DIFFUSION GENERATIVE MODELS FOR MOLECULE OPTIMIZATION

Xiaochuan Zha, Xinyan Gao, Wenxue Hui & Zonghua Luo
School of Biomedical Engineering, ShanghaiTech University
{zhaxch,gaoxy2023,huiwx,luozh}@shanghaitech.edu.cn

ABSTRACT

In the pursuit of novel drug molecules, the optimization stage for enhanced safety, efficacy, and pharmacokinetics presents a significant challenge. Deep generative models, particularly diffusion models, emerge as an effective strategy for this purpose. Our method integrates 3D protein structures with diffusion models, facilitating the generation of new ligands that consider both the molecular scaffold and the specific environment of the target protein's binding pocket. As a demonstration of its effectiveness, we applied this proposed approach to generate new ligands targeting the colony-stimulating factor 1 receptor. The outcomes highlight the method's ability to design potent inhibitors, achieving enhanced inhibitory efficacy in compared to existing inhibitors, as confirmed by *in vitro* assays.

1 INTRODUCTION

In the early stage of drug discovery, optimizing molecules is a critical phase. This process involves on modifying the scaffold of a molecule to improve key properties such as safety, efficacy, and pharmacokinetics. However, the complex chemical space presents significant challenges in identifying the optimal modifications, often leading to extensive time and resource consuming. To address these challenges, recent advancements have introduced deep generative models (DGMs) into the field of molecule generation, offering a promising approach for overcoming these obstacles Anstine & Isayev (2023) Renz et al. (2019).

Among the various types of DGMs, diffusion models have displayed superior performance, with sucessful applications in image generation Ho et al. (2020), molecule docking Corso et al. (2022), and protein engineering Watson et al. (2023). These models excel at navigating through complex, high-dimensional data distributions, which is crucial for exploring the intricate chemical space involved in molecule optimization Guo et al. (2023). Building on this foundation, our work aims to apply diffusion models for molecule optimization process. By integrating 3D structures of target proteins, our approach generate new ligand fragments that account for both the molecular scaffold and the specific environment of the target protein's binding pocket. This method is designed to enable a more precise and efficient exploration of chemical space, potentially accelerating the optimization process.

To demonstrate the effectiveness of our approach, we focused on optimizing molecules targeting the colony-stimulating factor 1 receptor (CSF1R), a tyrosine kinase linked to a variety of diseases, including neurodegeneration Anstine & Isayev (2023) and cancers Wen et al. (2023). By employing diffusion models, we sought to generate and refine molecules with enhanced ability to inhibit CSF1R. The results of this effort were encouraging, showcasing the practical utility and efficiency of diffusion model-based molecule generation in the context of optimization process.

2 RELATED WORKS

Small molecule Discovery with deep generative models. The use of the Simplified Molecular-Input Line-Entry System (SMILES) for representing molecules has allowed for the application of neural language models like Long Short-Term Memory (LSTM) networks in generating molecules. This method has notably advanced *de novo* molecule discovery for targets such as receptor interacting protein kinase 1 Li et al. (2022) and phosphoinositide 3-kinase gamma Moret et al.

(2023). AstraZeneca's researchers have utilized SMILES for molecule optimization Arús-Pous et al. (2020).Moreover, Tan et al. have validated this approach by identifying a highly potent molecule for the discoidin domain receptor 1, with an IC50 value of 10.6 nM Tan et al. (2021).

Advancements in molecular generation with diffusion models. The application of diffusion models in molecular generation has seen considerable progress. A key development was the Equivariant Diffusion Model (EDM) by Emiel Hoogeboom and colleagues for 3D molecular generation Hoogeboom et al. (2022). The field has recently shifted towards generating molecules for specific conditions.Notable advancements include DiffLiker for linker design Igashov et al. (2022) and SBDD for structure-based drug design Schneuing et al. (2022). Similar diffusiom models for 3D molecule generation are PMDM Huang (2023) and Diffbp Lin et al. (2022).

3 Method

The proposed method for designing small molecule inhibitors targeting specific biomarkers integrates two primary stages: molecule generation based on the target protein and scaffold structure, followed by docking to screen molecules.

Conditional Ligand Generation Using Equivariant Diffusion Models. Our approach diverges from other generative models that generate entire molecules within a protein pocket, focusing instead on generating molecular fragments only based on the scaffold as depicted in Figure 1a. Central to our method is an equivariant diffusion model, specifically designed to create small molecules conditioned on the protein pocket environment. This involves a two-phase process: a diffusion phase that incrementally introduces noise, transforming data into Gaussian noise, and a denoising phase that reverses this process, informed by a learnable function based on a modified E(3)-equivariant graph neural network (EGNN) Satorras et al. (2021).

As illustrated in Figure 1b, we start by selecting a ligand scaffold randomly in training stage. The protein pocket and this scaffold form the constant context through the diffusion and denoising phases. Our model was trained using the MOAD databaseHu et al. (2005) and comprises six layers of Equivariant Graph Convolutional Layers (EGCL) with 10.6 million parameters. Diffusion step set to 500.

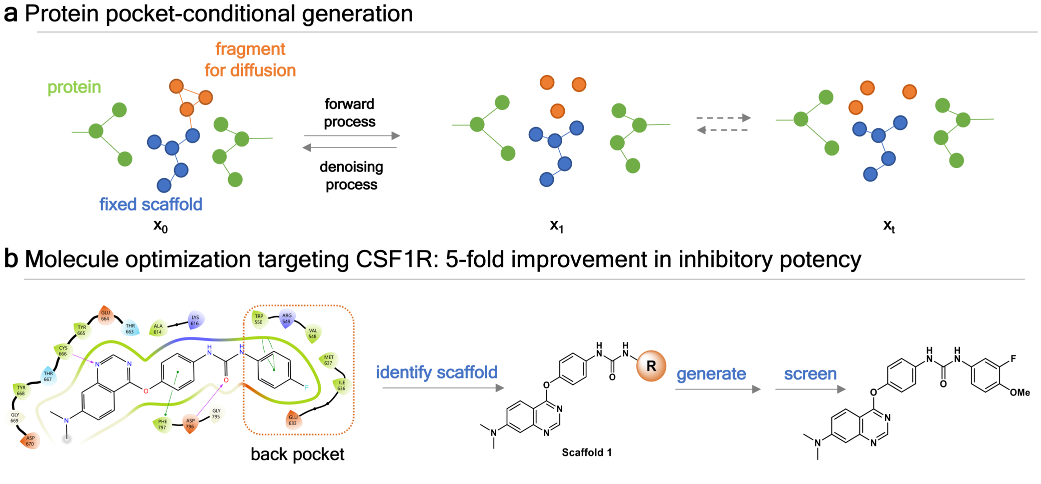


Figure 1: Schematic of Diffusion model based optimization process. Panel a illustrates the protein pocket-conditional generation using a diffusion model, where a fixed scaffold and a variable fragment undergo a forward diffusion process followed by a denoising process to generate the optimized molecule. Panel b outlines the molecule optimization workflow for the CSF1R target. The workflow starts with the identification of a suitable scaffold, followed by the generation of novel compounds, and concludes with the screening of these compounds, highlighting the final high-affinity molecular structure enhanced with a benzene ring fragment for optimal binding within the CSF1R back pocket.

For ligand generation, we prioritize scaffold selection to ensure compatibility with the protein pocket environment. The resulting ligand fragments, ranging from 4 to 13 atoms (excluding hydrogen), are subjected to structural feasibility checks before being converted into molecular structures.

Docking. We evaluated our model's ability to produce ligands with strong binding affinities using docking simulations, a widely adopt approach for molecule generationCieplinski et al. (2023). Docking scores, which are inversely related to binding affinity, provide a quantitative measure of a ligand's potential effectiveness in binding to the target protein. A lower docking score indicates a higher predicted affinity, marking the ligand as a potential therapeutic agent. We conducted docking with Glide from Schrödinger, Inc., using the CSF1R PDB code 6WXJ.

4 Results

In our pursuit of identifying novel inhibitors targeting CSF1R through molecule optimization, we selected compound 1 (depicted in Figure 1b) from a recent patent Hsieh et al. (2020) as our lead compound, noted for its potent effects and therapeutic promise. Compound 1 interacts with CSF1R through hydrogen bonds between its quinazoline moiety and Cys666 in the hinge region, and the urea moiety's interactions with Asp796 and Glu633, alongside $\pi-\pi$ stacking with Phe797's benzene ring. To enhance the inhibitory efficacy of compound 1 against CSF1R, our analysis led us to select scaffold 1 for for a targeted exploration of the back pocket. This decision was based on the hypothesis that the back pocket is a relatively undiscovered area, and that focusing on this area could greatly increase affinity and potentially enhance selectivity Lee et al. (2021).

To this end, we generated a diverse virtual library comprising 200 molecules for each specified atom number range from 8 to 13. This extensive collection underwent a meticulous assessment based on their physicochemical properties, followed by a filtration process utilizing docking simulations to identify the most promising candidates. From this process, six novel molecules were synthesized, culminating in the discovery of potent inhibitors.

4.1 Model Performance and Properties Distribuation

The efficacy of our model in the field of small molecule generation was rigorously evaluated through a set of established metrics: validity, uniqueness, and diversity, complemented by an in-depth analysis of attributes critical for therapeutic effectivenessBrown et al. (2019). Validity, which quantifies the model's compliance with structural and chemical benchmarks, was notably high, with Scaffold 1 achieving a 94.6% score. This underscores the model's adeptness at accurately replicating molecular configurations. Uniqueness, a measure critical to the discovery of novel therapeutic agents, was particularly impressive for scaffold 1, recorded at 80.0%. Moreover, the measure of diversity, reflecting the structural variance among the generated molecules, stood at 58.0% for scaffold 1, thereby ensuring a comprehensive exploration of the chemical space.

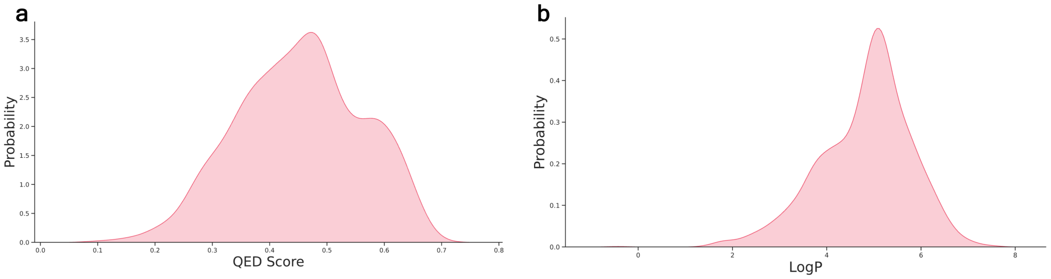


Figure 2: Properties distribution of generated molecules. Panels a and b depict the probability distributions for the Quantitative Estimate of Drug-likeness (QED) score and the octanol-water partition coefficient (LogP), respectively, illustrating the physicochemical profile of generated compounds. All properties was calculated by RDKit Landrum et al. (2013)

Additionally, the model's capacity to distribute key properties, such as the quantitative estimation of drug-likeness (QED score), LogP, and topological polar surface area (TPSA), was meticulously examined (Figure 2a, Figure 2b). The generated molecules based on scaffold 1, displayed a broad and heterogeneous distribution of these properties, underscoring the model's ability to accommodate various drug-likeness criteria, including lipophilicity and polarity.

4.2 Docking analysis

Figure 3a's kernel density estimates (KDE) for docking scores substantiate the effectiveness of our generative model. Notably, a significant number of ligands associated with scaffold 1 demonstrated lower docking scores, indicative of enhanced binding affinity. This phenomenon manifests as a left-skewed distribution in the KDE, suggesting a tendency towards more robust ligand-target interactions. Furthermore, the relationship depicted in Figure 3b between docking scores and molecular size, including the number of atoms generated, exhibits a discernible trend. Within this context, larger molecular sizes correlate with lower docking scores, underscoring the spatial limitations within the back pocket of CSF1R.

To summarize, our generative model has demonstrated superior performance across all critical metrics and physicochemical properties. Concurrently, the docking scores have been satisfactory. This foundation build confidence for next experimental endeavors.

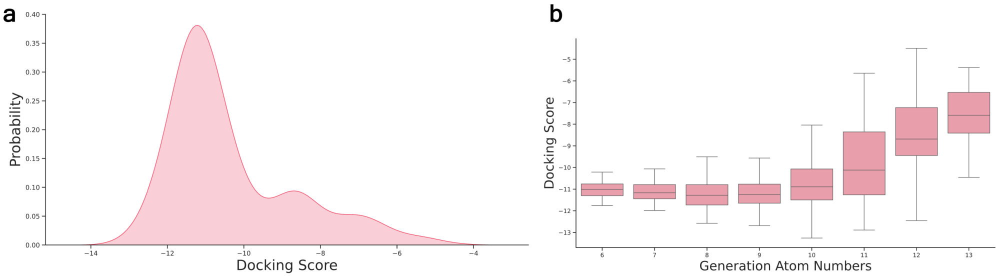


Figure 3: Distribution of Docking Score. Panel a shows the distribution of docking scores, reflecting the binding affinity of compounds to the CSF1R protein target. Panel b presents a box plot of the docking scores across different generation sizes of molecules.

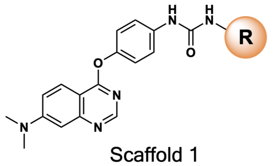
4.3 Experiment Validation

Upon utilizing guidance from our generative diffusion models and subsequent docking analyses, we performed docking for compound 1 and used its docking score as a benchmark for screening. Based on their likelihood of synthetic success, compound 2-7 were selected for synthesis. All six compounds were then synthesized and evaluated for their CSF1R inhibitory activity. Detailed descriptions of the synthetic and enzyme inhibition assay methodologies are provided in the Appendix.

As indicated in Figure 4, the results showed that all the six compounds exhibited excellent CSF1R inhibitory activity with IC_{50} values less than 50 nM, which is higher than the known CSF1R compounds, PLX647 Zhang et al. (2013) and GW2580 Conway et al. (2005). This represents a remarkable enhancement in inhibitory efficacy, with the most effective compound achieving more than a five-fold improvement. Our assay experiment results demonstrate that our methodology successfully achieved the initial goal of molecular optimization, significantly enhancing inhibitory efficacy.

5 Conclusion

Our research introduces generative machine learning as a novel method for molecular optimization in drug development. We began by constructing our diffusion models for molecule generation, followed by a validation process using CSF1R as a biological target. We have successfully identified a suite of novel inhibitors, which showed considerable improvements in potency compared to the lead compounds.



Scaffold 1

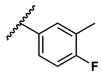
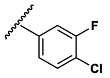
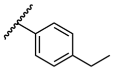
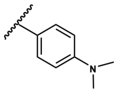
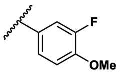
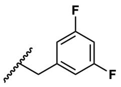
Compound	R	CSF1R IC$_{50}$ (nM)	Docking Score
2		28.4	-11.7
3		39.7	-11.9
4		48.2	-11.4
5		34.2	-12.0
6		26.9	-11.6
7		43.1	-11.8
Compound 1		147.1	-10.8
PLX647		78.6	-12.5
GW2580		162.1	-8.98

Figure 4: Comparative Analysis of Inhibitory Potency. The table in this figure presents novel inhibitors (compounds 2-7) with their molecular structures and corresponding inhibitory concentration (IC50) curves, reflecting the binding efficacy to CSF1R. These novel compounds demonstrate a range of IC50 values, indicating potent inhibitory activities. This table also displays known inhibitors including compound 1, PLX647, and GW2580 for comparison. The IC$_{50}$ values and activity profiles highlight the advancements achieved with the novel inhibitors in terms of increased binding affinity.

Our study provides a valuable tool that can speed up the molecule optimization process, reduce on resource use, and possibly shorten the time needed for drug development. We are convinced that applying generative machine learning approaches in the early stages of drug discovery will significantly benefit the development of new therapeutics.

References

Dylan M Anstine and Olexandr Isayev. Generative models as an emerging paradigm in the chemical sciences. *Journal of the American Chemical Society*, 145(16):8736–8750, 2023.

Josep Arús-Pous, Atanas Patronov, Esben Jannik Bjerrum, Christian Tyrchan, Jean-Louis Reymond, Hongming Chen, and Ola Engkvist. Smiles-based deep generative scaffold decorator for de-novo drug design. *Journal of cheminformatics*, 12(1):1–18, 2020.

Nathan Brown, Marco Fiscato, Marwin HS Segler, and Alain C Vaucher. Guacamol: benchmarking models for de novo molecular design. *Journal of chemical information and modeling*, 59(3): 1096–1108, 2019.

Tobiasz Cieplinski, Tomasz Danel, Sabina Podlewska, and Stanisław Jastrzebski. Generative models should at least be able to design molecules that dock well: A new benchmark. *Journal of Chemical Information and Modeling*, 2023.

James G Conway, Brad McDonald, Janet Parham, Barry Keith, David W Rusnak, Eva Shaw, Marilyn Jansen, Peiyuan Lin, Alan Payne, Renae M Crosby, et al. Inhibition of colony-stimulating-factor-1 signaling in vivo with the orally bioavailable cfms kinase inhibitor gw2580. *Proceedings of the National Academy of Sciences*, 102(44):16078–16083, 2005.

Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. *arXiv preprint arXiv:2210.01776*, 2022.

Zhiye Guo, Jian Liu, Yanli Wang, Mengrui Chen, Duolin Wang, Dong Xu, and Jianlin Cheng. Diffusion models in bioinformatics and computational biology. *Nature Reviews Bioengineering*, pp. 1–19, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Emiel Hoogeboom, Víctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3D. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8867–8887. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/hoogeboom22a.html`.

H. Hsieh, K. Lee, W. Lin, and C. Shih. Heterocyclic compounds as kinase inhibitors for therapeutic uses, 2020. Patent WO2020210481.

Liegi Hu, Mark L Benson, Richard D Smith, Michael G Lerner, and Heather A Carlson. Binding moad (mother of all databases). *Proteins: Structure, Function, and Bioinformatics*, 60(3):333–340, 2005.

Lei Huang. A dual diffusion model enables 3d binding bioactive molecule generation and lead optimization given target pockets. *bioRxiv*, pp. 2023–01, 2023.

Ilia Igashov, Hannes Stärk, Clément Vignac, Victor Garcia Satorras, Pascal Frossard, Max Welling, Michael Bronstein, and Bruno Correia. Equivariant 3d-conditional diffusion models for molecular linker design, 2022.

Greg Landrum et al. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*, 8:31, 2013.

Kun-Hung Lee, Wan-Ching Yen, Wen-Hsing Lin, Pei-Chen Wang, You-Liang Lai, Yu-Chieh Su, Chun-Yu Chang, Cai-Syuan Wu, Yu-Chen Huang, Chen-Ming Yang, et al. Discovery of bpr1r024, an orally active and selective csf1r inhibitor that exhibits antitumor and immunomodulatory activity in a murine colon tumor model. *Journal of Medicinal Chemistry*, 64(19):14477–14497, 2021.

Yueshan Li, Liting Zhang, Yifei Wang, Jun Zou, Ruicheng Yang, Xinling Luo, Chengyong Wu, Wei Yang, Chenyu Tian, Haixing Xu, et al. Generative deep learning enables the discovery of a potent and selective ripk1 inhibitor. *Nature Communications*, 13(1):6891, 2022.

Haitao Lin, Yufei Huang, Meng Liu, Xuanjing Li, Shuiwang Ji, and Stan Z Li. Diffbp: Generative diffusion of 3d molecules for target protein binding. *arXiv preprint arXiv:2211.11214*, 2022.

Michael Moret, Irene Pachon Angona, Leandro Cotos, Shen Yan, Kenneth Atz, Cyrill Brunner, Martin Baumgartner, Francesca Grisoni, and Gisbert Schneider. Leveraging molecular structure and bioactivity with chemical language models for de novo drug design. *Nature Communications*, 14(1):114, 2023.

Philipp Renz, Dries Van Rompaey, Jörg Kurt Wegner, Sepp Hochreiter, and Günter Klambauer. On failure modes in molecule generation and optimization. *Drug Discovery Today: Technologies*, 32:55–63, 2019.

Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *International conference on machine learning*, pp. 9323–9332. PMLR, 2021.

Arne Schneuing, Yuanqi Du, Charles Harris, Arian Jamasb, Ilia Igashov, Weitao Du, Tom Blundell, Pietro Lió, Carla Gomes, Max Welling, et al. Structure-based drug design with equivariant diffusion models. *arXiv preprint arXiv:2210.13695*, 2022.

Xiaoqin Tan, Chunpu Li, Ruirui Yang, Sen Zhao, Fei Li, Xutong Li, Lifan Chen, Xiaozhe Wan, Xiaohong Liu, Tianbiao Yang, et al. Discovery of pyrazolo [3, 4-d] pyridazinone derivatives as selective ddr1 inhibitors via deep learning based design, synthesis, and biological evaluation. *Journal of Medicinal Chemistry*, 65(1):103–119, 2021.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Jiachen Wen, Siyuan Wang, Rongxian Guo, and Dan Liu. Csf1r inhibitors are emerging immunotherapeutic drugs for cancer treatment. *European Journal of Medicinal Chemistry*, 245: 114884, 2023.

Chao Zhang, Prabha N Ibrahim, Jiazhong Zhang, Elizabeth A Burton, Gaston Habets, Ying Zhang, Ben Powell, Brian L West, Bernice Matusow, Garson Tsang, et al. Design and pharmacology of a highly specific dual fms and kit kinase inhibitor. *Proceedings of the National Academy of Sciences*, 110(14):5689–5694, 2013.

A Appendix

A.1 The synthesis of compound 2-7

Scheme 1. The synthesis of compound of 17. Reaction agents and condition: (a) dimethylamine, EGME, 140 °C, 30 hours; (b) POCl3, toluene, reflux, overnight; (c) 4-aminophenol, K2CO3, CuI, 1,10-phenanthroline, DMF, 90 °C, 3 hours; (d) 5-10, DPPA, Et3N, dioxane, 100oC, 3 hours.

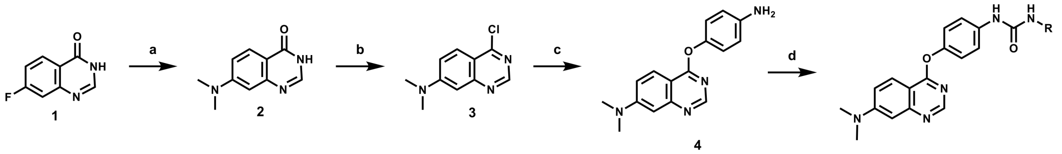


Figure 5: Scheme 1.

A.2 Inhibitory Assay Test

The recombinant protein CSF1R (residues L539-C972, CSR-H5543, Acro, China) containing kinase domain was diluted in kinase buffer (40 mM Tris-HCl, pH 7.4, 20 mM MgCl2, 0.1 mg/mL BSA, 50 μM DTT). The test compound and control were gradient diluted in DMSO. 600 ng CSF1R was incubated for 40 minutes at room temperature with 25 μM ATP, 2.5 mM MnCl2, 200 μg/mL poly (4: 1 Glu, Tyr) peptide (P2075, Merck, German), kinase buffer and inhibitors (10 μM-0.1 nM) in a final volume of 50 μL with 1 % DMSO. Following incubation, 50 μL of kinase activity assay reagent (S0155S, Beyotime, China) was added, and the mixture was incubated for 10minutes at 25 °C. The mixture was added in black Opaque 96-well Microplate (3915, Corning, USA) and the luminescence was measured on SorftMax Pro 7.1.2. Dose-reaction curve was fitted by Prism 9.0 (GraphPad Software Inc, USA).

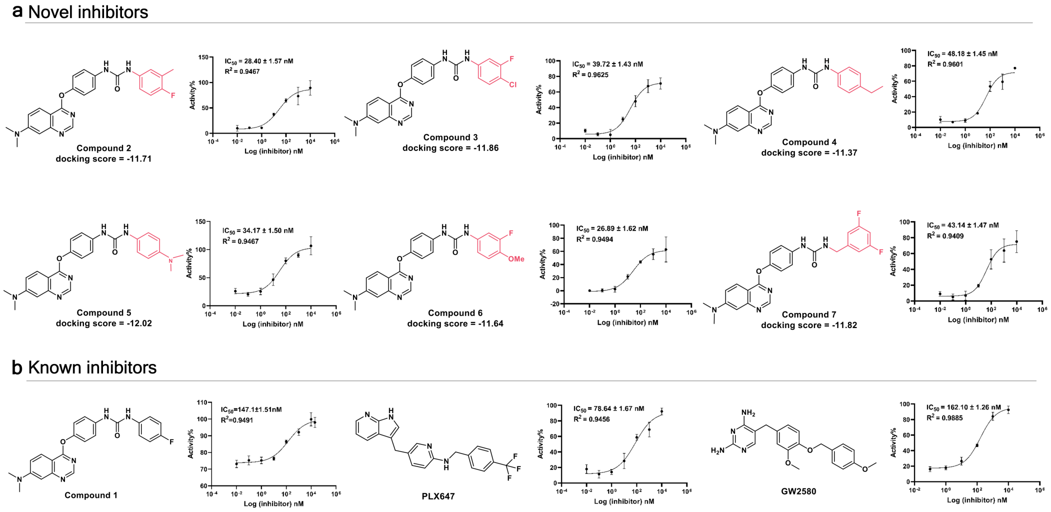


Figure 6: Structures of Compound 2-7