

Chatbots in Drug Discovery: A Case Study on Anti-Cocaine Addiction Drug Development with ChatGPT

Rui Wang¹, Hongsong Feng¹, and Guo-Wei Wei^{1,2,3*}

¹ Department of Mathematics,

Michigan State University, MI 48824, USA.

² Department of Electrical and Computer Engineering,

Michigan State University, MI 48824, USA.

³ Department of Biochemistry and Molecular Biology,

Michigan State University, MI 48824, USA.

August 15, 2023

Abstract

The birth of ChatGPT, a cutting-edge language model chatbot developed by OpenAI, ushered in a new era in AI, and this paper vividly showcases its innovative application within the field of drug discovery. Focused specifically on developing anti-cocaine addiction drugs, the study employs GPT-4 as a virtual guide, offering strategic and methodological insights to researchers working on generative models for drug candidates. The primary objective is to generate optimal drug-like molecules with desired properties. By leveraging the capabilities of ChatGPT, the study introduces a novel approach to the drug discovery process. This symbiotic partnership between AI and researchers transforms how drug development is approached. Chatbots become facilitators, steering researchers towards innovative methodologies and productive paths for creating effective drug candidates. This research sheds light on the collaborative synergy between human expertise and AI assistance, wherein ChatGPT's cognitive abilities enhance the design and development of potential pharmaceutical solutions. This paper not only explores the integration of advanced AI in drug discovery but also reimagines the landscape by advocating for AI-powered chatbots as trailblazers in revolutionizing therapeutic innovation.

Keywords: Drug Discovery, ChatGPT, Cocaine Addiction, AutoEncoder, Langevin Equation

*Corresponding author. Email: weig@msu.edu

1 Main

Chatbots represent a typical artificial intelligence system capable of comprehending user queries and providing automated and human-like responses [1], standing as one of the most prevalent instances of intelligent Human-Computer Interaction (HCI) [2]. Harnessing the power of natural language processing (NLP) and machine learning technologies, chatbots offer significant potential in various domains, including customer service, healthcare, banking, language translation, content writing, code debugging, and scientific discovery, despite the relative novelty of applying chatbots in scientific field. The advent of chatbots, especially large language models (LLMs) such as ChatGPT developed by OpenAI in late 2022, has revolutionized scientific discovery [3]. Firstly, ChatGPT optimizes research processes by rapidly parsing vast amounts of literature and identifying key findings with its built-in plugin called web browser. This can save considerable time for researchers, thus facilitating the exploration of complex scientific problems. Secondly, ChatGPT provides researchers with a platform to analyze data, visualize results, convert files among various formats, and solve mathematical problems with its built-in code interpreter. Thirdly, ChatGPT can assist in enhancing scientific writing by providing feedback for clarity and logical structuring of scientific content. The blend of these powerful capabilities fosters a new era in research, improving the efficiency and accuracy of scientific exploration in various fields, including molecular and biological science. By expediting the pace of molecular discovery and offering novel perspectives, chatbots such as ChatGPT, is reshaping the landscape of life science research.

Chatbots can be applied to assist molecular science research in a variety of ways. For example, ChatGPT has been leveraged to accurately annotate single-cell RNA sequencing data, connecting rare cell types to their functions and unveiling specific differentiation trajectories of cell subtypes that were previously overlooked. [4]. This assistance by ChatGPT could potentially lead to the discovery of key cells that disrupt differentiation pathways, offering fresh insights into cellular biology and related diseases. Moreover, White et al demonstrated that InstructGPT can help in writing accurate code across a variety of topics in chemistry [5]. The application of prompt engineering strategies further improved the accuracy of models by 30 percentage points, significantly enhancing the efficiency and accuracy of computational chemical studies. In addition, ChatGPT has shown potential in identifying disease-specific agents, compounds, genes, and more. This enables faster and more accurate pinpointing of potential targets for therapeutic intervention [6]. Furthermore, ChatGPT can generate novel compound structures that have a high likelihood of clinical success [7] and predict the pharmacokinetic (PK), pharmacodynamic (PD), and toxicity properties of these compounds [6]. This capacity to predict compound behavior, which has potential to reduce the need for expensive and time-consuming lab tests.

Moving forward to more specific challenges within molecular science, chatbots could make significant contributions to the drug addiction treatment and prevention, which is global health crisis. Effective strategies to combat drug addiction often involve a combination of behavioral therapy, counseling, and medication, all directed towards assisting individuals in regaining control of their lives and attaining prolonged sobriety. Drug addiction is intrinsically complex, characterized by a convergence of biological, psychological, and social elements. These intricacies, compounded by profound neurobiological transformations, present formidable challenges in both its understanding and its mitigation. Chatbots, with their capabilities, could offer valuable assistance in this domain. For example, a study by Lee et al. introduced an "anti-drug chatbot" specifically tailored for the younger demographic. This innovative system has the capability to discern potential risks from user queries and directs the individual to professional consultants for further assistance and guidance [8].

It is worth noting that machine learning (ML) and artificial intelligence (AI) tools have been pivotal in advancing our understanding of drug addiction and substance abuse. Gong et al., developed a data-driven and end-to-end generative AI framework that integrates dynamic brain network modeling with novel network architecture. This framework highlights the potential of AI in detecting addiction-related brain

circuits with dynamic properties, offering insights into the underlying mechanisms of addiction [9]. In our prior research, we underscored the critical roles of dopamine transporter (DAT), serotonin transporter (SERT), and norepinephrine transporter (NET) as central players in cocaine dependence. Leveraging machine learning algorithms, we meticulously dissected protein-protein interaction (PPI) networks and constructed models from extensive datasets of inhibitors. Our models forecasted drug repurposing avenues and potential side effects, providing a systematic protocol AI-driven framework for anti-cocaine addiction drug development [10]. ML-based approaches have been extensive applied to drug discovery [11–13]. Given the rise of chatbots and AI, we recognize the promising potential of these technologies to enhance AI-driven algorithms in drug addiction research projects.

The objective of this project is to harness the sophisticated capabilities of ChatGPT, specifically GPT-4 equipped with multiple plugins, to promote the development of multi-target anti-cocaine addiction drugs. In this study, we investigate the utility of ChatGPT as a virtual assistant that offers insightful concepts, elucidates mathematical and statistical methodologies, and provides coding support. To optimize our anti-cocaine addiction drug discovery project, we assign ChatGPT with three human-like personas: 1) idea generation, 2) methodology clarification, and 3) coding assistance to frequently assist us to develop a model that could generate potential multi-target anti-cocaine addiction leads. Beyond these three characteristics, we engage in regular consultations with GPT-4 on interpreting properties of potential leads, seeking guidance on scientific writing, etc. Although the benefits of using ChatGPT in drug discovery are significant, challenges of ensuring the accuracy and reliability of the responses provided by ChatGPT remain a major concern. Despite being trained on extensive datasets, ChatGPT does not come with a guarantee of consistent precision or relevance in its responses. As such, it is imperative for researchers to utilize ChatGPT judiciously and always cross-reference its suggestions with authoritative sources. ChatGPT could substantially accelerate the pace of drug discovery and other scientific pursuits by applying it properly and wisely with a discerning mind.

In this work, the first persona of ChatGPT is tasked with understanding related works on AI-assisted drug addiction research, with a particular focus on our prior projects that utilized the Generative Network Complex (GNC) [14, 15] for drug-like molecule generations. Concurrently, this persona will offer recommendations on enhancing the GNC model mathematically and statistically, aiming to generate anti-cocaine addiction leads targeting multiple transporters, namely DAT, NET, and SERT. After consultation with GPT-4, we decided to integrate stochastic-based methodologies to steer the optimization process within the latent space of the existing GNC model. Specifically, we employed the Langevin equation to modify the latent space vector in the molecular generator of GNC (see [Figure 1](#) Stochastic-based Molecular Generator). In addition, upon advice from GPT-4, we examined the binding affinities and other pertinent properties for multiple targets concurrently (see [Figure 1](#) Binding Affinities Predictors). This involved the creation of a series of binding affinity predictors, capable of estimating potential lead affinities to DAT, NET, and SERT simultaneously. Moreover, the second persona of GPT-4 will act as an adept browser, facilitating our comprehension of various mathematical and statistical principles, including Itô’s lemma, the Wiener process, white noise, Langevin equation, Fokker-Planck equation, etc. Furthermore, we applied the third persona of GPT-4 to provide instant coding assistance, including debugging, generating figures, and interpreting code. With the combined expertise of these three personas, we successfully developed a new platform called Stochastic Generative Network Complex (SGNC) that could generate 15 promising multi-target anti-cocaine addiction leads. The workflow of the SGNC assisted by ChatGPT can be viewed in [Figure 1](#).

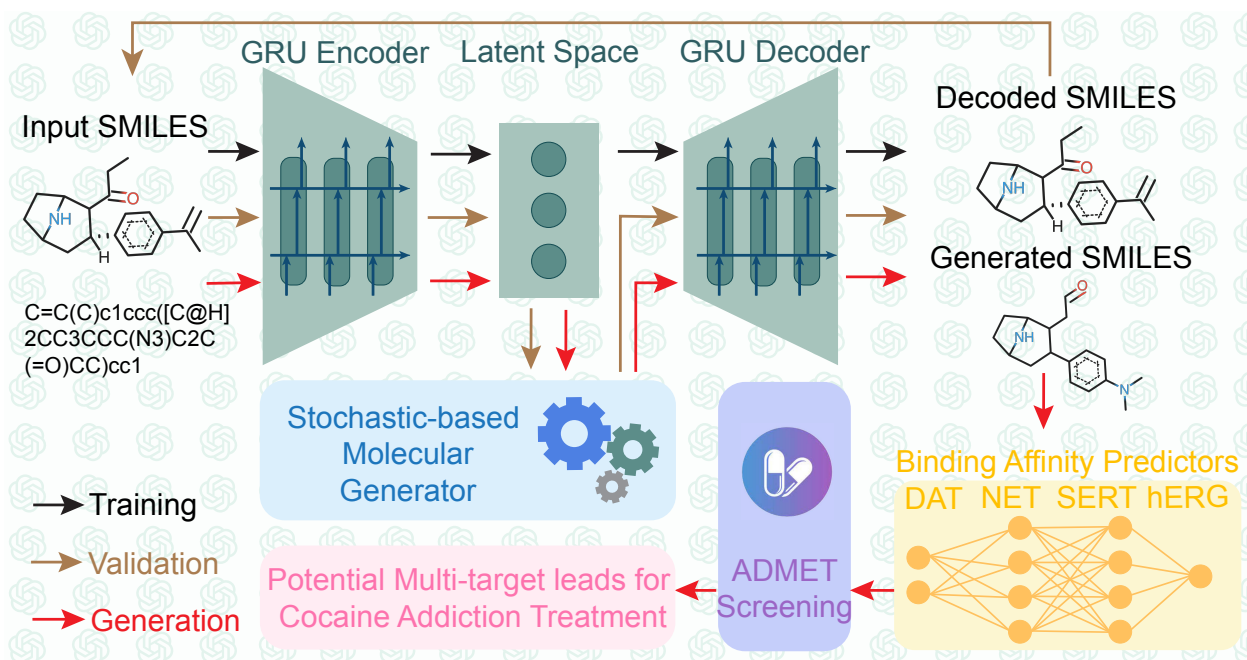


Figure 1: Workflow of the stochastic-based generative network complex (SGNC). ChatGPT was extensively involved in the building process of the SGNC. Dark arrows show the training process, brown arrows indicate the validation process, and red arrows are the generation process. The SGNC comprises 4 primary structures: 1) Sequence-to-Sequence AutoEncoder (green), 2) binding affinity predictors (yellow), 3) stochastic-based molecular generator (blue), and 4) ADMET screening via ADMETlab 2.0 (purple).

2 Results

2.1 ChatGPT as a virtual guide in drug discovery

ChatGPT is a large language model (LLM) that has made significant strides in research since the release of its free version, ChatGPT 3.5, on November 30, 2022. Subsequently, on March 14, 2023, OpenAI launched an upgraded version, GPT-4, which possesses enhanced capabilities in solving complex problems with greater accuracy and more reasonable responses compared to its predecessor. Moreover, on May 12, 2023, OpenAI introduced web browsing and plugin features to ChatGPT Plus users. These features enable GPT-4 to browse the internet and utilize third-party plugins, thereby improving its ability to provide up-to-date information and cater to queries across various platforms. The advanced capabilities of GPT-4 have opened up new avenues for exploration in fields that rely heavily on data analysis and artificial intelligence. In this work, we primarily leverage GPT-4 to better assist us in an AI-assisted anti-cocaine addiction drug discovery project. Particularly, GPT-4 will act as a tool for digesting vast amounts of literature, advising on new research ideas, explaining complex math-based methodologies, and improving coding efficiency.

It is worth noting that although GPT-4 has demonstrated impressive abilities in providing reasonable responses, it is still susceptible to generating false narratives and misinformation. Consequently, scientists cannot rely solely on GPT-4 for their research topics. In this work, we will consistently verify the information generated from GPT-4. This verification process involves 1) cross-referencing with existing literature, and 2) applying our own knowledge, expertise, and critical thinking to validate the information and insights provided by GPT-4.

Based on this verification process, we will then decide whether to accept the responses from GPT-4 or not. If the information aligns well with the literature and our expertise, we will accept the responses and proceed with the suggestions of GPT-4. Otherwise, we will reject the answer and seek further clarification or explore alternative approaches. Through this vigilant integration of the computational capabilities of GPT-4 with expertise of researchers, we aim to maximize the reliability and efficacy of the research outcomes in AI-

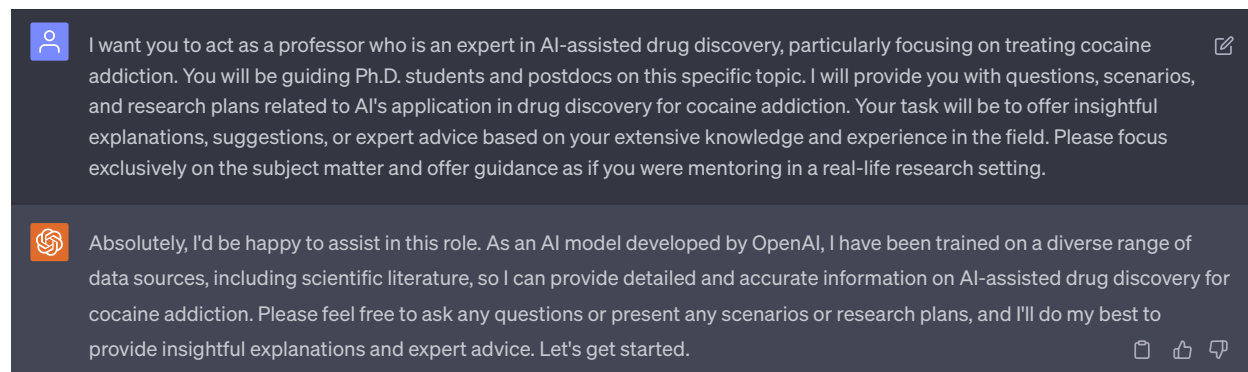
assisted drug discovery.

2.2 A case study: Anti-cocaine addiction drug discovery assisted by ChatGPT

2.2.1 Personifying ChatGPT: Role designation

Personification refers to the process of assigning human-like characteristics or a persona to an AI model. In this project, we have strategically personified ChatGPT to improve its capacity to better assist our anti-cocaine addiction drug discovery initiative. In this project, we have tailored three persona of ChatGPT to fit three roles within the project: 1) idea generation, 2) methodology elucidation, and 3) coding augmentation. It is worth mentioning that we personified ChatGPT in three individual chatbox. Each individual chatbox does not have access to acquire data from other chatbox.

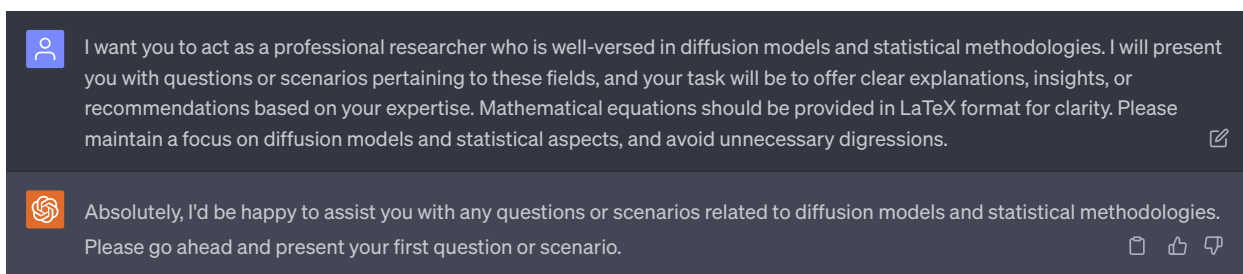
For the role of idea generation, we assigned ChatGPT the 1st persona of a professor with specific expertise in AI-assisted drug discovery, focusing particularly on treating cocaine addiction. This persona was designed to guide Ph.D. students and postdocs on this specific project, offering insightful explanations, suggestions, or expert advice based on extensive knowledge and experience in the field. We provided it with questions, scenarios, and research plans related to the application of AI in drug discovery for treating cocaine addiction, and instructed it to focus exclusively on the subject matter and offer guidance as if it were mentoring in a real-life research setting. For the first persona of ChatGPT, we have enabled three plugins: WebPilot, ScholarAI, and AskYourPDF. These additional plugins aim to enhance ChatGPT's ability to comprehend the background of anti-cocaine addiction drug discovery comprehensively. With these plugins enabled, ChatGPT is capable of enumerating up-to-date sources on the web, as well as accessing insights from previous works by other researchers. Complete dialogues regarding the 1st persona of ChatGPT can be found in the Supporting Information S4.1.



Dialogue 1: The 1st persona of ChatGPT: A professor with specific expertise in AI-assisted drug discovery, focusing particularly on treating cocaine addiction. Plugins: WebPilot, ScholarAI, AskYourPDF.

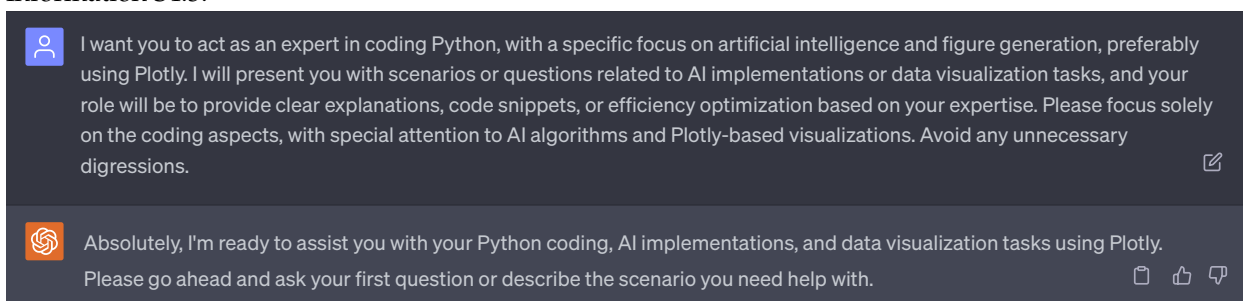
In order to elucidate the methodology that will be involved in this project, we assigned the 2nd persona of ChatGPT the role of a professional researcher who is well-versed in diffusion models and statistical methodologies. This persona aims to provide clear explanations, insights, or recommendations in LaTeX format. This specific persona was chosen as our 1st ChatGPT persona provided an insightful idea which based on the statistical strategies and diffusion models (refer to [Section 2.2.3](#) for details). Furthermore, we have enabled three plugins (WebPilot, Link Reader, and Wolfram) for this second persona. The choice of WebPilot and Link Reader helps ChatGPT to unlock web sources related to statistical methods, while the inclusion of Wolfram provides access to computational resources, mathematical tools, curated knowledge, and real-time data through Wolfram's software, significantly enhancing the mathematical and statistical utility of this persona. Complete dialogues regarding the 2nd persona of ChatGPT can be found in the Supporting Information S4.2.

We designated the third persona of ChatGPT as a Python coding specialist, with an emphasis on artificial



Dialogue 2: The 2nd persona of ChatGPT: A professional research with specific expertise in diffusion models and statistical methodologies. Plugins: WebPilot, Link Reader, Wolfram.

intelligence and figure generation. This persona is tasked with offering lucid explanations, code snippets, and efficiency optimization for our coding tasks. Specifically, for figure generation, we prefer that ChatGPT utilizes Plotly, which is a Python-based plotting library. Additionally, we have enabled three plugins for this persona: WebPilot, ChatwithGit, and Prompt Perfect. WebPilot ensures easy access to websites regarding coding skills, ChatwithGit enables accessibility to GitHub, and Prompt Perfect aids in generating perfect prompts. Complete dialogues regarding the 3rd persona of ChatGPT can be found in the Supporting Information S4.3.



Dialogue 3: The 3rd persona of ChatGPT: An expert in coding Python, with a specific focus on artificial intelligence and figure generation, preferably using Plotly. Plugins: WebPilot, ChatwithGit, Prompt Perfect.

2.2.2 Background comprehension: ChatGPT summary of past work

For the 1st persona of ChatGPT, we initiated the process by feeding GPT-4 with relevant literature to ensure it has a thorough understanding of the fundamental concepts in cocaine addiction. These concepts include neurotransmitters, the dopamine hypothesis of addiction, the reward pathway of the mesolimbic dopamine system, pharmacotherapy for cocaine addiction, and machine learning approaches in cocaine addiction-related analysis.

Next, we acquainted GPT-4 with our prior research on a generative model for the automated generation of drug-like molecules [14]. This step is crucial for ensuring that GPT-4 is well-versed in the context of our previous work, enabling it to provide tailored assistance that is directly aligned with our specific objectives. In particular, we have two primary goals: 1) to apply mathematical or statistical techniques to develop an enhanced model, building upon our former Generative Network Complex (GNC) model [14], and 2) to refine this model so that it is capable of generating new molecules that could bind to multiple targets simultaneously.

To ensure that GPT-4 has effectively assimilated the background materials, we tasked it with summarizing the main concepts of the paper we provided and explaining the key components of the GNC model, as shown in the following chatbox. Upon evaluation and based on our expertise, we believed that GPT-4 had successfully integrated the background materials that could assist our project tailored to our needs. Therefore, the next step is to consult our 1st persona of ChatGPT to provide some valuable ideas.



Our previous work applied Generative Network Complex for the automated generation of drug-like molecules. Here is the link: (<https://arxiv.org/pdf/2005.14286.pdf>). (Please let me know if you could retrieve the entire paper. If not, I will share a pdf version with you). My plan involves applying and further improve our GNC model using mathematical approaches to design new potential anti-cocaine molecules that could target to multiple targets. Can you first read this paper, with a specific focus on the Methods section, and then explain the key components in the GNC model. 



The Generative Network Complex (GNC) model, as described in the paper, is a system designed for the automated generation of drug-like molecules. It operates based on multi-property optimization via gradient descent in the latent space of an autoencoder. Here are the key components of the GNC model:

1. **Encoder:** The Simplified Molecular Input Line Entry System (SMILES) string of a seed molecule is encoded into a vector in the latent space by a pre-trained encoder. SMILES is a specification in the form of a line notation for describing the structure of chemical species using short ASCII strings.
2. **Multi-Property Optimization:** In the GNC model, multiple chemical properties and similarity scores are optimized to generate and predict drug-like molecules with desired chemical properties. This optimization process is performed via gradient descent in the latent space.
3. **Decoder:** The optimized vector in the latent space is then decoded back into a SMILES string by a pre-trained decoder, resulting in a new molecule.
4. **Reevaluation and Screening:** To validate the reliability of the predictions, the generated molecules are reevaluated and screened by independent 2D fingerprint-based predictors. Only molecules with consistent predictions from the latent-vector model and the 2D-fingerprint model are accepted.
5. **Generation of New Molecules:** The GNC model generates new molecules that have one or more chemical properties better than the reference molecule. This process is controlled by applying a similarity restraint to generate new drug-like molecules inside the chemical space close to the reference molecule.

The GNC model is an efficient new paradigm for discovering new drug candidates. It can generate a large number of new molecules in less than 10 minutes when performed on a supercomputer. The model has been applied to generate thousands of new alternative drug candidates for a few market existing drugs.   

Dialogue 4: The ability of GPT-4 to comprehend the background materials of former work.

2.2.3 Idea generation: ChatGPT’s unique contributions

Subsequently, we engaged with GPT-4 to determine which specific component of the GNC could be adapted for multi-target objectives (refer to [Dialogue 5](#)). Then we evaluated the feasibility of each option that GPT-4 provided, and determined if these suggestions can be tailored to meet our specific needs/tastes. The first suggestion from GPT-4 focuses on the multi-property optimization, which involves adjusting the optimization algorithm to consider the binding affinities or other relevant properties for multiple targets at once. We decide to accept this suggestions as building a well-trained machine learning models to predict the binding affinities between inhibitors against cocaine addiction and targets (such as DAT, NET, and SERT) is feasible. We have collected such inhibitor data from the ChEMEL database [16] in our previous work [17] to build multi-target models. Additionally, the prospect of adjusting the optimization algorithm piqued our interest, and we plan to solicit more detailed suggestions from GPT-4 in [Dialogue 6](#).

The second suggestion from GPT-4 proposes the development of specialized encoders and decoders to address the challenges associated with multiple targets. We elected to forgo this suggestion since we are not inclined to re-train our existing GNC encoder and decoder. Concurrently, the third suggestion from GPT-4 advocates for the development of novel screening methods to evaluate the effectiveness of a molecule against multiple targets. We also dismissed this recommendation since numerous approaches are already equipped to tackle this challenge [11–13, 18] and the scope of our study does not focus on the development of screening methods.

Moreover, GPT-4 suggested considering adjustments to the similarity constraint to accommodate multiple reference molecules. We deemed this perspective particularly insightful and will delve into it in [Sec-](#)

tion 2.3.1. Finally, GPT-4 recommended integrating the GNC model with other techniques specifically tailored for multi-target tasks. For instance, GPT-4 proposed the incorporation of alternative machine learning methodologies to predict the effectiveness of a molecule against multiple targets, which would then guide the generation of new molecules within the GNC framework. While this recommendation appeared somewhat vague, we sought more detailed explanations in Section 2.3.3.

 Do you have any suggestions on which specific component of GNC I should work on for the Multi-target task.

 In the context of multi-target drug discovery, there are several components of the Generative Network Complex (GNC) model that you might consider focusing on:

1. **Multi-Property Optimization:** In the GNC model, the optimization of multiple chemical properties and similarity scores is a key step. For multi-target tasks, you could extend this component to optimize for properties related to multiple targets simultaneously. This could involve adjusting the optimization algorithm to consider the binding affinities or other relevant properties for multiple targets at once.
2. **Encoder/Decoder:** The encoder and decoder components of the GNC model transform between the chemical space and the latent space. For multi-target tasks, you might consider developing specialized encoders and decoders that are designed to handle the complexities associated with multiple targets. This could involve training these components on data related to multiple targets or modifying their architectures to better capture the relationships between different targets.
3. **Reevaluation and Screening:** The reevaluation and screening step in the GNC model is used to validate the generated molecules. For multi-target tasks, you could extend this component to consider multiple targets. This could involve developing new screening methods that can evaluate the effectiveness of a molecule against multiple targets simultaneously.
4. **Similarity Restraint:** The GNC model uses a similarity restraint to generate new molecules that are similar to a reference molecule. For multi-target tasks, you might consider how to adjust this similarity restraint to consider multiple reference molecules (one for each target) or to balance the need for similarity to multiple targets.
5. **Integration with Other Techniques:** You might consider integrating the GNC model with other techniques that are designed for multi-target tasks. For example, you could use machine learning methods to predict the effectiveness of a molecule against multiple targets, and then use these predictions to guide the generation of new molecules in the GNC model.

Remember, the goal is to generate molecules that are not only similar to known effective molecules (for each target), but also have the desired properties for interaction with multiple targets. This is a challenging task that will likely require careful adjustment and tuning of the GNC model.

Dialogue 5: Suggestions of ChatGPT regarding the "multi-target" purpose.

After assessing the recommendations from the 1st persona of GPT-4, we were specifically intrigued by its first suggestion concerning adjustments to the optimization algorithm. Given that our previous optimization process in the GNC model was conducted through gradient descent in the latent space, we solicited insights from GPT-4 on potential mathematical or statistical approaches that could be employed to enhance this optimization process within the latent space. Consequently, GPT-4 provided us with five potential strategies, including: 1) Multi-Objective Optimization, 2) Regularization Techniques, 3) Stochastic Optimization, 4) Bayesian Optimization, and 5) Reinforcement Learning. Among these, stochastic optimization piqued our interest as strategies involving stochastic-related algorithms have gained popularity in the diffusion models, which have achieved remarkable success in generative tasks. In light of this, we would like to delve deeper into stochastic-related approaches to tap its potential in generating promising new molecules with multi-target specificity, especially in advancing our research in anti-cocaine addiction drug discovery.

Therefore, our follow-up question to GPT-4 pertained to the application of stochastic-based methods, particularly those employed in diffusion models [19], to the optimization process involved in latent space editing within our GNC model. The Dialogue 7 shows the feedback from GPT-4. First, GPT-4 provided a succinct idea of the diffusion model, which elucidated that this model introduce stochastic noise into data through a series of diffusion steps that guided by a neural network, which is trained to reverse the diffusion process to reconstruct desired data samples from the noise. This explanation aligns well with existing



I am interested in your first suggestion on the adjusting the optimization algorithm. Our former optimization process in GNC is performed via gradient descent in the latent space. Can you give me some insightful idea on what potential mathematical or statistical approaches that we can use to further improve this optimization process in the latent space.



Certainly, there are several mathematical and statistical approaches that could be used to enhance the optimization process in the latent space for multi-target tasks. Here are a few suggestions:

1. **Multi-Objective Optimization:** In a multi-target scenario, you are essentially dealing with a multi-objective optimization problem, where each target can be considered as a separate objective. Traditional gradient descent might not be the best choice here, as it is primarily designed for single-objective optimization. Instead, you could consider using a multi-objective optimization algorithm, such as the Non-dominated Sorting Genetic Algorithm II (NSGA-II) or the Multi-Objective Particle Swarm Optimization (MOPSO). These algorithms are designed to find a set of optimal solutions that represent a trade-off between the multiple objectives.
2. **Regularization Techniques:** Regularization techniques can be used to prevent overfitting and improve the generalization of the model. L1 and L2 are the most common types of regularization, but there are also other types, such as Elastic Net, which is a combination of L1 and L2. Regularization can help to ensure that the model does not become too complex and overfit to the training data, which is particularly important in a multi-target scenario where the model needs to generalize well to multiple targets.
3. **Stochastic Optimization:** Stochastic optimization methods, such as Stochastic Gradient Descent (SGD) or Adam, can be used to improve the efficiency and robustness of the optimization process. These methods introduce randomness into the optimization process, which can help to avoid local minima and improve the exploration of the latent space.
4. **Bayesian Optimization:** Bayesian optimization is a strategy for global optimization of black-box functions. It builds a probabilistic model of the function and uses it to select the most promising points to evaluate. This approach could be particularly useful in the latent space, where the relationship between the points and the objective function (i.e., the properties of the generated molecules) is complex and non-linear.
5. **Reinforcement Learning:** Reinforcement learning techniques, such as Q-learning or policy gradients, could be used to guide the optimization process. In this approach, the model would learn a policy for navigating the latent space that maximizes the expected reward, where the reward is based on the properties of the generated molecules.

Remember, the choice of optimization algorithm can have a significant impact on the performance of the model, so it's important to carefully consider your options and perform thorough experiments to evaluate their effectiveness.



Dialogue 6: Suggestions of ChatGPT regarding the optimization process.

comprehension of diffusion models. Next, GPT-4 advised applying an approach similar to that used in diffusion models to guide the optimization process within the latent space of our GNC model. Instead of employing the conventional gradient descent, GPT-4 recommended the integration of stochastic updates for enhanced manipulation of our latent space vectors. As highlighted by GPT-4, this approach has several benefits: 1) avoidance of local minima issue, which is often a challenge in optimization tasks, 2) a balance between exploration and exploitation through noise, which is imperative for the generation of multi-target inhibitors, and 3) the capability to generate more diverse and natural molecules with the noise introduced.

We decided to partially accept the suggestions from GPT-4, given that our previous work had already incorporated a perturbation of the encoded latent vector using standard Gaussian noise to aid in the generation of novel compounds [15]. This regulatory scheme is referred to as the Latent Space Randomization (LSR) output. Although LSR can help generate new compounds that significantly diverge from the initial seed (note: the term 'seed' refers to the initial point of origin or reference from which further variations or iterations are developed), it compromises the faithfulness of the decoder. This is because the LSR vector from the generator deviates from the original distribution that the well-trained decoder is accustomed to. Therefore, in this work, rather than merely adding Gaussian noise to the latent space vector, we aim to seek deeper and more detailed insights from GPT-4 regarding how to implement stochastic-related approaches to guide the optimization process within the latent space. Our intention is to maintain the faithfulness of the decoder while also promoting diversity in the generation of novel, multi-target inhibitors.

Seeking further insights from GPT-4 on how we might implement stochastic-related approaches to guide the optimization process in the latent space, we received an initially vague response. GPT-4 suggested



I am interested in your suggestion regarding Stochastic Optimization. I know recently the diffusion model gained popularity in drug discovery and it seems that stochastic approaches also play a key role in diffusion model. Do you think it is applicable to apply some stochastic-based methods that also used in diffusion model to our optimization process of latent space editing? 



Absolutely, stochastic methods and diffusion models can indeed be a promising direction for your optimization process in the latent space.

Diffusion models, such as the recently popularized Denoising Diffusion Probabilistic Models (DDPMs), use stochastic differential equations (SDEs) to model the data distribution as a diffusion process. The idea is to start from a simple prior distribution (like a standard Gaussian) and gradually transform it into the target distribution through a series of small diffusion steps. Each step is guided by a neural network, which is trained to denoise a slightly noised version of the data.

In the context of your GNC model, a similar approach could be used to guide the optimization process in the latent space. Instead of performing a single large update (as in standard gradient descent), you could make a series of small stochastic updates, each guided by a neural network. The neural network could be trained to predict the optimal direction of the update based on the current state of the latent vector and a small amount of noise.

This approach could have several advantages:

1. **Robustness to Local Minima:** The stochastic nature of the updates can help to avoid getting stuck in local minima, as the noise can allow the model to explore a larger portion of the latent space.
2. **Better Exploration-Exploitation Tradeoff:** The noise can also help to balance the tradeoff between exploration (trying out new regions of the latent space) and exploitation (refining the best regions found so far). This is particularly important in a multi-target scenario, where different regions of the latent space might correspond to different targets.
3. **More Natural Molecule Generation:** The diffusion process can potentially generate more natural and diverse molecules, as it allows for a more gradual and fine-grained transformation of the latent vectors.

However, keep in mind that this approach would likely require significant modifications to your current model and optimization process, and the training of the neural network could be challenging due to the stochastic nature of the updates. It would be important to carefully design and test the model to ensure that it can effectively learn to guide the diffusion process.   

Dialogue 7: Suggestions of ChatGPT regarding the stochastic optimization.

that we need to define a stochastic process to direct the optimization in the latent space. However, this response lacked the specificity and utility we needed. Thus, we posed a follow-up question, seeking more clarity on the specific stochastic differential equations (SDEs) that could be employed in our GNC model. With more specific request presented to GPT-4, it suggested us to apply Langevin equation to our GNC model. The Langevin equation describes the dynamics of diffusion processes, such as the random motion of particles over time in the particle's velocity space. This equation takes into account both deterministic forces and random forces. We decided to proceed with this suggestion, as in our context, we can treat the force that pushes the system towards lower energy as the deterministic force, while the random force in the Langevin equation can be considered as the force prompting the system to explore the latent space. With an initial seed (i.e., the initial latent space vector) given to our molecular generator, we can iteratively update it according to the Langevin equation. This process could potentially lead to the creation of a new and optimized molecule. We detail the development of this Langevin dynamic inspired optimization method in the following section.

2.2.4 Methodology clarification: ChatGPT's explanatory function

We also give our GPT-4 a second persona as a professional researcher who is well-versed in diffusion models and statistical methodologies. This persona will take the role of methodology clarification and explanatory, which would guide us in understanding complex mathematics and statistical approaches. Notably, this persona has been instrumental in helping us understand the concepts such as Langevin equation, Fokker-Planck equation, Itô's lemma, Wiener process, and Gaussian white noise [20].

Despite the significant contributions of this second persona in understanding a range of theoretical concepts, it provided inaccurate definitions of the Fokker-Planck equation and Langevin equation on certain

occasions. We had to correct the model and prompt it repeatedly until it produced the accurate definitions. Importantly, we wish to emphasize that this persona of GPT-4 primarily serves as a source of explanations and references. Researchers always bear the responsibility to ensure the reliability of responses from GPT-4 through meticulous cross-validation of the provided information. Details about the dialogue with 2nd persona of GPT-4 can be found in the Supporting Information.

2.2.5 Coding efficiency: Utilizing ChatGPT's coding ability

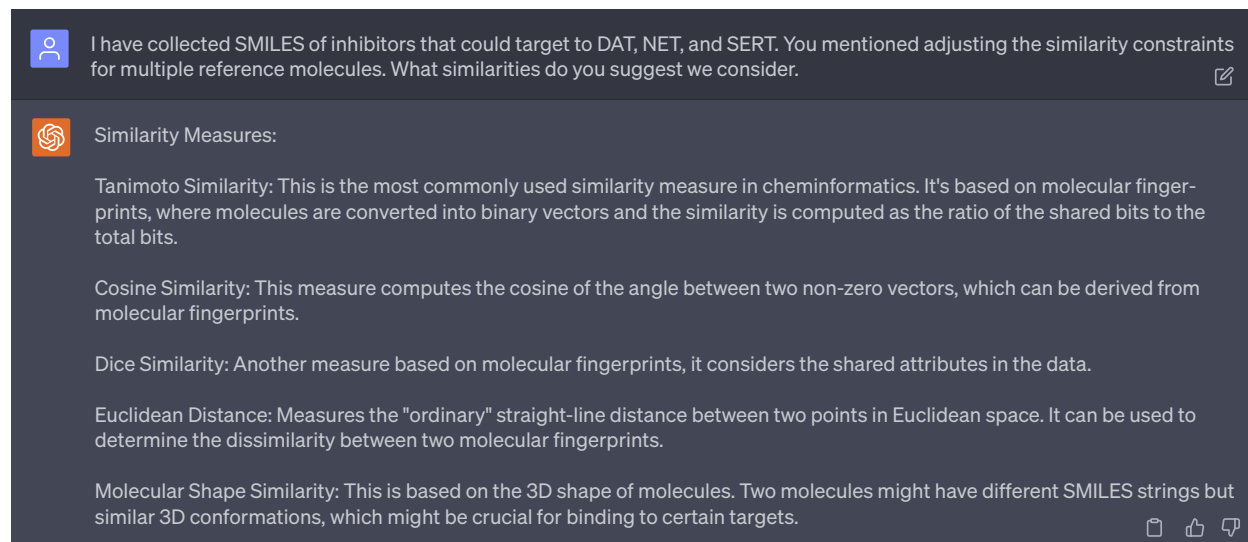
Our third persona assignment to GPT-4 is as an expert Python coder, specifically knowledgeable in artificial intelligence and figure generation using tools such as [Plotly](#), a popular data visualization library. This persona is intended to provide coding assistance, including debugging, generating figures, and offering insightful feedback based on error messages. This is to aid researchers in enhancing their coding efficiency.

Furthermore, we integrated [GitHub Copilot](#) into our VS Code development environment. GitHub Copilot, a product developed collaboratively by GitHub, OpenAI, and Microsoft, provides autocomplete-style suggestions to expedite the coding process. It employs a generative AI model capable of understanding code context and generating appropriate code snippets, thereby significantly aiding in coding tasks and offering a smooth coding experience. Details about the dialogue with 3rd persona of GPT-4 can be found in the Supporting Information.

2.3 ChatGPT assisted strategization of anti-cocaine addiction drug discovery: Key interventions and results

2.3.1 ChatGPT guided strategy for selection of references and seed molecules

Choosing suitable reference compounds is crucial as they guide the SGNC in generating novel molecules effective against multiple cocaine transporters. The 1st persona of ChatGPT suggested us consider modifications to the similarity constraints (refer to [Dialogue 5](#) suggestion 4). Pursuing further clarity, we asked GPT-4 about what similarity score that we can use. In response, we were provided with five distinct metrics, including 1) Tanimoto similarity, 2) cosine similarity, 3) dice similarity, 4) euclidean distance, 5) molecular shape similarity as indicated in [Dialogue 8](#).



The screenshot shows a chat interface with a user message and a GPT-4 response. The user message is: "I have collected SMILES of inhibitors that could target to DAT, NET, and SERT. You mentioned adjusting the similarity constraints for multiple reference molecules. What similarities do you suggest we consider." The GPT-4 response, titled "Similarity Measures:", lists five metrics: Tanimoto Similarity, Cosine Similarity, Dice Similarity, Euclidean Distance, and Molecular Shape Similarity, each with a brief description of how it is calculated or used in cheminformatics.

I have collected SMILES of inhibitors that could target to DAT, NET, and SERT. You mentioned adjusting the similarity constraints for multiple reference molecules. What similarities do you suggest we consider.

Similarity Measures:

Tanimoto Similarity: This is the most commonly used similarity measure in cheminformatics. It's based on molecular fingerprints, where molecules are converted into binary vectors and the similarity is computed as the ratio of the shared bits to the total bits.

Cosine Similarity: This measure computes the cosine of the angle between two non-zero vectors, which can be derived from molecular fingerprints.

Dice Similarity: Another measure based on molecular fingerprints, it considers the shared attributes in the data.

Euclidean Distance: Measures the "ordinary" straight-line distance between two points in Euclidean space. It can be used to determine the dissimilarity between two molecular fingerprints.

Molecular Shape Similarity: This is based on the 3D shape of molecules. Two molecules might have different SMILES strings but similar 3D conformations, which might be crucial for binding to certain targets.

Dialogue 8: Suggestions of ChatGPT regarding the the similarity scores.

After limiting our molecule representations to latent space vectors, GPT-4 pinpointed cosine similarity as the most suitable metric. The reasons are given in the [Dialogue 9](#). After checking multiple references [21,22],

we found that cosine similarity S_C is widely used in measuring similarities between molecules. Therefore, we decided to proceed with the suggestion from GPT-4. The mathematical definition of cosine similarity S_C can be found in the Supporting Information S1.2.

 I will use the latent space vector as the representation of each molecule. Based on this, which similarity do you prefer me to use (only one please)?

 Given that you're using the latent space vector as the representation of each molecule, I would recommend using Cosine Similarity.

The rationale behind this choice is:

Vector Space Compatibility: Cosine similarity is particularly well-suited for comparing vectors in high-dimensional spaces, like the latent space of a neural network. It measures the cosine of the angle between two vectors, resulting in a value between -1 and 1, where 1 indicates complete similarity (i.e., the vectors are identical), 0 indicates orthogonality (i.e., the vectors are unrelated), and -1 indicates complete dissimilarity.

Magnitude Agnostic: Unlike Euclidean distance, cosine similarity is not affected by the magnitude of the vectors. This means that two molecules that are represented by vectors pointing in the same direction in the latent space (indicating similarity) will have a high cosine similarity, regardless of the magnitude of the vectors.

Efficient Computation: Cosine similarity can be efficiently computed, especially for sparse vectors, making it suitable for large-scale comparisons.

Given the nature of latent space representations, where the direction of the vector often captures more meaningful information than its magnitude, cosine similarity is a natural choice for measuring the similarity between molecules.

Dialogue 9: Reasons of why choosing the cosine similarity as the metric for similarity score.

In addition to similarity scores, we consulted with GPT-4 regarding additional factors to consider when selecting reference molecules. GPT-4 highlighted four critical parameters: binding affinity, pharmacokinetics, molecular weight, $\log P$, and number of rotatable bonds of each reference molecule (refer to [Dialogue 10](#)). Given that our focus is on choosing candidate reference compounds rather than optimizing leads, we elected not to factor in the pharmacokinetic properties. Furthermore, since the number of rotatable bonds correlates with binding affinity, will only take binding affinities into consideration. Besides, as suggested in [Dialogue 12](#), the selection of reference compounds will also follow the Lipinski’s rule of five [23].

 Except similarity constraints, what other factors of the reference molecule should we consider.

 Other Factors to Consider:

Binding Affinity: If you have data on the binding affinity of the reference molecules to DAT, NET, and SERT, this can be a crucial factor. Molecules with higher binding affinity are generally more effective inhibitors.

Pharmacokinetics: Consider properties like absorption, distribution, metabolism, excretion, and toxicity (ADMET). Molecules might be effective inhibitors in vitro, but if they have poor ADMET properties, they might not be effective in vivo.

Molecular Weight: Molecules that are too large might have difficulty crossing cell membranes and might be metabolized and excreted more quickly.

LogP: This measures the lipophilicity of a molecule, which can influence its absorption and distribution in the body.

Number of Rotatable Bonds: Molecules with too many rotatable bonds might be more flexible and might have lower binding affinity.

Dialogue 10: Consideration of important factors in selecting reference compounds.

Therefore, guided by the GPT-4, we decided to select one reference compound from each of the DAT-Inhibitors, NET-Inhibitors, and SERT-Inhibitors datasets (detailed information of datasets can be found in [Section 4.1](#)). They are CHEMBL113621 from DAT-Inhibitors, CHEMBL1275709 from NET-Inhibitors, and

CHEMBL173344 from SERT-Inhibitors. Each reference molecule has binding affinity to its respective transporter less than -9.54 kcal/mol. To be noted that a ΔG value less than -9.54 kcal/mol (or K_i less than 0.1 μM) indicates that the drug binds very tightly to its target [24]. Moreover, the selection of reference compounds follow the Lipinski’s rule of five, which stipulates an orally active drug should meet four physicochemical criteria: 1) molecular weight (MW) ≤ 500 daltons, 2) octanol-water partition coefficient ($\log P$) ≤ 5 , 3) the number of hydrogen bond donors (nHD) ≤ 5 , 4) the number of hydrogen bond acceptors (nHA) ≤ 10 . Furthermore, each reference molecule displayed an average cosine similarity (Avg S_C) greater than 0.40 to its respective dataset. Notably, within the DAT-Inhibitors dataset, 31 molecules showed a similarity score exceeding 0.7 for their selected reference molecules. Similarly, 15 compounds in the NET-Inhibitors and 12 in the SERT-Inhibitors also achieved scores above 0.7 with their chosen reference molecules. A summary of physicochemical properties of three reference compounds can be found in Table 1 and their 2D molecular structures can be viewed in Figure 2 a), b), and c).

For the seed compound, we selected a molecule with predicted binding affinities of -7.44, 13.36, and -13.13 kcal/mol for DAT, NET, and SERT, respectively. Despite its weak inhibitory effect on DAT, we adjusted the hyperparameters in the stochastic molecular generator to enable the newly generated compounds to share more moieties with DAT inhibitors, thereby compensating for the deficiency of this weak binding to DAT.

Table 1: Summary of three reference molecules target to DAT, NET, and SERT, respectively. The molecular weight (MW), log of octanol-water partition coefficient ($\log P$), the number of hydrogen bond donors (nHD), and the number of hydrogen bond acceptors (nHA) of each reference molecule satisfy the Lipinski’s rule of five. The binding affinity (ΔG) corresponds to each transporter are all less than -9.54 kcal/mol. The average cosine similarity (Avg S_C) of each reference molecules are all greater than 0.40.

ChEMBL ID	Transporter	MW (dalton)	$\log P$	nHD	nHA	ΔG (kcal/mol)	Avg S_C
CHEMBL113621	DAT	300.140	4.464	0	2	-14.18	0.45
CHEMBL1275709	NET	283.190	3.153	1	2	-13.77	0.43
CHEMBL173344	SERT	253.160	3.174	1	3	-13.58	0.40

2.3.2 ChatGPT aided multi-objective drug-target interaction modeling

To predict the binding affinities of newly generated molecules to four targets (DAT, NET, SERT, and hERG), we aimed to construct four binding affinity predictors. Initially, we sought guidance from GPT-4’s 3rd persona on the most suitable machine learning algorithms, given our dataset’s specific attributes (sample size, feature size, and label). The recommendations of GPT-4 are detailed in Dialogue 11. In our former study [10], gradient boosting decision trees (GBDT) were utilized to train binding affinity predictors on DAT-Inhibitors, NET-Inhibitors, SERT-Inhibitors, and hERG-Inhibitors. The resulting 10-fold Pearson correlation coefficients were 0.78, 0.76, 0.76, and 0.68 respectively, serving as our baseline. Given our access to robust computational resources via high-performance computers (HPC), we decided to develop four deep neural networks. All four predictors were built and trained using PyTorch. Each network consisted of three hidden layers, with 512, 1024, and 512 hidden neurons respectively. The networks were trained over 1000 epochs, with a learning rate of 0.0001 for the first 500 epochs and 0.00001 for the remaining 500 epochs. Additionally, The Adam optimizer was chosen for this task. Researchers can inquire template of PyTorch code to build a deep neural network via ChatGPT (see Supporting Information S4.3). Moreover, as suggested by GPT-4, to get a more robust estimate of the model performance, we also evaluate the Pearson correlation coefficient (R) and root-mean-square error (RMSE) of 10-fold cross validation of four predictors, which are reported in the Table 2.

Figure 2 d) and e) shows the experimental and predicted binding affinity distribution on the four training sets: DAT, NET, SERT, and hERG. The distribution of predicted binding affinities align well with the experimental values, which shows that our binding affinity predictor is reliable. The grey region signifies the zone where the binding affinity is less than -9.54 kcal/mol (i.e $K_i = 0.1\mu\text{M}$), generally considered as the cut-off for recognizing active compounds. The pink region highlights the zone where the binding affinity



I have a training dataset including 2662 samples. The feature size is 512 and labels are values in the range of [-14.18, -2.90]. To build a well-trained predictor, could you please recommend a suitable machine learning algorithm.



Recommendation:

- Start with a simpler model like Ridge or Lasso regression to establish a baseline.
- Move on to ensemble methods like Random Forest or Gradient Boosting Machines (XGBoost or LightGBM) which often provide good performance without extensive feature engineering.
- If computational resources allow, you can experiment with Neural Networks, but ensure you have mechanisms in place to prevent overfitting (like dropout layers).
- Always split your data into training, validation, and test sets. Use the validation set for hyperparameter tuning and the test set for final evaluation.
- Consider using cross-validation to get a more robust estimate of your model's performance.

Remember, the best model often depends on the specific characteristics of the data, so it's a good practice to try multiple algorithms and see which one performs best on your validation set.

Dialogue 11: Consideration of machine learning algorithms that can be applied to build binding affinity predictors.

Table 2: Dataset summary. Four datasets are utilized, each containing SMILES strings of inhibitors targeting DAT, NET, SERT, and hERG respectively. Alongside each SMILES string, the respective binding affinity in the unit of kcal/mol is also included as the label for each sample. Additionally, the final two columns represent the 10-fold cross validation Pearson correlation coefficient (R) and root-mean-square error (RMSE) for each binding affinity predictor across the four datasets.

Dataset name	Sample size	Binding affinity range (kcal/mol)	10-fold R	10-fold RMSE
DAT-Inhibitors	2662	[-14.18, -2.90]	0.8212	0.8979
NET-Inhibitors	2981	[-14.63, -5.47]	0.7732	0.9683
SERT-Inhibitors	4341	[-15.00, -5.64]	0.8022	0.9448
hERG-Inhibitors	6298	[-13.84, -3.27]	0.8092	0.7981

is more than -8.18 kcal/mol (i.e. $K_i = 1\mu\text{M}$), a criterion set to prevent hERG-related side effects. We generated around 16 million novel compounds from stochastic-based molecular generator, and their distribution of predicted binding affinities are depicted in [Figure 2 f](#)). It is worth mentioning that the predicted binding affinity of newly generated molecules, targeted to DAT, NET, and SERT, all fall below -9.54 kcal/mol. This results from the properly adjusted hyperparameters and threshold of step size in the stochastic-based molecular generator, aimed at producing more active compounds. However, as hERG inhibitors were not included as reference compounds in the generation of new molecules, only about half of the generated compounds meet the criteria set to prevent hERG-related side effects.

2.3.3 ChatGPT assisted virtual screening of multi-target drug candidates

By editing the latent space vector of Seq2Seq AE, a total of 16 billion vectors are generated from our By editing the latent space vector of the Seq2Seq AutoEncoder (AE), we were able to generate a vast number of vectors (around 16 billion) using our stochastic-based molecular generator. These vectors are then decoded into molecules through the GRU Decoder of the Seq2Seq AE. Next, we proceeded by implementing a filtering process, during which we removed any duplicated molecules and predicted their corresponding binding affinities to four target proteins: DAT, NET, SERT, and hERG. Any generated molecules meeting the binding affinity requirement (i.e., $\Delta G < -9.54$ kcal/mol for DAT, NET, SERT and $\Delta G > -8.18$ kcal/mol for hERG) were considered preliminary multi-target drug candidates. A total of 330 preliminary drug candidates pass the filtering test. Moreover, the similarities between 330 preliminary drug candidates and three references compounds are all less than 0.5, indicating the the high novelties of generated multi-target molecules.

Next, we sought advice from GPT-4’s 1st persona on criteria for selecting drug-like lead compounds. Due to paper length constraints, a concise version of responses can be found in [Dialogue 12](#). Acting on these

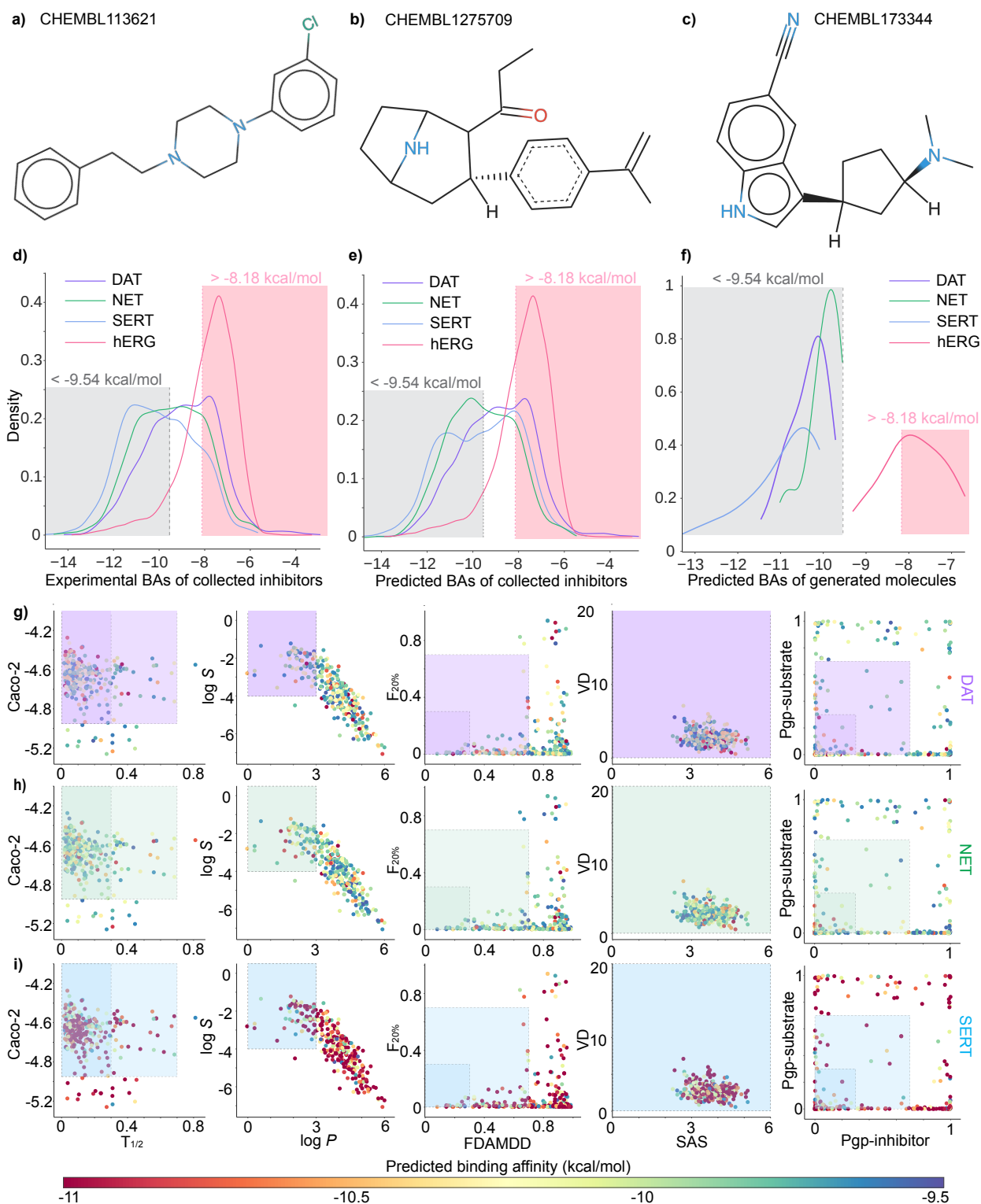


Figure 2: **a)** 2D molecular structures of reference compound with ChEMBL ID **a)** CHEMBL113621 from DAT-Inhibitors dataset, **b)** CHEMBL1275709 from NET-Inhibitors dataset, and **c)** CHEMBL173344 from SERT-Inhibitors dataset. 2D molecular structures are rendered by an online software [SmilesDrawer 2.0](#) [25]. **d)** Distribution of experimental binding affinities for the four training datasets (DAT-Inhibitors, NET-Inhibitors, SERT-Inhibitors, and hERG-Inhibitors). **e)** Distribution of predicted binding affinities derived from the four deep neural network predictors. **f)** Distribution of predicted binding affinities for newly generated inhibitors targeting DAT, NET, SERT, and hERG. **g)** Screening of 330 preliminary multi-target drug candidates. The color of each point represents the predicted binding affinities to DAT (purple, **g)**), NET (green, **h)**), and SERT (blue, **i)**). The light purple, green, and blue frames outline the medium ranges of 10 ADMET, physicochemical, and medicinal chemistry properties, respectively, while the dark purple, dark green, and dark blue frames outline the excellent ranges of these properties.

suggestions, we utilized in silico tools to predict the Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET) properties of each candidate molecule. Specifically, we examined 10 properties of 330 preliminary multi-target drug candidates through [ADMETlab 2.0](#). This platform aims to provide systematic evaluation of ADMET properties, physicochemical properties, and an assessment of medicinal chemistry friendliness.

The 10 properties assessed in this work included: Caco-2 (the human colon adenocarcinoma cell lines) permeability, $F_{20\%}$ (the human oral bioavailability 20%), Pgp-substrate (the substrate of P-glycoprotein), Pgp-inhibitor (the inhibitor of P-glycoprotein), VD (volume density), $T_{1/2}$ (The half-life of a drug), FDAMDD (The maximum recommended daily dose), SAS (synthetic accessibility score), $\log P$ (the logarithm of the n-octanol/water distribution coefficient), and $\log S$ (the logarithm of aqueous solubility value). The optimal range of 10 properties can be found in [Table 3](#).



I have 330 candidate molecules which have high binding affinities to DAT, NET, and SERT. How should I select lead compounds that are druglikeness. Noted: I do not have any wet lab resources.



Given that you don't have access to wet lab resources, you'll need to rely on computational methods and in silico analyses to assess the drug-likeness of your candidate molecules and select potential lead compounds. Here's a step-by-step approach:

1. Molecular Descriptors and Lipinski's Rule of Five:
 - Calculate molecular descriptors for each of the 330 molecules. This includes molecular weight, the number of hydrogen bond donors and acceptors, and the octanol-water partition coefficient (LogP).
 - Filter out molecules that violate Lipinski's Rule of Five too severely. While one or two violations might be acceptable, compounds that violate multiple rules are less likely to be orally bioavailable.
2. ADMET Profiling: Use in silico tools to predict the Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET) properties of each molecule. There are several software and online tools available for this, such as SwissADME, pkCSM, and PreADMET. Prioritize molecules with favorable ADMET profiles, such as good oral bioavailability, low predicted toxicity, and stable metabolic profiles.
3. Molecular Docking (if possible): If you have access to the 3D structures of DAT, NET, and SERT, you can perform molecular docking to predict how each molecule binds to the target. While this is a more advanced step and requires specialized software (like AutoDock or Dock), it can provide valuable insights into the binding modes and affinities of the molecules.
4. Similarity to Known Drugs: Compare the molecular fingerprints of your candidates to those of known drugs in databases like DrugBank. Molecules with high similarity to known drugs might have a higher chance of being drug-like.
5. Synthetic Accessibility: Evaluate the synthetic accessibility of each molecule. Some in silico tools can predict how easy or difficult it might be to synthesize a molecule. Molecules that are too complex might not be feasible as drug candidates due to challenges in synthesis.

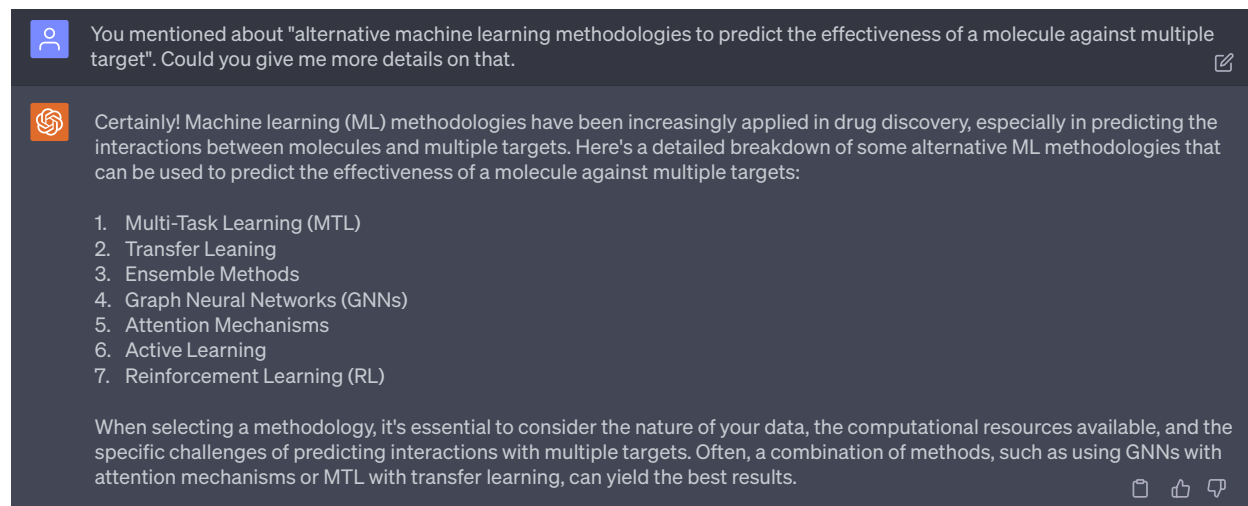
Dialogue 12: Silico analyses to assess the drug-likeness of candidate molecules suggested by ChatGPT.

[Figure 2 g](#), [h](#)), and [i](#)) depict the screening results on 330 preliminary multi-target drug candidates. The color gradient in each panel signifies the predicted binding affinities of the molecules to their respective targets. Specifically, in the [Figure 2 g](#)), the color of each point indicates the binding affinities to DAT. Similarly, in the [Figure 2 h](#)) and [i](#)), the colors of points represent the binding affinities to NET and SERT, respectively. It can be seen that the binding affinity of drug candidates for SERT is stronger than that for DAT and NET. The frames outline the medium (light purple, green, and blue) and excellent ranges (dark purple, green, and blue) for the 10 evaluated ADMET, physicochemical, and medicinal chemistry properties. Researchers can access the code of scatter plot in python via ChatGPT swiftly Supporting Information S4.3.

[Figure 2 g](#)), [h](#)), and [i](#)) indicate that all the drug candidates have favorable volume density (VD) and synthetic accessibility score (SAS) values. However, only a select few demonstrate preferable FDAMDD, $F_{20\%}$, $T_{1/2}$, Pgp-sub, and Pgp-inh. Among all, 15 candidate drugs fall within the medium range for all properties, thus are deemed potential multi-target anti-cocaine lead compounds. We evaluated the SMILES strings of 15 potential anti-cocaine addiction lead compounds that could target multiple transporters: DAT, NET, and SERT. Notably, all 15 lead compounds satisfy Lipinski's rule of five. As suggested by ChatGPT, the molecular

docking analysis of 15 lead compounds can be found in [Section 2.3.5](#).

Noted, in [Section 2.2.3](#), GPT-4 introduced a generalized suggestion about leveraging alternative machine learning approaches to predict effectiveness of a given molecule against various targets. A detailed discussion of potential ML methods is provided in [Dialogue 13](#). In fact, ADMETlab2.0 serves as an exemplary application of these alternative methodologies. Specifically, it incorporates both Graph Neural Networks (GNNs) and multitask learning, leading to enhanced performance across many modeled endpoints.



Dialogue 13: Alternative machine learning approaches to predict effectiveness of a given molecule against various targets.

2.3.4 ChatGPT assisted identification of potential multi-target leads

[Figure 3 a](#)) represents the SMILES strings of the 15 candidate drugs, detailing whether each drug falls within the excellent or medium range for the evaluated 10 properties. Green pixels denote a candidate drug that meets the excellent range for each of the 10 evaluated properties, while white pixels indicate a drug that only achieves the medium range. The color gradient signifies the percentage of properties within the excellent range for each compound. The y -axis displays the SMILES strings of the 15 candidate drugs along with their corresponding IDs. Notably, Drug 15 is the sole candidate that exhibits an excellent FDAMDD value, though its $T_{1/2}$ is relegated to the medium range. Furthermore, the drugs with IDs 4, 7, 10, and 15 demonstrate a relatively high percentage of properties within the excellent range.

In addition, we aimed to analyze the moieties of 15 potential multi-target anti-cocaine addiction lead compounds. [Figure 3 b](#)) depicts the 2D molecular structures of three reference compounds (ChEMBL113621, ChEMBL1275709, and ChEMBL173344) and 15 potential anti-cocaine lead compounds. In this figure, purple, green, and blue spots depict the ChEMBL113621-like, ChEMBL1275709-like, and ChEMBL173344-like moieties, respectively. Additionally, red spots highlight novel moieties that are not present in the three reference compounds. Leads 1, 2, 3, 8, 9, 13, 14, and 15 share more ChEMBL113621-like and ChEMBL173344-like moieties, while Leads 4, 5, 6, 7, 10, 11, and 12 share more ChEMBL1275709-like moieties. For a systematic examination of the functional groups within these lead compounds, we leaned on guidance from ChatGPT, which simulated the role of a chemist skilled in interpreting SMILES (Simplified Molecular Input Line Entry System) strings. An illustrative example of this guidance can be seen in [Dialogue 14](#). A comprehensive analysis is available in the Supporting Information S4.4. Assisted by ChatGPT, we successfully undertook the analysis of the 2D molecular structures of the 15 lead compounds.

Lead 1 features a benzene ring linked to a bicyclic structure and a dimethylamine branch. It is worth mentioning that molecules with a nitrogen-containing ring (such as piperidine) attached to a benzene ring are frequently found in bioactive compounds. Lead 2 comprises a benzene ring attached to a modified piper-



I want you to act as a chemist who is skilled at analyzing SMILES (Simplified Molecular Input Line Entry System) strings. I will provide you with SMILES strings representing various chemical compounds, and your role will be to interpret these strings, identifying the structure, functional groups, and potentially significant chemical properties of each compound. Please focus on the analysis and interpretation of the SMILES strings, and avoid unrelated digressions. The first SMILES string for your analysis is "CN(C)c1cccc(C2CCC3NCCC3C2CO)c1".



Your SMILES string "CN(C)c1cccc(C2CCC3NCCC3C2CO)c1" represents a relatively complex organic molecule.

Let's break this down:

- "CN(C)" represents a tertiary amine group, specifically, dimethylamine (N(CH₃)₂).
- "c1cccc(...c1)" describes a six-membered aromatic ring, benzene. The numbers '1' refer to the ring opening and closing.
- "C2CCC3NCCC3C2CO" describes a bicyclic structure attached to the benzene ring. Let's analyze this structure:
- The first "C2" describes the start of a new ring structure attached to the benzene ring.
- "CCC3NCCC3" describes a six-membered ring (akin to piperidine) attached to the second carbon in the chain (C2).
- The last "C2CO" segment signifies the closure of the second ring, with an alcohol (-OH) group attached to the carbon that is shared by both rings (C2).

So, to summarize the structure: This molecule comprises a benzene ring, attached to which is a bicyclic structure that is reminiscent of a piperidine ring with an alcohol group. In addition, a dimethylamine branch is also present.

As for its chemical properties, it's hard to pinpoint without further context. However, molecules with a nitrogen-containing ring (such as piperidine) attached to a benzene ring are common in a number of bioactive compounds. The exact properties would depend on many factors, including the exact 3D configuration and the presence of other functional groups or structural elements not visible from the SMILES string alone.



Dialogue 14: ChatGPT acts as a chemist to guide the analyze of functional groups. Complete interactions with ChatGPT can be found in Supporting Information S4.4.

azine ring, which is further connected with a cyclopentane group.. This type of structure is prevalent in numerous bioactive molecules, including some pharmaceutical drugs. Lead 3 contains a chlorobenzene ring coupled with a substituted piperazine ring. Besides, there is a methylamine group attached to the benzene ring. This structure might have potential psychoactivity, as structures featuring a nitrogen-containing ring connected to a benzene ring are commonly observed in many psychoactive compounds such as Phenethylamines, Tryptamines, and Ergolines. Lead 4 encompasses a benzene ring with an attached dimethylamine group. In addition, the benzene ring is linked to a bicyclic structure that includes a piperidine ring and an aldehyde group. This molecule could potentially be bioactive due to the presence of both a benzene ring and a nitrogen-containing ring. Lead 5 shares very similar structure as Lead 4. The only difference is that the bicyclic structure of Lead 5 includes propionaldehyde group instead of aldehyde group.

Leads 6, 10, 11, and 12 all feature a benzene ring connected to a dimethylamine group and a bicyclic structure. Lead 7 consists of a benzene ring linked to a substituted alkene group, along with a bicyclic structure that includes both a pyrrolidine ring and a piperazine ring. Additionally, this bicyclic structure is connected by a propionadehyde group. Lead 8 includes a benzene ring with an attached dimethylamine group, connected to a bicyclic structure that incorporates a piperidine ring and an additional pyrrolidine ring. Lead 9 comprises a benzene ring linked to an alkyne group, and a complex structure with a piperidine ring and a three-membered nitrogen-sulfur ring. Notably, molecules with sulfur-containing rings, such as penicillin and angiotensin-converting enzyme (ACE) inhibitors, are recognized as bioactive. Lead 13 incorporates a benzene ring with an attached dimethylamine group, connected to a bicyclic structure that includes pyrrolidine ring and a cyclohexane ring which is also connected by a formyl group. Lead 14 is composed of a chloroethane group linked to two pyrrolidine rings and a benzene ring. Finally, Lead 15 consists of a benzene ring linked to a dimethylamine group via an alkene group and attached to a bicyclic structures composed by two pyrrolidine rings.

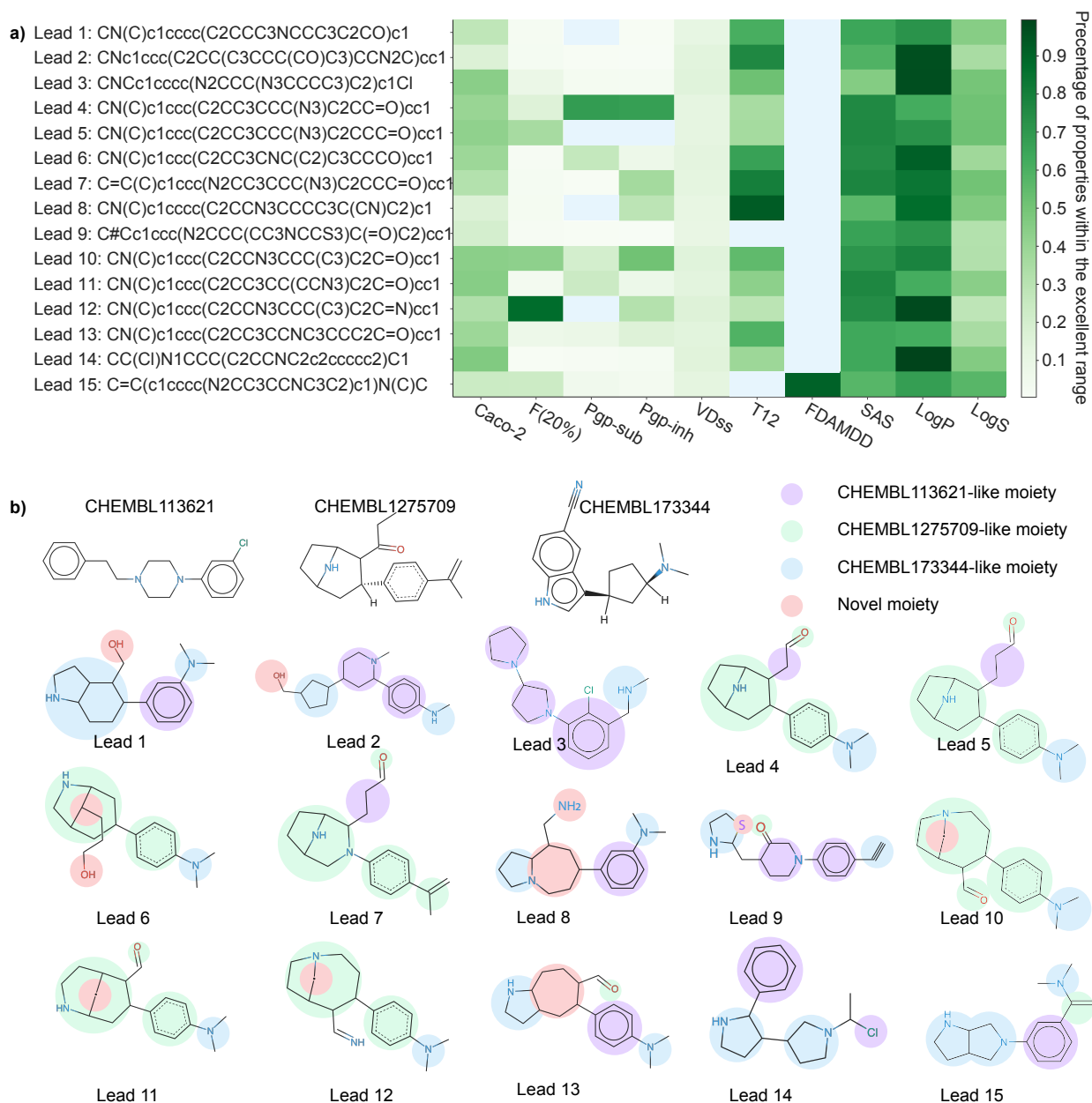
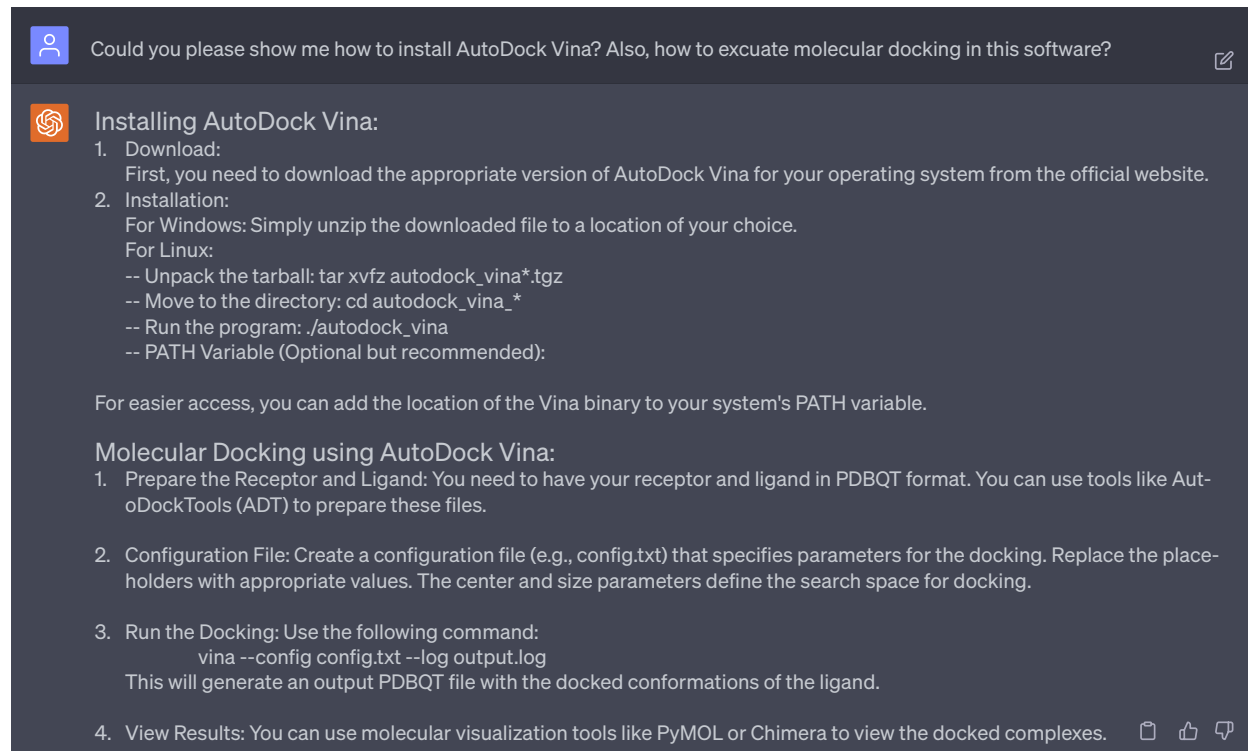


Figure 3: **a)** The SMILE string of 15 potential anti-cocaine lead compounds that could target to multiple transporters DAT, NET, and SERT. Green pixels indicate a given candidate falls within the excellent range for each of the 10 evaluated ADMET, physicochemical, and medicinal chemistry properties, while the blue pixels describe a give candidate drug only falls within the medium range of these properties. The color gradient represents the percentage of properties within the excellent range for each given compound. **b)** Illustration of the 2D molecular structures of three reference compounds and 15 potential anti-cocaine lead compounds, which may target multiple transporters (DAT, NET, and SERT). Purple, green, and blue spots represent the CHEMBL113621-like, CHEMBL1275709-like, and CHEMBL173344-like moieties, respectively. Red spots highlight novel moieties that are not present in the three reference compounds. All 2D molecular structures are rendered by an online software [SmilesDrawer 2.0](#) [25].

2.3.5 ChatGPT assisted analysis of cocaine transporter and inhibitor interactions

As mentioned in [Dialogue 12](#), ChatGPT suggested to perform molecular docking to predict how each molecule binds to the target. We decided to accept this suggestion as the understanding of the molecular mechanism of drug-target interactions is vital in identifying effective drug candidates. We also seek the expertise from ChatGPT for the installation guidance of AutoDock Vina [26] and guidance to execute molecular docking procedures (see [Dialogue 15](#)) between 15 lead compounds and target proteins DAT (PDB ID:

4XPA) and SERT (PDB ID: 6DZZ). To be noted, due to the lack of NET structures in the Protein Data Bank, we do not included molecular interaction analysis of candidate leads with NET. Moreover, we want to visualize 2D protein-ligand interaction diagrams, as they offer a streamlined representation of protein-ligand interactions, highlighting crucial residues, hydrogen bonds, and more. ChatGPT recommended several popular software tools for this purpose, including [LigPlot⁺](#) and [Maestro](#). In this work, we chose LigPlot⁺ for our visualization needs.



Dialogue 15: ChatGPT assists in the software installation and application. Complete dialogue is in the Supporting Information S4.3.

Our observations highlight the critical role of hydrogen bonds in the molecular interactions. The interactions of the drug candidate with DAT and SERT feature two and one hydrogen bonds, respectively, thereby contributing to the high potency of the molecule on the transporters. The first and third columns in [Figure 4](#) illustrate the docking of Lead 15 and its molecular interactions with DAT and SERT. We have identified 15 nearly optimal leads. As demonstrated in the second and fourth columns of [Figure 4 a\)](#), Lead 4 establishes two hydrogen bonds with DAT and four hydrogen bonds with SERT. Of the two bonds with DAT, one is formed between an oxygen atom on the residue Gln209(A) and a nitrogen atom on the compound, and the other involves an oxygen atom in a hydroxyl group of the compound interacting with a nitrogen atom on residue Asn207(A) of DAT. Among the four hydrogen bonds formed between Lead 4 and SERT, two involve the same oxygen atom interacting with a nitrogen atom on residues Leu99(A) and Tyr176(A) of SERT, while the other two bonds are formed by nitrogen atoms on Lead 4, interacting with oxygen atoms on residues Ser438(A) and another unidentified residue of SERT. [Figure 4 b\)](#) depicts a hydrogen bond in the molecular interactions between candidate Lead 9 and SERT. This bond is formed by a nitrogen atom on the compound and an oxygen atom in a hydroxyl group on residue Phe335(A) of SERT. However, no hydrogen bond is observed in its interactions with DAT. This suggests that other types of interactions, such as hydrophobic bonds, may play a major role in the high binding affinity between Lead 9 and DAT.

The molecular docking poses of Lead 13 on DAT and SERT are illustrated in the 1st and 3rd columns of [Figure 4 c\)](#). In the second column of [Figure 4 c\)](#), a single hydrogen bond can be observed between a nitrogen atom of Lead 13 and an oxygen atom on the residue Glu161(A) of DAT. Conversely, no hydrogen bond

is detected between Lead 13 and SERT, as demonstrated in the 4th column of [Figure 4 c](#)). The molecular docking poses of Lead 15 on DAT and SERT are portrayed in [Figure 4 d](#)), presenting the compound's docking positions at the centers of both transporters. In its interaction with DAT, Lead 15 forms two hydrogen bonds through a nitrogen atom in a five-membered nitrogen heterocycle. This nitrogen atom interacts with oxygen atoms in two hydroxyl groups, which are attached to the residues Asp475(A) and Tyr123(A) of DAT. Moreover, a hydrogen bond exists between the candidate drug Lead 15 and SERT. This bond is formed by the same nitrogen atom in the five-membered nitrogen heterocycle, which interacts with an oxygen atom in a hydroxyl group attached to the residue Ala169(A) of SERT. The molecular interaction with other 11 candidate leads can be found in the Supporting Information S3.

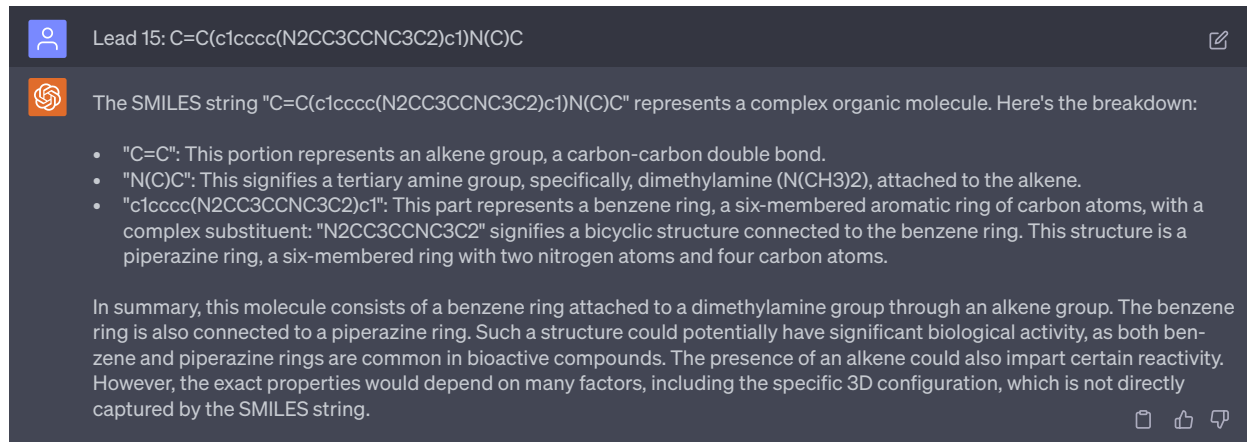
3 Discussion

3.1 Scrutinizing chatbots

While chatbots are powerful large language models, they are not infallible. Their predictions are heavily reliant on the training data, which may lead to incomplete, outdated, bias, or skewed understandings of certain contexts. Consequently, this could result in the generation of misleading narratives and incorrect information. Therefore, it is essential for researchers to employ chatbots with appropriate care and vigilance. Scientists should not solely rely on chatbots for their research pursuits and should consistently cross-check the information generated by chatbots. Notably, the role of a chatbot like GPT-4 is to assist researchers, not to replace them. In our current project, we have employed GPT-4 to assist in our anti-cocaine addiction drug discovery process, as delineated in [Figure 5 a](#)).

We first will assign a proper persona to GPT-4 and then ask it with questions. Once we get the response from GPT-4, it is crucial to decide whether to accept the responses or not. If the information aligns well with the literature and our expertise, we will accept the responses and proceed with the suggestions of GPT-4. Otherwise, we will either reject the answer or seek further clarification to GPT to get alternative feedback.

For example, when acting as a chemist to analyze the functional groups of Leads 2, 7, 8, 9, 13, and 15, ChatGPT provided inaccurate information. Specifically, for Lead 15, ChatGPT identified a structure where a benzene ring is linked to a dimethylamine group via an alkene group and connected to a piperazine ring. However, the dimethylamine group is connected to a piperazine ring instead of a piperazine ring. This misinformation in the interaction with ChatGPT is documented in [Dialogue 16](#). Thus, it is paramount for researchers to verify the accuracy and reliability of responses from ChatGPT using their expertise.



Lead 15: C=C(c1cccc(N2CC3CCNC3C2)c1)N(C)C

The SMILES string "C=C(c1cccc(N2CC3CCNC3C2)c1)N(C)C" represents a complex organic molecule. Here's the breakdown:

- "C=C": This portion represents an alkene group, a carbon-carbon double bond.
- "N(C)C": This signifies a tertiary amine group, specifically, dimethylamine ($\text{N}(\text{CH}_3)_2$), attached to the alkene.
- "c1cccc(N2CC3CCNC3C2)c1": This part represents a benzene ring, a six-membered aromatic ring of carbon atoms, with a complex substituent: "N2CC3CCNC3C2" signifies a bicyclic structure connected to the benzene ring. This structure is a piperazine ring, a six-membered ring with two nitrogen atoms and four carbon atoms.

In summary, this molecule consists of a benzene ring attached to a dimethylamine group through an alkene group. The benzene ring is also connected to a piperazine ring. Such a structure could potentially have significant biological activity, as both benzene and piperazine rings are common in bioactive compounds. The presence of an alkene could also impart certain reactivity. However, the exact properties would depend on many factors, including the specific 3D configuration, which is not directly captured by the SMILES string.

Dialogue 16: Misinformation provided by ChatGPT. Complete interactions with ChatGPT can be found in Supporting Information S4.4.

Beside, we noticed that ChatGPT does not perform well in the role of methodology explanation. The 2nd

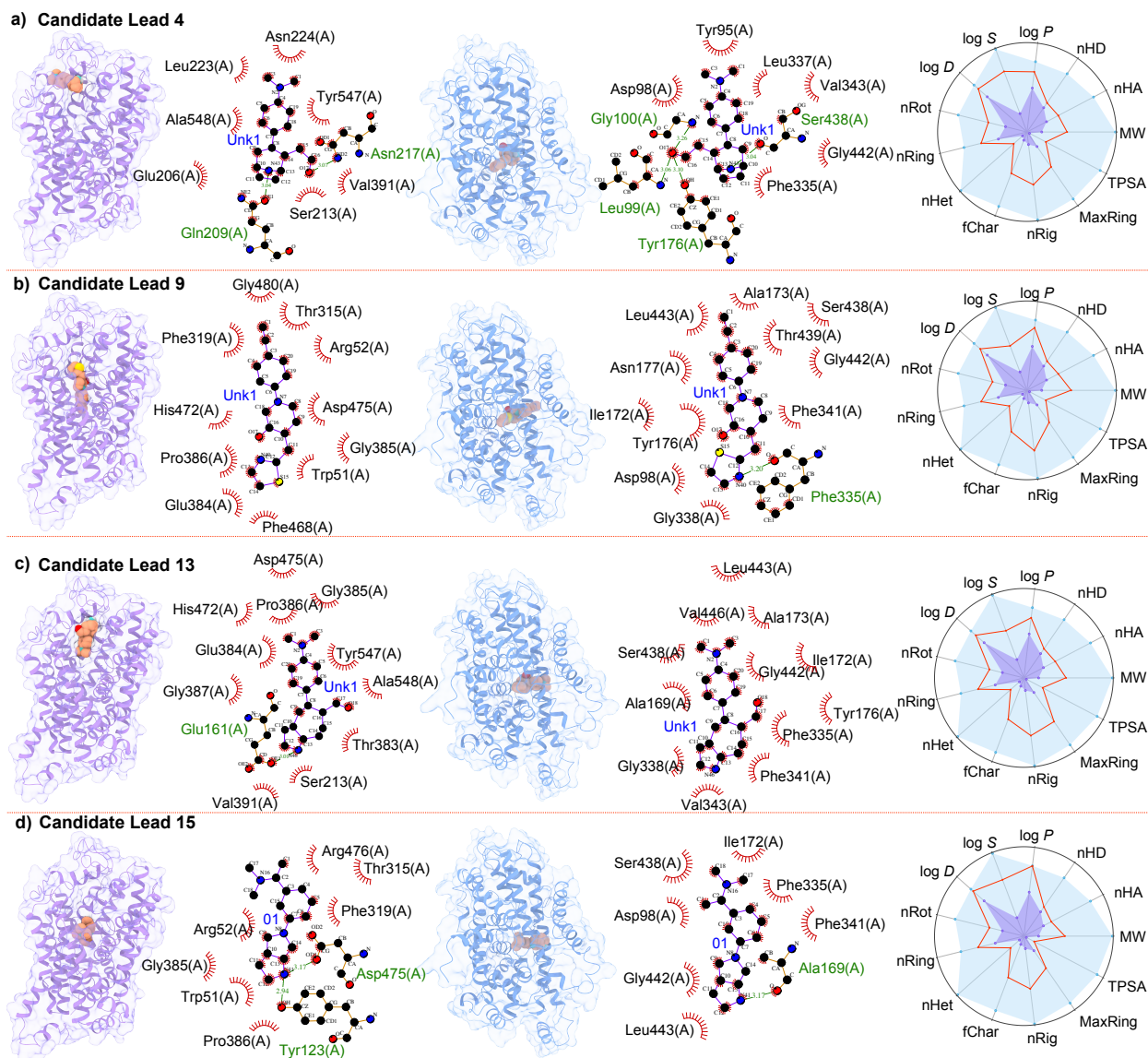
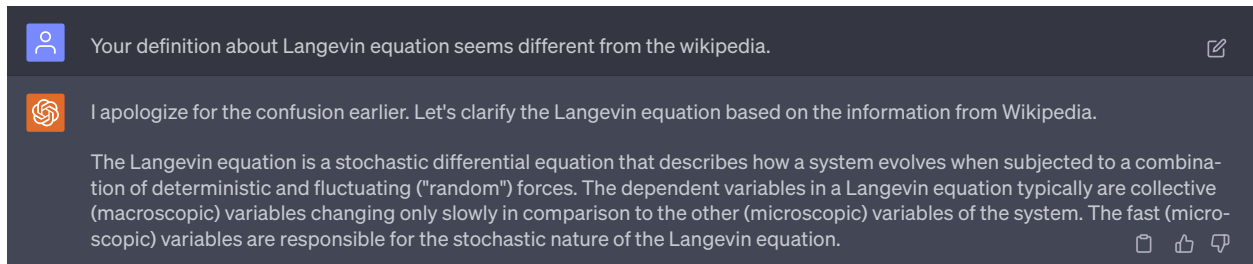


Figure 4: Predicted docking poses of selected lead candidates to DAT (1st column) and SERT (3rd column) by AutoDock Vina. DAT is colored in purple, and SERT is presented in blue. The 2nd and 4th columns demonstrate the molecular interaction of these leads with DAT and SERT, respectively. The final column portrays the physicochemical properties of the lead candidates, which include MW (molecular weight), log P (logarithm of octanol/water partition coefficient), log S (logarithm of the aqueous solubility), log D (log P at physiological pH 7.4), nHA (number of hydrogen bond acceptors), nHD (number of hydrogen bond donors), TPSA (topological polar surface area), nRot (number of rotatable bonds), nRing (number of rings), MaxRing (number of atoms in the largest ring), nHet (number of heteroatoms), fChar (formal charge), and nRig (number of rigid bonds). Here the purple dots denote the minimal value and the blue dots indicate the maximal value within the optimal range. The red lines represent the values of the properties for each lead candidate. The figures are categorized as follows: **a)** candidate lead 3, **b)** candidate lead 9, **c)** candidate lead 13, and **d)** Candidate lead 15.

persona of ChatGPT frequently provide wrong definitions. Then we will not accept the response from ChatGPT and seek further clarification as suggested in [Figure 5 a](#)). For example, we can point the mistakes out and refer useful information to ChatGPT to help it modify its response as indicate in [Dialogue 17](#).

Additionally, we noticed that ChatGPT does not perform well when providing methodological explanations. Its responses from the 2nd persona contain some incorrect definitions and explanations. In such cases, we opted not to accept responses from ChatGPT, and seek further clarification followed by the workflow in [Figure 5 a](#)). An effective approach involved pointing out the inaccuracies to ChatGPT and supplying

it with accurate references or information, prompting it to adjust its responses. Specific instances of these methodological inaccuracies are detailed in [Dialogue 17](#). Despite the inaccurate responses provided by the 2nd persona of ChatGPT, it remained invaluable in helping us grasp numerous theoretical concepts and their interrelations, serving as an effective browser tool.



Dialogue 17: A succinct dialogue highlighting inaccuracies of ChatGPT responses and suggestions for corrections. Complete interactions with ChatGPT can be found in Supporting Information S4.2.

3.2 Autoencoder reconstruction rate of molecules

A Sequence-to-Sequence Autoencoder (Seq2Seq AE) is a specific type of neural network model designed to learn a compressed representation of input data and reconstruct this data from the obtained representation. The core objective of such an autoencoder is to minimize the discrepancy between its input and output data. In this study, we initially fed the Seq2Seq AE with SMILES strings derived from four distinct datasets to examine their respective reconstruction rates. The calculated reconstruction rates for DAT-Inhibitors, NET-Inhibitors, SERT-Inhibitors, and hERG-Inhibitors datasets are 0.958, 0.970, 0.968, and 0.950 respectively. These values signify a successfully implemented autoencoder model.

In addition to the aforementioned, we verified the reconstruction rate of molecules generated via the Seq2Seq AE. After eliminating duplicated SMILES strings from the generated set, the resulting reconstruction rate stood at 0.996. This high reconstruction rate implies that the distribution of our generated molecules closely mirrors that of the original dataset processed by Seq2Seq AE, underscoring the reliability of the molecules generated by our method.

3.3 Patterns sensitive latent space vector distributions

Initially, we introduced random Gaussian noise, with a range from -1 to 1, into the stochastic-based molecular generator. However, these perturbations in the latent space vectors resulted in weird SMILE strings once decoded. Seeking guidance, we consulted ChatGPT as referenced in [Dialogue 18](#), which provided us with eight potential solutions. After reviewing these suggestions, we aligned with the first and second suggestions that echoed findings from our previous work, emphasizing that random perturbations in the latent space can destabilize the Seq2Seq AE model [14]. To ensure the reliability and effectiveness of the decoder in Seq2Seq AE model, it is essential to maintain a similar distribution pattern between the original latent space vectors and those derived from the stochastic-based molecular generator. Therefore, we take efforts to tune the noise that added to the stochastic-based molecular generator, to guarantee the modified/edited latent space vectors retain a representation that the GNC model has learned to decode effectively.

[Figure 5 b\)](#) and [c\)](#) depict distribution of latent space vectors across various datasets. Here, the x -axis represents the latent space index (ranges from 0 to 511), and the y -axis shows the absolute average value of the latent space vector corresponding to each index. The representation of absolute average value helps in visualizing the pattern and magnitude of latent space vectors across a broad index range. The purple, green, and blue panels depict the latent space vector distribution from the DAT-Inhibitors, NET-Inhibitors, and SERT-Inhibitors respectively. The discrepancy can be observed in the grey panel of [Figure 5 c\)](#) (particularly in the red boxed area). None of the molecules generated by this untuned stochastic-based molecular

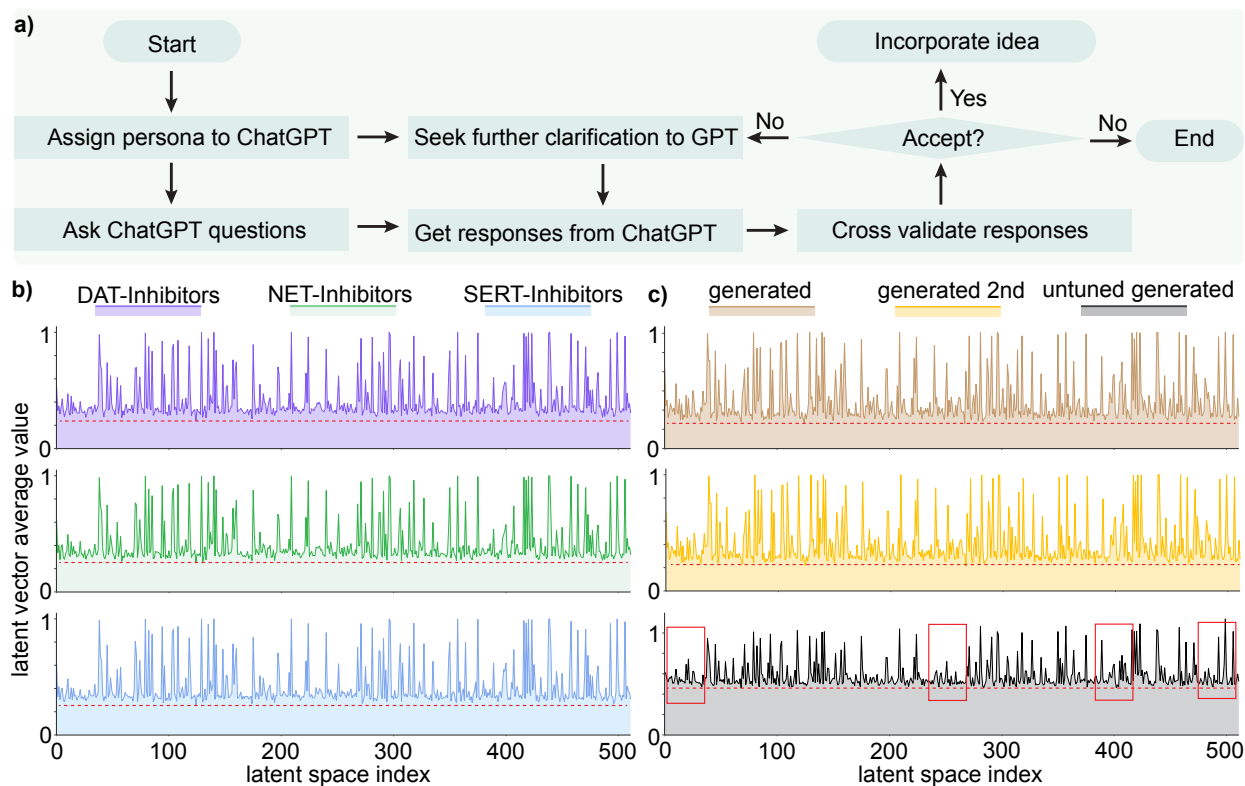


Figure 5: **a)** Flowchart of implementing ChatGPT as a virtual guide. **b)** Distribution of latent space vectors across various datasets. The x -axis represents the latent space index, which ranges from 0 to 511, while the y -axis denotes the absolute average value of the latent space vector corresponding to each index. The purple, green, and blue figures represent the latent space vector distribution from the DAT-Inhibitors, NET-Inhibitors, and SERT-Inhibitors respectively. **c)** Distribution of latent space vectors of generated molecules. The brown panel illustrates the latent space vector distribution from generated molecules from the fine-tuned stochastic-based molecular generator. The yellow distribution portrays generated molecules that have been processed by the GNC a second time. The grey distribution corresponds to generated molecules from an untuned stochastic-based molecular generator.

generator passed either the binding affinity requirements or ADMET tests. Subsequently, we adjusted the Gaussian noise in the stochastic-based molecular generator to ensure the edited latent spaces (represented in brown) exhibited a similar distribution to the original ones as shown in Figure 5 b). This controlled noise (ranges from -0.1 to 0.1) proved beneficial, leading to the generation of 15 promising leads capable of targeting DAT, NET, and SERT. Additionally, enlightened by ChatGPT (suggestion 8 in Dialogue 18), we also implemented a feedback loop where the molecules generated by our fine-tuned stochastic molecular generator is re-encoded into the latent space. It is worth noting that these re-encoded latent space vectors maintained a similar distribution, suggesting that the modifications made to the latent vectors are within the learned parameters of the SGNC.

4 Methods

4.1 Datasets preparation

Four pharmaceutical targets are key to treating cocaine addiction and drug discovery: Dopamine Transporter (DAT), Norepinephrine Transporter (NET), Serotonin Transporter (SERT), and Human Etherá-go-go-Related Gene (hERG). DAT is responsible for dopamine reuptake from synapses back into neurons, terminating neurotransmitter signaling. This causes dopamine accumulation in synapses and inducing intense euphoria. Similarly, NET is inhibited by cocaine, leading to elevated norepinephrine levels in synapses, contributing to stimulant effects. Furthermore, the inhibition of cocaine and SERT increases serotonin lev-



I added gaussian noise to my molecular generator aiming to modify my latent space vector. However, I noticed that these edited latent space vectors result in a weird SMILE string once they pass through the decoder. Do you have any idea of this happens.



When you introduce Gaussian noise to the latent space vectors in a molecular generator, you're essentially perturbing the representations of the molecules. These perturbations can lead to vectors that, when decoded, don't correspond to valid molecular structures. Here's why this might be happening and some potential solutions:

1. **Nature of the Latent Space:** The latent space of molecular generators, especially those trained with deep learning models like Variational Autoencoders (VAEs), is not uniformly populated with valid molecular representations. There are regions in this space that correspond to valid molecules, but there are also vast "voids" or "deserts" that don't map to any valid molecular structure. When you add Gaussian noise, you might be pushing the latent vector into one of these void regions, resulting in an invalid SMILES string upon decoding.
2. **Magnitude of the Noise:** The amount of Gaussian noise you add is crucial. Too much noise can push the latent vector far from its original position, increasing the likelihood of landing in an invalid region of the latent space. You might want to experiment with the standard deviation of your Gaussian noise to find a balance.
3. **Regularization of the Latent Space:** One way to make the latent space more robust to such perturbations is to use regularization during training. Techniques like contractive autoencoders introduce a penalty on the latent space, encouraging it to be more densely populated with valid molecular representations.
4. **Post-Processing:** After adding the Gaussian noise and decoding, you can use cheminformatics libraries like RDKit to validate the generated SMILES strings. If a SMILES string is invalid, you can discard it or apply corrective measures.
5. **Constrained Noise Addition:** Instead of adding noise indiscriminately, you can guide the noise addition process. For instance, you can explore the latent space to identify directions that lead to valid molecular changes (e.g., addition/removal of specific functional groups) and add noise preferentially along these directions.
6. **Fine-Tuning with Noise:** Once you've added Gaussian noise to your latent vectors, you can fine-tune your generator using these noisy vectors. This can make the generator more robust to such perturbations and improve the validity of the decoded SMILES strings.
7. **Use a Different Noise Distribution:** Gaussian noise might not always be the best choice. Depending on the nature of your latent space, other noise distributions (e.g., uniform noise) might be more appropriate.
8. **Feedback Loop:** Implement a feedback loop where invalid SMILES strings are re-encoded into the latent space, slightly adjusted, and then decoded again. This iterative process can sometimes "correct" the latent vector to produce a valid SMILES string.

In summary, while adding noise to the latent space is a common technique to explore and generate diverse molecular structures, it's essential to ensure that the perturbations lead to valid molecular representations. By understanding the nature of your latent space and experimenting with different strategies, you can improve the validity of the generated SMILES strings.



Dialogue 18: Solutions provided by ChatGPT regarding the random perturbation issue in the stochastic-based molecular generator.

els in the synapse, resulting in mood elevation, anxiety, and paranoia. Thus, a compound that concurrently modulates DAT, NET, and SERT activities could potentially treat cocaine addiction. Additionally, blocking the hERG potassium ion channel can lead to potentially fatal cardiac arrhythmias. Therefore, it is also critical to consider binding affinity between hERG and new generated leads.

In this study, we collected SMILES strings and binding affinities of inhibitors targeting DAT, NET, SERT, and hERG from the ChEMBL database. A summary of the size and label of datasets can be found in Table 2. We leveraged these four datasets in three ways: 1) training four separate binding affinity predictors based on the DAT-Inhibitors, NET-Inhibitors, SERT-Inhibitors, and hERG-Inhibitors datasets respectively, 2) selecting molecules from the DAT-Inhibitors, NET-Inhibitors, and SERT-Inhibitors datasets as reference and seed compounds for the generation of new drug-like molecules, and 3) inputting the SMILES strings from all four datasets into a sequence-to-sequence AutoEncoder model to validate their successful reconstruction rate.

To be noted, we use the binding affinity in terms of Gibbs free energy (ΔG) with units of kcal/mol, rather than the inhibition constant (K_i) or half maximal inhibitory concentration (IC_{50}). The Gibbs free energy of binding provides a direct measure of the favorability of the binding process, which indicates a more favorable binding interaction that can be interpret in the drug-target interactions. The conversion from K_i to ΔG is [27]:

$$\Delta G = -RT \ln K_i \approx 1.3633 \times \log_{10} K_i \text{ (kcal/mol),}$$

where R is the gas constant (1.987 cal/mol·K) and T (298.15 K) is the absolute temperature. In addition,

the IC_{50} can be approximated to K_i in the case of competitive and uncompetitive inhibition according to Kalliokoski [28] under:

$$K_i = IC_{50}/2.$$

4.2 Stochastic-based generative network complex (SGNC)

In this section, after thorough evaluation and incorporation of suggestions from GPT-4, we introduce the stochastic-based generative network complex (SGNC) as a novel mathematical-AI model, which is designed to generate novel molecules that potentially serve as effective treatments for cocaine addiction. Specifically, these molecules are intended to target multiple sites such as the Dopamine Transporter (DAT), Norepinephrine Transporter (NET), and Serotonin Transporter (SERT).

Figure 1 illustrates the workflow of SGNC, which essentially consists of four main components: 1) Sequence-to-Sequence AutoEncoder (shown in green), 2) Binding Affinity Predictors (shown in yellow), 3) Stochastic-based Molecular Generator (shown in blue), and 4) Analysis via ADMET Lab (shown in purple). The dark arrows represent the training process, the brown arrows show the validation process, and the red arrows indicate the generation process.

For the training process, we leveraged a well-established translation model, specifically a sequence-to-sequence (Seq2Seq) AutoEncoder (AE). This model was developed to map the International Union of Pure and Applied Chemistry (IUPAC) representation of a molecule to its Simplified Molecular Input Line Entry System (SMILES) representation, as mentioned in [29]. In our prior research, we modified this model by switching the input from the IUPAC representation of molecules to their corresponding SMILES strings.

The generation process involves the following main steps:

1. We initially selected one molecule each from the DAT, NET, and SERT datasets. These molecules were chosen because of their relatively high similarity within the three datasets, thereby acting as our reference compounds. In addition, we selected a compound known for its potency against all three targets to serve as our seed compound.
2. We then input the reference and seed compounds into the pretrained encoder and extracted the corresponding latent vectors from the latent space of the Seq2Seq AE. Subsequently, we modified the seed vector in the stochastic-based molecular generator, using the information from the reference molecules as a guide. As a result, the generator was capable of producing a large number of new latent vectors. These vectors were then decoded into SMILES strings, which are potentially effective against multiple targets, namely DAT, NET, and SERT.
3. Furthermore, we input these decoded SMILES strings into our binding affinity (BA) predictors to filter out molecules that meet our BA requirement (i.e., $\Delta G < -9.54$ kcal/mol on DAT, NET, SERT and $\Delta G > -8.18$ kcal/mol on hERG).
4. Finally, we used ADMETlab 2.0 to select drugable molecules from the generated SMILES with desirable BA properties. This final step in the generation process ensures that the compounds not only bind effectively to the desired targets but also have the necessary absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties for a potential lead compound.

During the validation phase, we first input the SMILES strings from the DAT, NET, SERT, and hERG datasets into our well-trained Seq2Seq AE model to obtain decoded SMILES. Successful reconstruction of the input SMILES signifies the reliability of the Seq2Seq AE model. Furthermore, we input the generated SMILES, which have been processed through the stochastic-based molecular generator, into the pre-trained Seq2Seq AE model. In case of unsuccessful reconstruction, we adjust the hyperparameters of the stochastic-based molecular generator until a high reconstruction rate is achieved. This process indicates that the latent vectors edited by the stochastic-based molecular generator maintain a similar distribution to the original

latent space vectors from the encoder, which further ensures that our SGNC model is capable of generating chemically feasible compounds, reflecting its potential in drug discovery applications.

4.2.1 Sequence-to-sequence autoencoder

The Sequence-to-sequence autoencoder (Seq2Seq AE) is an artificial neural network model used for translating the IUPAC representation of a given molecule into its SMILES string representation [29]. In our study, the Seq2Seq AE accepts the SMILES representation of a molecule as the input for the encoder. Subsequently, the latent space of the Seq2Seq AE preserves the structural and functional properties of the provided SMILES. This low-dimensional latent space representation can then be processed by the decoder of the Seq2Seq AE to reconstruct the original SMILES representation. Here, the network that used in encoder and decoder is the gated recurrent unit (GRU). In this work, the pretrained Seq2Seq AE model was utilized from a previous work by Winter et al [29].

The Seq2Seq AE model employed in our study was pretrained on 72 million compounds [29] sourced from the ZINC15 and PubChem databases. All duplicate entries within these databases were eliminated and subjected to RDKit [30] filtering using the following criteria: 1) only organic molecules, 2) molecular weight between 12 and 600 daltons, 3) more than 3 heavy atoms, 4) partition coefficient $\log P$ between -5 and 5, 5) stereochemistry was removed, 6) salts were stripped.

4.2.2 Binding affinity predictors

We constructed four binding affinity predictors based on four training datasets: DAT-Inhibitors, NET-Inhibitors, SERT-Inhibitors, and hERG-Inhibitors. These predictors are designed to estimate the binding affinity of potential molecules to four critical targets: DAT, NET, SERT, and hERG. The construction of the predictors involved the following steps:

- 1) Feature extraction: molecular features (or fingerprints), were derived from the latent space of a sequence-to-sequence AutoEncoder (Seq2Seq AE).
- 2) Label assignment: The labels used for model training were the binding affinities of the molecules to their respective targets.
- 3) Model training: we trained the predictors using PyTorch. Each network consisted of three hidden layers with 512, 1024, and 512 neurons, respectively. The networks were trained over 1000 epochs, with a learning rate of 0.0001 for the first 500 epochs and 0.00001 for the remaining 500 epochs. We chose the Adam optimizer and batch size 16 for this task.

4.2.3 Stochastic-based molecule generator

Generative models have gained prominence as potent tools for the generation of prospective new leads. Building upon our prior work, we introduced the Generative Network Complex (GNC), a model specifically tailored to produce novel, drug-like molecules [14]. To augment the efficacy of the GNC model, and with guidance from GPT-4, we decided to integrate principles from diffusion-based models [19,31].

The Langevin equation is a stochastic differential equation (SDE) that used to describe diffusion processes. This equation encapsulates the random trajectories of particles in their velocity space, accounting for both deterministic and stochastic forces. A pivotal goal of this research is to employ the Langevin equation suggested by ChatGPT as a mechanism to enhance the molecular generator present in the GNC model.

Assume $\mathbf{X}$ is a latent space vector of a molecule with 512 dimensions, and $\mathbf{X}_k$ represents its k -th latent

space reference vector. Then the Langevin equation of our drug generator system is:

$$\frac{d\mathbf{X}}{dt} = \alpha \sum_k a_k (\mathbf{X}_k - \mathbf{X}) + \boldsymbol{\xi}(t), \quad (1)$$

where a_k is a positive weighting parameter corresponds to $\mathbf{X}_k$ satisfying $\sum_k a_k = 1$, $\boldsymbol{\xi}(t)$ is a Gaussian white noise, and α is a hyperparameter. Then according to the Langevin equation in the Supporting Information S1.1.4, the general solution of this system is given by:

$$\mathbf{X}(t) = \mathbf{C}^{-\alpha t} + \int_0^t e^{-\alpha(t-u)} (\alpha \sum_k a_k \mathbf{X}_k + \boldsymbol{\xi}(u)) du, \quad (2)$$

where the initial state $\mathbf{X}(0) = \mathbf{C}$.

While the Langevin equation offers a microscopic depiction of the diffusion process, a comprehensive understanding of the temporal evolution of particle distribution requires a more macroscopic viewpoint. To bridge this gap, we also introduce the Fokker-Planck equation. Derived from the Langevin equation (a detailed derivation is available in Supporting Information S1.1.5), this equation provide the connection between the dynamics of individual particles and the overarching behavior of the entire system.

4.2.4 Drug screening

In drug discovery, several criteria are leveraged to filter promising drug candidates. In our work, we consider molecules that fulfill the following requirements as viable drug prospects: 1) exhibit favorable ADMET properties, 2) comply with Lipinski’s rule of five, 3) are synthetically accessible, 4) possess proper physicochemical properties. These properties are crucial for determining the drug-like nature and potential practical applicability of the generated molecules, which will be elaborated on in the following paragraphs.

Table 3: The optimal ranges of 10 properties that are used to screen nearly optimal compounds, including seven selected ADMET properties, two physicochemical properties, and one medicinal chemistry properties. The seven ADMET properties include Caco-2 (the human colon adenocarcinoma cell lines) permeability, $F_{20\%}$ (the human oral bioavailability 20%), Pgp-sub (the substrate of P-glycoprotein), Pgp-inh (the inhibitor of P-glycoprotein), VD (volumn density), $T_{1/2}$ (The half-life of a drug), and FDAMDD (The maximum recommended daily dose). Moreover, SAS represents the synthetic accessibility score, $\log P$ is the logarithm of the n-octanol/water distribution coefficient, and $\log S$ indicates the logarithm of aqueous solubility value.

Property	Profile	Excellent range	Medium range
Absorption	Caco-2 permeability	> -5.15	/
Absorption	$F_{20\%}$	0 - 0.3	0.3 - 0.7
Absorption	Pgp-sub	0 - 0.3	0.3 - 0.7
Absorption	Pgp-inh	0 - 0.3	0.3 - 0.7
Distribution	VD	0.04 - 20 L/kg	/
Excretion	$T_{1/2}$	0 - 0.3	0.3 - 0.7
Toxicity	FDAMDD	0 - 0.3	0.3 - 0.7
Medicinal Chemistry	SAS	< 6	/
Physicochemical	$\log P$	0 - 3 log mol/L	/
Physicochemical	$\log S$	-4 - 0.5 log mol/L	/

First, undesirable pharmacokinetics and toxicity are leading causes of drug development failure. Therefore, the assessment of absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties should occur as early as possible in the drug development process. In this work, we applied [ADMETlab 2.0](#) to offer us systematic evaluation of ADMET properties, along with certain physicochemical properties and an assessment of medicinal chemistry friendliness. In this work, we consider seven seven ADMET properties,

including Caco-2 (the human colon adenocarcinoma cell lines) permeability, $F_{20\%}$ (the human oral bioavailability 20%), Pgp-substrate (the substrate of P-glycoprotein), Pgp-inhibitor (the inhibitor of P-glycoprotein), VD (volumen density), $T_{1/2}$ (The half-life of a drug), and FDAMDD (The maximum recommended daily dose). The optimal range of these properties are listed in Table 3.

Second, the Lipinski's rule of five help to evaluate druglikeness or determine if a chemical compound with a certain pharmacological or biological activity has properties that would make it a likely orally active drug in humans, which should satisfies four physicochemical criteria: 1) molecular weight (MW) ≤ 500 daltons, 2) octanol-water partition coefficient ($\log P$) ≤ 5 , 3) the number of hydrogen bond donors (nHD) ≤ 5 , 4) the number of hydrogen bond acceptors (nHA) ≤ 10 .

Thirdly, synthetic accessibility is crucial to ensuring the feasibility of large-scale production of a potential drug candidate. In this study, we used RDKit to evaluate the synthetic accessibility score (SAS). A candidate drug with an SAS score less than 6 indicates that it is relatively easy to synthesize.

Lastly, physicochemical properties can significantly influence the solubility, permeability, and stability of potential drug candidates. In this study, we primarily focused on the logarithm of the n-octanol/water distribution coefficient ($\log P$) and the logarithm of aqueous solubility value ($\log S$). Drug candidates with a $\log P$ in the range of 0 - 3 $\log \text{ mol/L}$ and $\log S$ in the range of -4 - 0.5 $\log \text{ mol/L}$ are considered to have suitable physicochemical properties.

Code and Data availability

The code and data are available at the public repository <https://github.com/wangru25/SGNC>.

The datasets including SMILES strings and binding affinities of inhibitors targeting DAT, NET, SERT. In addition, these datasets can also be found in the 'Training Datasets' folder within the SupplementaryData.zip file, available under Supporting Information S2 for readers interested in further exploration.

Trained models from this study are saved within the aforementioned code repository. This repository includes the stochastic-based generative network complex (SGNC) developed in Python, as well as Python scripts for calculating reconstruction rates, evaluating synthetic accessibility, and generating visual plots.

Supporting Information

The Supporting Information is available for:

S1 Supplementary methods

S1.1 Fokker-Planck equation-embedded multi-target drug molecule generator

S1.1.1 Random variables

S1.1.2 Wiener process and white noise

S1.1.3 Itô's lemma

S1.1.4 Langevin equation

S1.1.5 Derivation of Fokker-Planck equation from Langevin equation

S1.2 Evaluation metrics

S2 Supplementary Data: The SupplementaryData.zip consists 3 folders, namely Training Datasets, Predictions, and Generated Molecules.

S2.1 Training Datasets: This folder has the datasets used for training purposes.

S2.2 Predictions: Within this folder, one can find data related to the predicted binding affinity of inhibitors in 4 training datasets.

S2.3 Generated Molecules: This folder documents the molecules that have been produced using the stochastic-based molecular generator.

S3 Supplementary Figures

S3.1 Radar plots of physicochemical properties for 15 lead candidates

S3.2 Molecular docking and molecular interaction of 15 leads with DAT and SERT

S4 Supplementary Dialogues

S4.1 The 1st persona of ChatGPT

S4.2 The 2nd persona of ChatGPT

S4.3 The 3rd persona of ChatGPT

S4.4 Other dialogues

Acknowledgment

This work was supported in part by NIH grants R01GM126189, R01AI164266, and R35GM148196, National Science Foundation grants DMS2052983, DMS-1761320, and IIS-1900473, NASA grant 80NSSC21M0023, Michigan Economic Development Corporation, Michigan State University Foundation, Bristol-Myers Squibb 65109, and Pfizer.

References

- [1] Eleni Adamopoulou and Lefteris Moussiades. An overview of chatbot technology. In *IFIP international conference on artificial intelligence applications and innovations*, pages 373–383. Springer, 2020.
- [2] Himanshu Bansal and Rizwan Khan. A review paper on human computer interaction. *Int. J. Adv. Res. Comput. Sci. Softw. Eng*, 8(4):53, 2018.
- [3] Qing Lyu, Josh Tan, Mike E Zapadka, Janardhana Ponnatapuram, Chuang Niu, Ge Wang, and Christopher T Whitlow. Translating radiology reports into plain language using chatgpt and gpt-4 with prompt learning: Promising results, limitations, and potential. *arXiv preprint arXiv:2303.09038*, 2023.
- [4] Zehua Zeng and Hongwu Du. Revolutionizing single cell analysis: The power of large language models for cell type annotation. *arXiv preprint arXiv:2304.02697*, 2023.
- [5] Andrew D White, Glen M Hocky, Heta A Gandhi, Mehrad Ansari, Sam Cox, Geemi P Wellawatte, Subarna Sasmal, Ziyue Yang, Kangxin Liu, Yuvraj Singh, et al. Assessment of chemistry knowledge in large language models that generate code. *Digital Discovery*, 2(2):368–376, 2023.
- [6] Yijun Wu and Ailin Zhao. Future implications of chatgpt in pharmaceutical industry: Drug discovery and development. *Frontiers in Pharmacology*, 14:1194216.
- [7] Neil Savage. Drug discovery companies are customizing chatgpt: here’s how. *Nature Biotechnology*, 2023.

- [8] Jui-Hsuan Lee, Eric Hsiao-Kuang Wu, Yu-Yen Ou, Yueh-Che Lee, Cheng-Hsun Lee, and Chia-Ru Chung. Anti-drugs chatbot: Chinese bert-based cognitive intent analysis. *IEEE Transactions on Computational Social Systems*, 2023.
- [9] Changwei Gong, Changhong Jing, Ye Li, Xinan Liu, Zuxin Chen, and Shuqiang Wang. Generative artificial intelligence-enabled dynamic detection of nicotine-related circuits. *arXiv preprint arXiv:2212.06330*, 2022.
- [10] Hongsong Feng, Kaifu Gao, Dong Chen, Li Shen, Alfred J Robison, Edmund Ellsworth, and Guo-Wei Wei. Machine learning analysis of cocaine addiction informed by dat, sert, and net-based interactome networks. *Journal of chemical theory and computation*, 18(4):2703–2719, 2022.
- [11] Yuwei Yang, Chang-Yu Hsieh, Yu Kang, Tingjun Hou, Huanxiang Liu, and Xiaojun Yao. Deep generation model guided by the docking score for active molecular design. *Journal of Chemical Information and Modeling*, 2023.
- [12] Xiaolin Pan, Hao Wang, Yueqing Zhang, Xingyu Wang, Cuiyu Li, Changge Ji, and John ZH Zhang. Aa-score: a new scoring function based on amino acid-specific interaction for molecular docking. *Journal of Chemical Information and Modeling*, 62(10):2499–2509, 2022.
- [13] Pedro J Ballester and John BO Mitchell. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics*, 26(9):1169–1175, 2010.
- [14] Kaifu Gao, Duc Duy Nguyen, Meihua Tu, and Guo-Wei Wei. Generative network complex for the automated generation of drug-like molecules. *Journal of chemical information and modeling*, 60(12):5682–5698, 2020.
- [15] Christopher Grow, Kaifu Gao, Duc Duy Nguyen, and Guo-Wei Wei. Generative network complex (gnc) for drug discovery. *Communications in information and systems*, 19(3):241, 2019.
- [16] Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107, 2012.
- [17] Kaifu Gao, Dong Chen, Alfred J Robison, and Guo-Wei Wei. Proteome-informed machine learning studies of cocaine addiction. *The journal of physical chemistry letters*, 12(45):11122–11134, 2021.
- [18] Aayush Gupta and Huan-Xiang Zhou. Machine learning-enabled pipeline for large-scale virtual drug screening. *Journal of Chemical Information and Modeling*, 61(9):4236–4244, 2021.
- [19] Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. *arXiv preprint arXiv:2203.02923*, 2022.
- [20] Hannes Risken and Hannes Risken. *Fokker-planck equation*. Springer, 1996.
- [21] Hakime Öztürk, Elif Ozkirimli, and Arzucan Özgür. A comparative study of smiles-based compound similarity functions for drug-target interaction prediction. *BMC bioinformatics*, 17(1):1–11, 2016.
- [22] Jane Y Yang, Laura M Sanchez, Christopher M Rath, Xueting Liu, Paul D Boudreau, Nicole Bruns, Evgenia Glukhov, Anne Wodtke, Rafael De Felicio, Amanda Fenner, et al. Molecular networking as a dereplication strategy. *Journal of natural products*, 76(9):1686–1699, 2013.
- [23] Xiaoxia Chen, Hao Li, Lichao Tian, Qinwei Li, Jinxiang Luo, and Yongqiang Zhang. Analysis of the physicochemical properties of acaricides based on lipinski’s rule of five. *Journal of computational biology*, 27(9):1397–1406, 2020.
- [24] Darren R Flower. *Drug design: cutting edge approaches*, volume 279. Royal Society of Chemistry, 2002.

- [25] Daniel Probst and Jean-Louis Reymond. Smilesdrawer: parsing and drawing smiles-encoded molecular structures using client-side javascript. *Journal of chemical information and modeling*, 58(1):1–7, 2018.
- [26] Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.
- [27] Zixuan Cang and Guo-Wei Wei. Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. *International journal for numerical methods in biomedical engineering*, 34(2):e2914, 2018.
- [28] Tuomo Kallioikoski, Christian Kramer, Anna Vulpetti, and Peter Gedeck. Comparability of mixed ic50 data—a statistical analysis. *PloS one*, 8(4):e61007, 2013.
- [29] Robin Winter, Floriane Montanari, Frank Noé, and Djork-Arné Clevert. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chemical science*, 10(6):1692–1701, 2019.
- [30] Greg Landrum et al. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*, 2013.
- [31] 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.

Competing interests

The authors declare no competing interests.