendency of a compound to dissolve in fats, oils, and lipids over aqueous (water-based) solutions. It is often expressed as LogP, which is the logarithm of the partition coefficient between n-octanol and water. It indicates the compound's ability to penetrate cell membranes, affecting its absorption, distribution, metabolism, and excretion properties.

- Molecular Diversity (Tanimoto Similarity): assesses the structural variety among generated molecules. The Tanimoto coefficient is calculated using the formula:

$$\text{Tanimoto Similarity} = \frac{c}{a + b - c}$$

  where a and b are the counts of features present in molecule A and B respectively, and c is the count of features common to both molecules. A higher Tanimoto score indicates greater similarity.

- Molecular Weight: its analysis is crucial for evaluating the molecule's atomic composition. This factor plays a vital role in determining a drug's pharmacokinetics, making it an indispensable element in pharmaceutical research.

- Synthetic Accessibility Score (SAS): It is calculated as the sum of fragment scores plus a complexity penalty. This metric assesses the ease with which a chemical compound can be synthesized, providing a crucial insight into the feasibility of its production on a larger scale.

- Toxicity: In the evaluation of potential drug candidates, assessing the toxicity profile is crucial to determine the safety and viability of molecules for further development. This study measured the toxicity across twelve different classes, using computational predictions to gauge potential adverse effects. The toxicity classes include interactions with various nuclear receptors (e.g., androgenic, estrogenic), responses to environmental and endogenous stress signals (e.g., oxidative stress response, heat shock response), and effects on key cellular processes (e.g., DNA damage response). Each class represents a distinct mechanism through which a compound might exhibit toxicological effects, providing a comprehensive view of its safety profile.

To facilitate these computations, the Chemprop algorithm (17), which employs a directed message-passing neural network to predict molecular properties effectively, was utilized.

Nuclear Receptor Panel (biomolecular targets)	**Stress Response Panel**
ER-LBD: estrogen receptor, luciferase	ARE: nuclear factor (erythroid-derived 2)-like 2 antioxidant responsive element
ER: estrogen receptor alpha	HSE: heat shock factor response element
aromatase	ATAD5: genotoxicity indicated by ATAD5
AhR: aryl hydrocarbon receptor	MMP: mitochondrial membrane potential
AR: androgen receptor	p53: DNA damage p53 pathway
AR-LBD: androgen receptor, luciferase	
PPAR: peroxisome proliferator-activated receptor gamma	

Table 2: Summary of toxicity classes measured in the study.

4 Results

In this section, we thoroughly analyse the performance of various molecular generation models.

It is important to note that the findings presented in this section are related to the protein DYRK1A, which was obtained from the Protein Data Bank (18). In recent years, the DYRK1A enzyme has been identified as a promising target for therapeutic intervention in Alzheimer's disease (AD), as it is involved in multiple biological functions. Studies have shown that DYRK1A undergoes alterations that are linked to the progression of AD, such as the phosphorylation of proteins like TAU and APP. Therefore, DYRK1A is a highly promising enzyme as a therapeutic target for designing new drugs that could potentially be used to treat AD.

The evaluation includes a range of metrics described in the Experiments section designed to give a comprehensive view of each model's capabilities. Collectively, these metrics offer insights into the models' ability to generate structurally innovative and diverse molecules that may be potential therapeutic agents. The results are crucial for understanding the current state of the art in molecular generation technologies, highlighting the strengths and areas for improvement for each model.

	DiffSBDD	DrugGPT	Lingo3DMol	Pocket2Mol	RGA
Docking (Mean ↓)	-6.77	-8.72	-8.71	**-9.21**	-8.53
Docking Top1 (↓)	-9.2	-10.80	-11.30	**-12.51**	-11.2
Docking Top10 (↓)	-8.77	-10.47	-10.64	**-12.08**	-10.61
pCHEMBL (Mean ↑)	5.87	5.99	6.08	5.96	**6.18**
pCHEMBL Top1 (↑)	6.83	7.08	8.03	7.13	**9.01**
pCHEMBL Top10 (↑)	6.59	6.86	7.69	6.89	**8.22**
logP (Mean 2-4)	0.43	**3.98**	**2.46**	**2.68**	**2.80**
QED (Mean ↑)	0.48	0.53	**0.65**	0.67	0.62
SAS (Mean ↓)	4.76	**2.57**	2.83	3.25	3.52
Diversity (↑)	**1**	**1**	0.99	**1**	0.85
Molecular Weight (< 500)	268.67	418.80	335.16	309.73	314.41
Molecules evaluated	271	400	600	562	416

Table 3: Models performance evaluation

Table 3 showcases the algorithmic performance in virtual docking and binding affinity, as exemplified by the pCHEMBL values, along with mean logP and QED scores. The results underline the efficacy of Pocket2Mol in docking score averages and top-ranking molecules, suggesting a superior fit for potential drug candidates within the target protein's binding site. Meanwhile, RGA distinguishes itself in predicting binding affinities, hinting at its utility in identifying potent drug molecules.

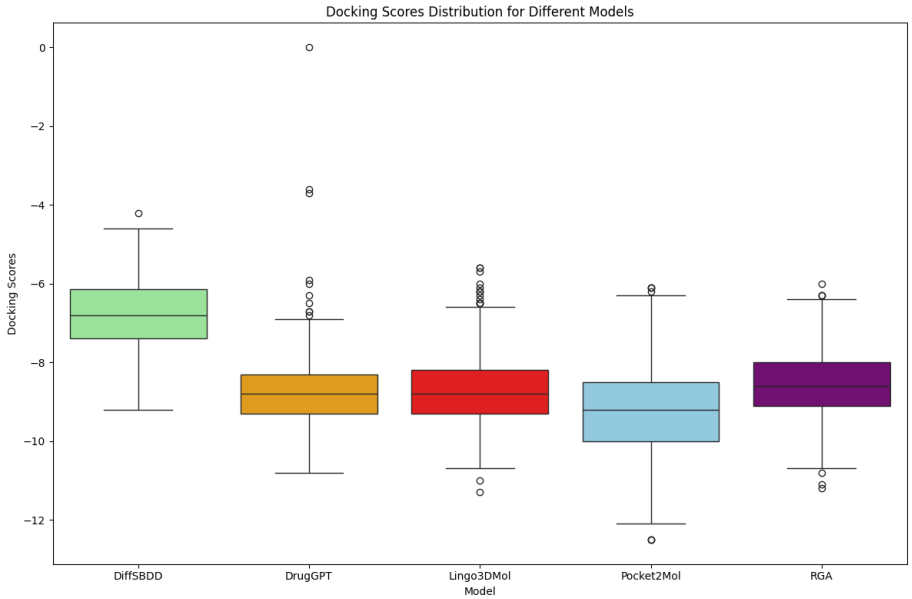

Figure 2: Docking Score Comparison by Model

Figure 2 presents a visual comparison of the docking score distributions for different molecular generation models. The median docking score, indicated by the horizontal line within each

box, serves as a robust indicator of central tendency among the scores achieved by each model. Notably, Pocket2Mol achieves the most favorable median docking score, suggestive of its adeptness in generating molecules that could form stable complexes with target proteins. The spread and range of scores, as denoted by the boxes and whiskers, also offer insights into the consistency and reliability of each model's predictions. For instance, while RGA shows a comparatively tight distribution signaling consistency in scoring, Lingo3DMol demonstrates a broader range, indicating greater variability in its docking score predictions. Outliers, represented by individual dots, highlight exceptional cases where molecules exhibit either particularly high or low docking scores compared to the typical range of the model's output. Collectively, this graphical representation underscores the variegated performance landscape of these models, guiding researchers to prudent selections depending on the specific requirements of their drug design endeavors.

Ring Size	DrugGPT	DiffSBDD	RGA	Pocket2Mol	Lingo3DMol
3	0.27%	36.71%	1.36%	0.05%	-
4	0.27%	5.27%	0.41%	-	0.28%
5	25.72%	21.94%	31.45%	23.96%	36.44%
6	72.61%	30.38%	66.78%	72.21%	63.14%
7	0.94%	4.22%	-	2.37%	0.14%
8	0.07%	0.84%	-	0.33%	-
9	-	0.42%	-	0.09%	-
≥ 10	0.20%	1.91%	-	2.15%	-

Table 4: Ring size distribution

The distribution of ring sizes across the generated molecules by different models, as presented in Table 4, offers a glimpse into the structural diversity each algorithm brings to the table. The models display varied propensity towards certain ring sizes, which could be indicative of their bias towards or against specific molecular frameworks. This diversity in structural motifs is essential in the exploration of the vast chemical space.

Number of Non-Toxic Classes	DrugGPT	DiffSBDD	RGA	Pocket2Mol	Lingo3DMol
6	0.00%	0.00%	3.85%	0.36%	0.00%
7	2.00%	1.11%	7.45%	2.85%	1.50%
8	9.00%	0.37%	19.47%	9.96%	8.33%
9	18.25%	2.58%	25.72%	21.89%	24.83%
10	13.00%	7.01%	14.90%	12.46%	8.00%
11	17.25%	15.13%	13.94%	22.60%	28.83%
12	40.50%	**73.80**%	14.66%	29.89%	28.50%

Table 5: Toxicity evaluation

Table 5 and Figure 3 illustrates the percentage of molecules across non-toxic classes for different models provides an insight into the safety profile of the molecules generated. Models like DiffSBDD and DrugGPT demonstrate a substantial percentage of molecules falling into non-toxic classes, particularly in higher non-toxic class counts, which may reflect a favorable safety profile for subsequent drug development stages.

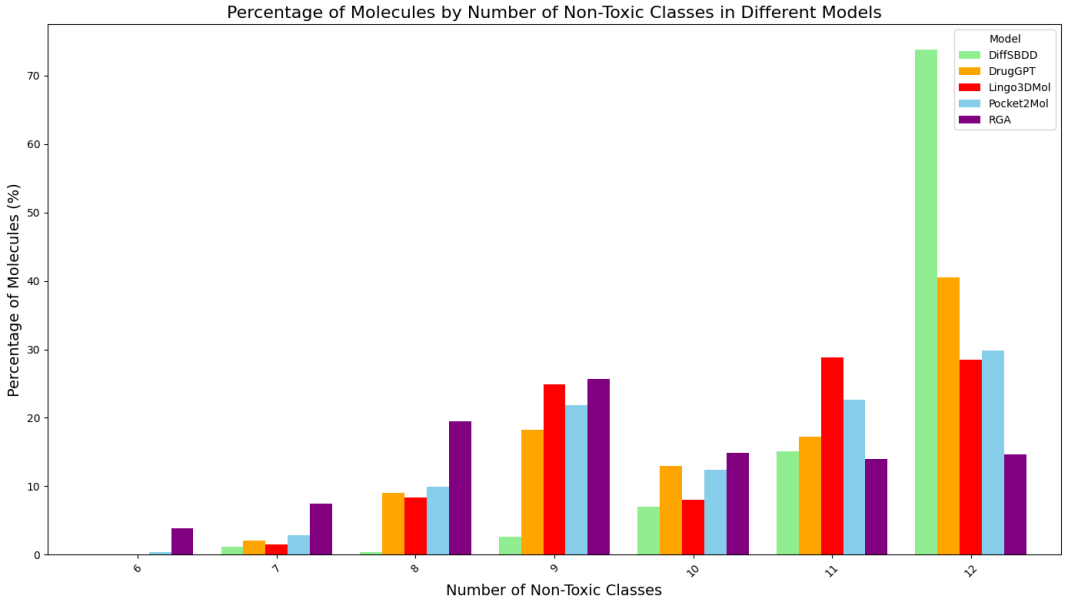

Figure 3: Toxicity Comparison by Model

Figure 4 utilizes a t-Distributed Stochastic Neighbor Embedding (t-SNE) plot to illustrate the molecular diversity generated by various computational models. Each point in the plot represents a molecule, with the proximity between points corresponding to their structural similarity.

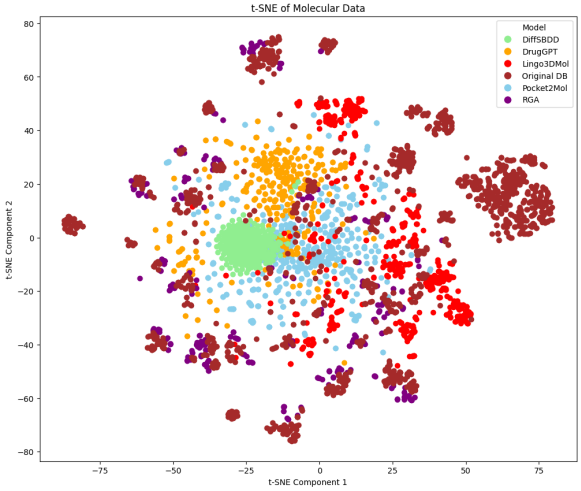

Figure 4: Diversity Comparison by Model

The distinct clustering of points suggests that each model has a unique signature in terms of molecular generation, with some models, as indicated by the concentrated clusters, producing

molecules with higher structural homogeneity. In contrast, the more dispersed clusters reveal models that generate a more structurally diverse set of molecules. The presence of distinct and well-defined clusters also implies that certain models may specialize in particular regions of the chemical space, potentially aligning with specialized drug discovery objectives. The original molecules are marked separately, serving as a baseline reference for the diversity introduced by each model. Overall, this figure highlights the importance of diversity in molecular design and the capacity of different models to explore the vast expanse of chemical space.

The RGA model produces molecules that are grouped in a molecular space quite similar to that of the original database. This is expected since the genetic algorithm forms molecules by molecular subgroups.

Similarly, the DIFSBDD molecules are grouped in a specific chemical space, similar to those of DrugGPT. On the other hand, both Pocket2Mol and Lingo3DMol molecules achieve greater molecular diversity by exploring different areas of chemical space that are distant from the molecules in the original database.

5 Conclusions

The use of artificial intelligence (AI) tools like diffusive models, genetic algorithms, or transformers in designing molecules has gained significant attention in recent years. This is due to their ability to generate new molecules related to a given protein.

In this article, the effectiveness of five algorithms was evaluated. These algorithms proved to be highly effective in developing new molecules and exploring new areas of the vast chemical space. They were able to create molecules with a high affinity for the DYRK1A protein, which is involved in Alzheimer's disease. The results of docking with Smina were later corroborated with traditional docking tools.

Furthermore, an evaluation framework was proposed that contains different metrics to help evaluate the performance of these models. As seen in the article, for a molecule to be evaluated by chemists in the laboratory, it must pass numerous filters that make it a candidate molecule.

Finally, since running these models can be complicated, a simple code is provided so that anyone can generate new molecules given the PDB of a protein and the coordinates of its protein pocket.

6 Code availability

The GitHub repository contains links to run the models used in the article, as well as the tools used to evaluate their quality using the metrics described in the article.

7 Acknowledgments

References

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

[2] H. R. Henze and C. M. Blair, "The number of isomeric hydrocarbons of the methane series," *Journal of the American Chemical Society*, vol. 53, no. 8, pp. 3077–3085, 1931.

[3] D. Weininger, "Combinatorics of small molecular structures," *Encyclopedia of Computational Chemistry*, vol. 1, 2002.

[4] C. M. Blair and H. R. Henze, "The number of stereoisomeric and non-stereoisomeric paraffin hydrocarbons," *Journal of the American Chemical Society*, vol. 54, no. 4, pp. 1538–1545, 1932.

[5] A. C. Kaushik, A. Kumar, S. Bharadwaj, R. Chaudhary, S. Sahi, A. C. Kaushik, A. Kumar, S. Bharadwaj, R. Chaudhary, and S. Sahi, "Ligand-based approach for in-silico drug designing," *Bioinformatics Techniques for Drug Discovery: Applications for Complex Diseases*, pp. 11–19, 2018.

[6] F. Imrie, T. E. Hadfield, A. R. Bradley, and C. M. Deane, "Deep generative design with 3d pharmacophoric constraints," *Chemical science*, vol. 12, no. 43, pp. 14 577–14 589, 2021.

[7] W. Feng, L. Wang, Z. Lin, Y. Zhu, H. Wang, J. Dong, R. Bai, H. Wang, J. Zhou, W. Peng *et al.*, "Generation of 3d molecules in pockets via a language model," *Nature Machine Intelligence*, pp. 1–12, 2024.

[8] A. Khan, S. Khan, A. Jan *et al.*, "Health complication caused by protein deficiency," *J. Food Sci. Nutr*, vol. 1, pp. 645–647, 2017.

[9] C. Soto, "Unfolding the role of protein misfolding in neurodegenerative diseases," *Nature Reviews Neuroscience*, vol. 4, no. 1, pp. 49–60, 2003.

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

[11] Y. Li, C. Gao, X. Song, X. Wang, Y. Xu, and S. Han, "Druggpt: A gpt-based strategy for designing potential ligands targeting specific proteins," *bioRxiv*, pp. 2023–06, 2023.

[12] X. Peng, S. Luo, J. Guan, Q. Xie, J. Peng, and J. Ma, "Pocket2mol: Efficient molecular sampling based on 3d protein pockets," in *International Conference on Machine Learning*. PMLR, 2022, pp. 17 644–17 655.

[13] T. Fu, W. Gao, C. Coley, and J. Sun, "Reinforced genetic algorithm for structure-based drug design," *Advances in Neural Information Processing Systems*, vol. 35, pp. 12 325–12 338, 2022.

[14] D. E. Graff and C. W. Coley, "pyscreener: A python wrapper for computational docking software," *arXiv preprint arXiv:2112.10575*, 2021.

[15] P. Varas, "pchembl prediction," 2024, Último acceso el 11 de abril de 2024. [Online]. Available: https://github.com/diegolfor9/pCHEMBL-prediction.git

[16] ChEMBL Database, "Chembl," 2024, Último acceso el 11 de abril de 2024. [Online]. Available: https://www.ebi.ac.uk/chembl/

[17] E. Heid, K. P. Greenman, Y. Chung, S.-C. Li, D. E. Graff, F. H. Vermeire, H. Wu, W. H. Green, and C. J. McGill, "Chemprop: A machine learning package for chemical property prediction," *Journal of Chemical Information and Modeling*, vol. 64, no. 1, pp. 9–17, 2023.

[18] U. Rothweiler, "Dyrk1a in complex with xmd7-117," *Protein Data Bank*, 2018, classification: TRANSFERASE. Organism(s): Homo sapiens. Expression System: Escherichia coli. Mutation(s): No. Deposited: 2017-09-19. Released: 2018-08-29. URL: https://www.rcsb.org/structure/6eif.

A1 DiffSBDD

Algorithm 1: Equivariant Diffusion Model for Structure-Based Drug Design (DiffSBDD)

Data: Protein pockets P and training set of ligands L
Result: Generate new ligands L^* optimized for binding to protein pockets
Initialize diffusion parameters θ, noise levels σ_t, and set total timesteps T;
Train tokenizer $\mathcal{T}$ and prepare data embeddings using SE(3)-equivariant GNNs;
`// Training Phase`
for *each epoch* **do**
  for *each protein pocket* $p \in P$ *and corresponding ligand* $l \in L$ **do**
    Represent p and l as point clouds z_p and z_l;
    Initialize diffusion process with $z_{l,0} = z_l$ and apply noise to get $z_{l,t}$;
    for $t = T$ ***downto*** 1 **do**
      Predict noise $\epsilon_\theta(z_{l,t}, t)$ using the model;
      Perform denoising step to estimate $z_{l,t-1} \sim p_\theta(z_{l,t-1}|z_{l,t})$;
      Update θ by minimizing loss between $z_{l,t-1}$ and $z_{l,t-1}^{predicted}$;
    end
  end
end
`// Generation Phase`
for *each protein pocket* $p^* \in P$ **do**
  Sample a new ligand point cloud $z_{l^*,T} \sim \mathcal{N}(0, I)$;
  for $t = T$ ***downto*** 1 **do**
    Refine $z_{l^*,t-1} \sim p_\theta(z_{l^*,t-1}|z_{l^*,t}, p^*)$;
  end
  Convert the final point cloud $z_{l^*,0}$ back to molecular structure L^*;
end

A2 DrugGPT

Algorithm 2: Simplified DrugGPT for Ligand Design

Data: Set of protein-ligand pairs (S_p, S_l) for training, protein sequence S_p^* for inference
Result: Generate optimized ligand S_l^* for given protein S_p^*

Initialize tokenizer $\mathcal{T}$ using Byte Pair Encoding (BPE);
Train tokenizer $\mathcal{T}$ on protein and ligand sequences to optimize vocabulary;
Initialize DrugGPT model $\mathcal{M}$ with parameters θ;
`// Training Phase`
for *each epoch* **do**
  for *each protein-ligand pair* (S_p, S_l) **do**
    Tokenize (S_p, S_l) to $(T_p, T_l) = \mathcal{T}(S_p, S_l)$;
    Prepare input sequence $X =$ '¡start¿' $+ T_p + T_l +$ '¡end¿';
    Train $\mathcal{M}$ to predict sequence $\hat{T}_l$ similar to T_l;
    Update $\mathcal{M}$ parameters θ to minimize prediction loss;
  end
  Validate model performance on unseen data;
end
`// Inference Phase`
Tokenize protein sequence $T_p^* = \mathcal{T}(S_p^*)$;
Start generation with $Y^* =$ '¡start¿' $+ T_p^*$;
while *not end of sequence* **do**
  Predict next token $\hat{y}$ using $\mathcal{M}$;
  Append $\hat{y}$ to Y^*;
  if $\hat{y} ==$ *'¡end¿'* **then**
    break;
  end
end
Convert Y^* to ligand structure S_l^* using inverse tokenizer $\mathcal{T}^{-1}$;

A3 Lingo3DMol

Algorithm 3: Generation of a 3D molecule using Lingo3DMol

Data: Protein pocket P and parameters μ
Result: Generate optimized 3D molecule M for the target pocket
Initialize model parameters μ;
Pre-train model on large-scale molecule data to optimize μ;
Fine-tune model on specific protein-ligand data set;
`// Generation Phase`
Input target pocket P;
for *each atom in pocket* P **do**
  Compute input features $f_i = \text{Encoder}(P_i; \mu)$ for each atom P_i in pocket P;
end
Initialize molecule M with empty set of atoms;
Select starting position in pocket based on NCI/Anchor prediction;
while *not complete molecule* **do**
  Predict next atom's FSMILES token and corresponding coordinates using current state of M;
  $M \leftarrow M \cup \{\text{new atom}\}$;
  Update model state for next prediction;
end
Convert FSMILES tokens and coordinates into a 3D molecular structure;

A4 Pocket2Mol

Algorithm 4: Efficient Molecular Sampling with Pocket2Mol

Data: Set of protein pockets P and known ligands L
Result: Generate new ligands L^* optimized for binding to protein pockets

Initialize model ϕ with parameters θ;
Train the model on dataset $\{(P_i, L_i)\}$ to learn molecular configurations;
// Training Phase
for *each pocket* P_i *and ligand* L_i *in training data* **do**
    Represent P_i and L_i as graphs G_{P_i} and G_{L_i};
    Mask random parts of L_i to simulate incomplete data;
    Use ϕ to predict missing parts of L_i based on G_{P_i};
    Update θ by minimizing loss between predicted and actual configurations of L_i;
end
// Generation Phase
for *each new protein pocket* P^* **do**
    Initialize ligand graph $G^0_{L^*}$ with minimal structure;
    while *not full ligand generated* **do**
        Use ϕ to predict next atom or group to add to $G^t_{L^*}$;
        Update $G^{t+1}_{L^*}$ with new atom or group;
        If model predicts end of generation, terminate;
    end
    Convert final graph G_{L^*} to molecular structure L^*;
end

A5 RGA

Algorithm 5: Reinforced Genetic Algorithm for Structure-Based Drug Design

Data: Population of molecules $\mathcal{P}$, Binding site information of target protein
Result: Optimized molecules with enhanced binding affinity

Initialize population $\mathcal{P}(0)$ randomly;
for *generation* $t = 1$ **to** T **do**
    for *each molecule* $m \in \mathcal{P}(t-1)$ **do**
        Select parents p_1, p_2 from $\mathcal{P}(t-1)$ based on fitness scores;
        Generate offspring o by crossover of p_1, p_2;
        Mutate offspring o to produce mutants o';
        Evaluate binding affinity of o' using docking simulations;
    end
    Select top N offspring to form new population $\mathcal{P}(t)$;
    if *convergence or* $t = T$ **then**
        break;
    end
    Optionally refine neural model with reinforcement learning using performance feedback;
end

B1 Protocol

- **Toxicity Filter**: The toxicity filter ensures that a molecule G is classified as nontoxic in all 12 toxicity classes:

$$\sum_{i=1}^{12} f_{\text{tox}}^{(i)}(G) = 0$$

- **QED Filter**: The QED filter ensures that the quantitative estimate of drug-likeness (QED) for molecule G exceeds a predefined threshold $QED_{\text{threshold}}$:

$$QED(G) > QED_{\text{threshold}}$$

- **LogP Filter**: The LogP filter ensures that the octanol-water partition coefficient (LogP) for molecule G falls within a predefined range $[\text{LogP}_{\text{min}}, \text{LogP}_{\text{max}}]$:

$$\text{LogP}_{\text{min}} \leq \text{LogP}(G) \leq \text{LogP}_{\text{max}}$$

- **Similarity Filter**: The similarity filter ensures that the similarity between a molecule G and a dataset D is below a predefined threshold $\text{similarity}_{\text{threshold}}$:

$$s_D(G) < \text{similarity}_{\text{threshold}}$$

- **Molecular Weight Filter**: The molecular weight filter ensures that the molecular weight of molecule G falls within a predefined range $[MW_{\text{min}}, MW_{\text{max}}]$:

$$MW_{\text{min}} \leq MW(G) \leq MW_{\text{max}}$$

- **Synthetic Accessibility Filter**: The synthetic accessibility filter ensures that the synthetic accessibility score (SAS) for molecule G is below a predefined threshold $SAS_{\text{threshold}}$:

$$SAS(G) < SAS_{\text{threshold}}$$

- **HBA Filter**: The HBA filter ensures that the number of hydrogen bond acceptors (HBA) for molecule G is below a predefined threshold HBA_{max}:

$$\text{HBA}(G) \leq \text{HBA}_{\text{max}}$$

- **HBD Filter**: The HBD filter ensures that the number of hydrogen bond donors (HBD) for molecule G is below a predefined threshold HBD_{max}:

$$\text{HBD}(G) \leq \text{HBD}_{\text{max}}$$