Target-aware Molecule Generation for Drug Design Using a Chemical Language Model*

Yingce Xia[1*†], Kehan Wu[2†], Pan Deng[1†], Renhe Liu[3†], Yuan Zhang[3], Han Guo[3], Yumeng Cui[3], Qizhi Pei[4], Lijun Wu[1], Shufang Xie[1], Si Chen[3], Xi Lu[3], Song Hu[3], Jinzhi Wu[3], Chi-Kin Chan[3], Shuo Chen[3], Liangliang Zhou[3], Nenghai Yu[2], Haiguang Liu[1], Jinjiang Guo[3*], Tao Qin[1*] and Tie-Yan Liu[1]

[1]Microsoft Research AI4Science.
[2]University of Science and Technology of China.
[3]Global Health Drug Discovery Institute.
[4]Renmin University of China.

*Corresponding author(s). E-mail(s): yingce.xia@microsoft.com; jinjiang.guo@ghddi.org; taoqin@microsoft.com;
†These authors contributed equally to this work.

Abstract

Generative drug design facilitates the creation of compounds effective against pathogenic target proteins. This opens up the potential to discover novel compounds within the vast chemical space and fosters the development of innovative therapeutic strategies. However, the practicality of generated molecules is often limited, as many designs focus on a narrow set of drug-related properties, failing to improve the success rate of subsequent drug discovery process. To overcome these challenges, we develop TamGen, a method that employs a GPT-like chemical language model and enables target-aware molecule generation and compound refinement. We demonstrate that the compounds generated by TamGen have improved molecular quality and viability. Additionally, we have

*This work is an upgraded version of [1]: While the previous work primarily emphasized the AI model and computational verification, this version accentuates the findings related to ClpP inhibitors.

integrated TamGen into a drug discovery pipeline and identified 7 compounds showing compelling inhibitory activity against the Tuberculosis ClpP protease, with the most effective compound exhibiting a half maximal inhibitory concentration (IC_{50}) of **1.9** µM. Our findings underscore the practical potential and real-world applicability of generative drug design approaches, paving the way for future advancements in the field.

Keywords: Generative drug design, Structure-based drug design, chemical language model, Generative AI, GPT, Tuberculosis

1 Introduction

Generative drug design, a promising avenue for drug discovery, aims to create novel molecules/compounds with desired pharmacological properties from scratch, without relying on existing templates or molecular frameworks [2, 3]. While conventional screening-based approaches, such as high-throughput screening, virtual screening, and emerging deep learning-based screening [4–7] usually hunt for drug candidates from libraries with 10^4 to 10^8 molecules [8–10], generative drug design enables exploration of the vast chemical space, which is estimated to contain over 10^{60} feasible compounds [11]. Consequently, it holds potential to identify underexplored classes of compounds, and novel compounds that are not in any existing library. This is especially important for target proteins without hit compounds (starting point for drug design) and those having developed resistance to current drugs.

Generative modeling techniques greatly empowers drug design. In recent years, a growing number of approaches have been proposed to guide the generation of drug-like compounds given the information of target proteins [12–17], stemming from creative artificial intelligence techniques such as autoregressive models [18], generative adversarial networks (GAN) [19], variational autoencoders (VAE) [20], and diffusion models [12]. These approaches, by exploring the chemical space conditioned on the target of interest, have demonstrated the feasibility of target-based generative drug design with deep learning. However, validations with biophysical or biochemical assays are often missing [21], as most of the generated compounds lack satisfying physiochemical properties for drug-like compounds such as synthetic accessibility. In other words, despite generating a large number of novel compounds, existing approaches struggle to demonstrate their capability to provide effective candidates that can improve the real-world drug discovery effectiveness.

We therefore propose a method named TamGen (Target-aware molecular generation). TamGen features a GPT-like chemical language model aiming for drug-like compound generation, inspired by the success of large language models [22]. The Generative Pre-trained Transformer [23] (GPT), backbone of large language models, has demonstrated its effectiveness in generating not only text [22] but also images [24] and speech [25], as well as understanding

and solving scientific problems [26]. Here, we demonstrate that a GPT-like architecture and training strategy are also effective for generating chemical compounds, as these compounds can be represented using Simplified Molecular Input Line Entry System (SMILES) [27], a sequential representation akin to text. In addition, we introduce two modules to encode target protein and compound information, which allow target-aware generation of compounds based on protein structures and compound refinement based on seeding compounds, respectively. With benchmark test, we show that TamGen not only produces compounds with higher plausibility, but also enhances the balance between pharmacological activity and synthetic accessibility.

We applied TamGen to generate compounds against tuberculosis (TB), an infectious disease caused by *Mycobacterium tuberculosis* (Mtb). TB was responsible for 1.3 million fatalities and 10.6 million new cases in 2022 [28, 29], and the rising antimicrobial resistance (AMR) in tuberculosis necessitates urgent therapeutic innovation to tackle the disease [30, 31]. We focused on Caseinolytic protease P (ClpP), an essential serine protease in bacterial protein degradation system and an emerging novel target for antibiotic development [32–35]. Using a Design-Refine-Test pipeline powered by TamGen, we discovered 7 candidate compounds showing promising potency against Mtb ClpP, with half maximal inhibitory concentrations (IC_{50}) ranging from 1.88 µM to 19.9 µM. Significantly, the compounds generated by TamGen not only enrich candidate pool for further optimization, but also provide effective anchors for hit expansion and structure-activity relationship (SAR) synthesis. These findings highlight the broad applicability and considerable potential of TamGen in target-aware drug design.

2 Results

2.1 TamGen enables target-aware compound design and refinement

We implemented TamGen with three modules: (1) compound decoder, a GPT-like chemical language model and the core component of TamGen, which lays the foundation for compound generation in chemical space; (2) protein encoder, a Transformer-based model used to encode the binding pockets of target proteins; and (3) a contextual encoder for compound encoding and refinement.

The compound decoder was pre-trained on 10 million SMILES randomly sampled from PubChem. The compound decoder adopts the autoregressive pre-training objective used in GPT, aiming to predict the next SMILES token based on preceding tokens (Fig. 1a). This training strategy allows for the sequential generation of compounds in both unconditional and conditional manners, depending on whether target information is provided or not. With this pre-training strategy, TamGen is able to learn general and diverse knowledge about a multitude of compounds from chemical databases (e.g., PubChem), without requiring any additional information such as binding

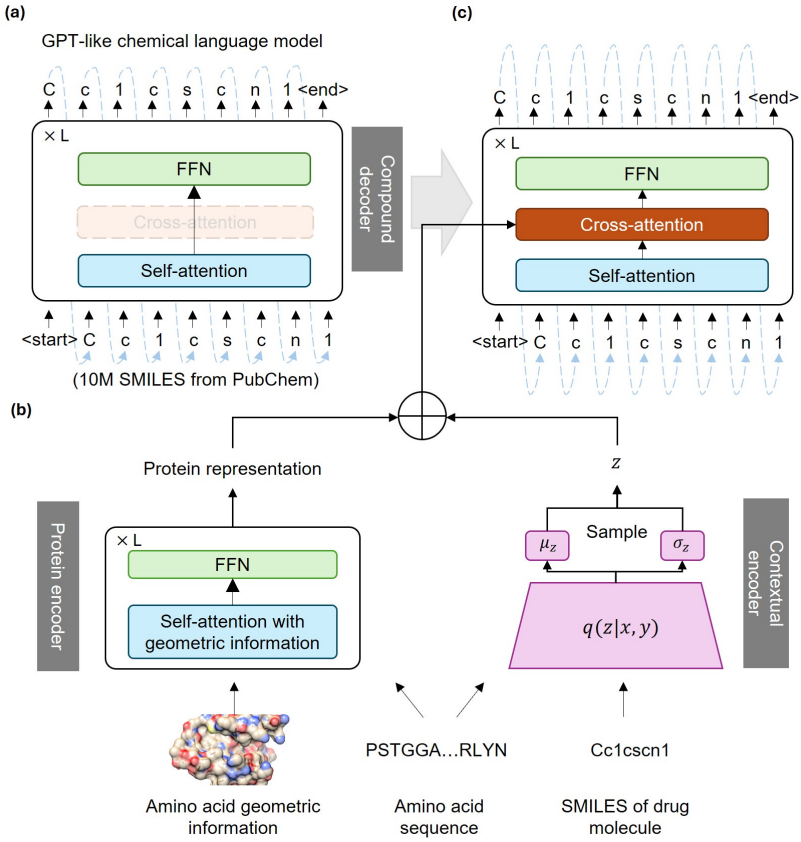


Fig. 1 The architecture of TamGen. **(a)** The pre-training phase of the compound decoder, a GPT-like chemical language model. The model adopts standard GPT architecture, which autoregressively generates the SMILES tokens from the input. 10 million compounds randomly selected from PubChem were used for pre-training. **(b-c)** The overall framework of TamGen during the fine-tuning and inference stages. **(b)** A Transformer-based protein encoder and a VAE-based contextual encoder to facilitate target-aware drug generation and seeding molecule-based compound refinement. See Methods and Figure S1 for details. **(c)** The outputs from the protein encoder and the contextual encoder are integrated and forwarded to the compound decoder via a cross-attention module.

proteins. This strategy enhances the generation capability of the compound decoder and improves the chemical properties of the generated compounds.

The protein encoder was developed to comprehend target protein information and to facilitate the generation of drug-like compounds in a target-aware manner (Fig. 1b left). The Transformer architecture adopted by the protein encoder features a self-attention mechanism, which gathers and processes information from input sequences. Here, we designed a variant of self-attention to capture both the sequential and geometric data of target proteins (Fig. S1, see Methods for details). The protein encoder's outputs are then directed to the

compound decoder via a cross-attention module (Fig. 1c), activated only when target proteins are provided. Therefore, we are able to generate compounds from the 3D conformation of target proteins via the protein encoder-compound decoder framework.

A Variational Autoencoder (VAE)-based contextual encoder was employed to encode compounds and assist the generation process. VAEs are commonly used to create new data by learning the input data's probability distribution and sampling from it [36]. In TamGen, the VAE-based contextual encoder determines the mean (μ) and standard deviation (σ) for any given compound $\mathbf{y}$ and protein sequence $\mathbf{x}$ pair (Fig. 1b right). Later, a vector z is sampled from the distribution determined by μ and σ and added to the output of protein encoder, before directed to the compound decoder (Fig. 1b right). In the training stage, the model's objective is to recover the input compound $\mathbf{y}$, whereas during application, the contextual encoder facilitates compound refinement once a seeding molecule is provided. The incorporation of this encoder enhances control over compound generation, enabling TamGen to be seamlessly integrated into multi-round drug optimization pipelines with human feedback. This interactive and iterative drug design capability holds the potential to increase the success rate of designed compounds and accelerate the drug discovery process.

2.2 TamGen is effective and efficient for generative drug design

To benchmark the overall performance of TamGen, we compared our methods against five approaches proposed recently: liGAN [37], 3D-AR [38] (there is no abbreviation for the proposed method, so we refer to it as 3D-AR), Pocket2Mol [14], ResGen [39] and TargetDiff [12]. These approaches focus on direct generation of compounds in the 3D space to match protein binding pockets with diverse deep learning techniques. Following previous practices, we evaluated these methods and TamGen on CrossDocked2020 dataset [40], a well-established benchmark dataset curated from PDBbind. CrossDocked2020 is composed of a train set with about 100,000 drug-target pairs and a test set with 100 protein binding pockets. For fair comparison with previous work, we used the same training and test data as those used in [12, 14] to fine-tune TamGen.

We generated 100 compounds for each target protein in CrossDocked2020 test set with each method respectively. Then, we evaluated the designed compounds using a comprehensive set of metrics: binding affinity to target proteins, estimated by docking scores from Autodock-Vina [41]; drug-likeness, assessed using both the Quantitative Estimate of Drug-likeness (QED) [42] and Lipinski's Rule of Five [43] based on calculated molecular physicochemical properties; synthetic accessibility scores (SAS), estimated by RDKit as a proxy for the ease of synthesis of a compound [44]; and LogP, an indicative of molecular lipophilicity, with an optimal range of 0-5 for oral administration [45]. In addition, we quantified the ability to generate diverse compounds

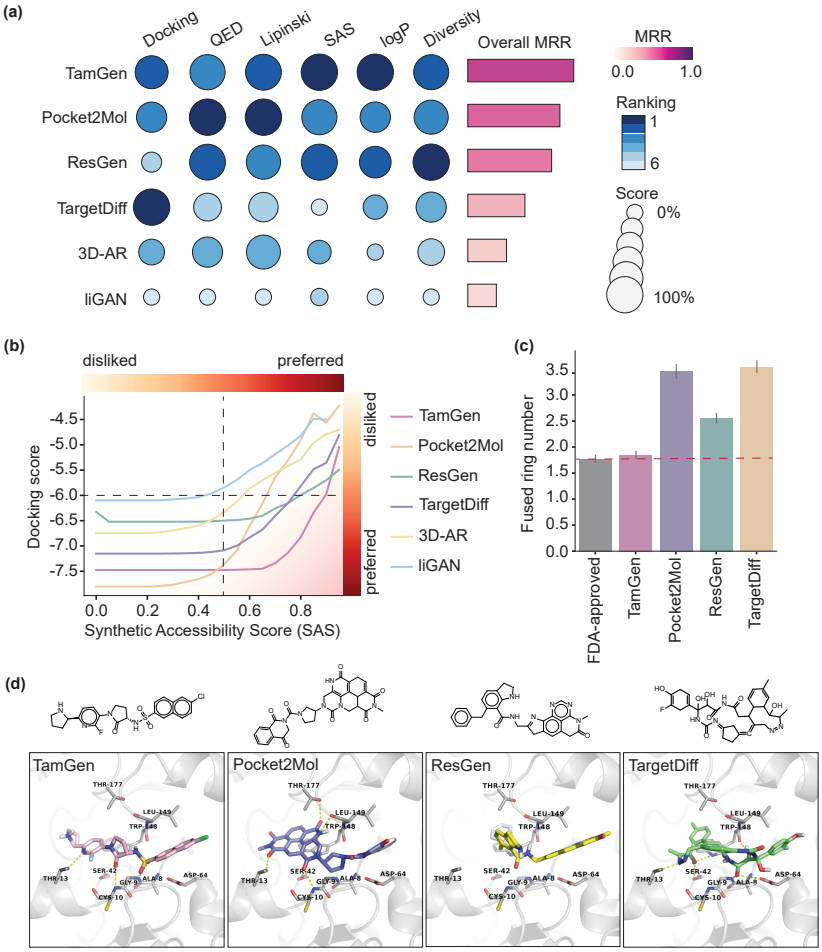


Fig. 2 TamGen achieves the state-of-the-art performance on compound generation. (a) Overview of generative drug design methods ranked by overall scores for the CrossDocked2020 task. Metrics include docking score (lower scores indicate better binding affinity), quantitative estimation of drug-likeness (QED), Lipinski's Rule of Five, Synthetic accessibility scores (SAS), LogP, and molecular diversity (Div). Scores were normalized to 0%-100% for each metric. Absolute values were used for docking score normalization. Overall scores were calculated with mean reciprocal rank (see Methods for details). See also Figure S2 and Table S1. **(b)** Average docking scores against SAS for TamGen and alternate methods. TamGen achieves more favorable docking scores for compounds with higher SAS and lower docking scores (bottom-right corner). **(c)** Barplot of the number of fused rings (see Methods for details) in FDA-approved drugs and top-ranked compounds generated by selected methods. For each method, a statistics of 1,000 compounds (100 targets $\times$ 10 compounds with the highest docking scores against each corresponding target) were plotted. The dashed line represents the average number of fused rings in FDA-approved drugs. Error bar, 95% confidence interval. **(d)** Example compounds generated by selected methods, and their binding poses to ClpP protein (shown as ribbons, with key residues shown as sticks).

of each method with molecular diversity. Molecular diversity is derived from the Tanimoto similarity between Morgan fingerprints of compounds. This set of metrics provides a broad and complementary assessment of compound properties, indicating the overall efficacy of a drug design method.

While each method demonstrates strengths across certain metrics, TamGen is consistently top ranked. For example, TamGen achieves either the first or the second place in 5 out of 6 metrics and exhibits the best overall performance (Fig. 2a, Fig. S2 and Table S1). This finding shows that TamGen is capable of simultaneously optimizing multiple aspects of compounds during the generation process.

Among the metrics, synthetic accessibility is an important factor affecting the practicality of a drug candidate, especially for novel compounds. It is worth pointing out that TamGen performs the best in terms of SAS for compounds with high binding affinity (reflected on docking scores, Fig. 2b), which are likely to possess superior bioactivity against target proteins. To discern why TamGen generates compounds with both high binding affinity and favorable SAS, we examined the top-scoring compounds generated by TamGen and other methods. Our analysis reveals that TamGen tends to produce compounds with fewer fused rings (Fig. 2c and Fig. S3). Notably, the number of fused rings in compounds generated by TamGen aligns closely with FDA-approved drugs, averaged to 1.78 (Fig. 2c and Fig. S3). Conversely, while methods involving direct 3D generation can sometimes create compounds with superior poses within binding pockets, these compounds often feature multiple fused rings (Fig. 2c-d). Prior research indicates that a higher number of fused rings may lead to lower SAS [46–48], potentially accounting for the subpar SAS scores of other methods. Moreover, a high count of fused rings is linked with increased cellular toxicity and decreased developability [48, 49]. In line with this understanding, compounds generated by TamGen display a higher similarity score to FDA-approved drugs (Fig. S4). We hypothesize that pre-training on natural compounds and employing a sequence-based generation strategy enhance the overall plausibility of compounds produced by TamGen.

TamGen also achieves the best efficiency compared to alternate methods (Fig. S5). We benchmarked the wall time to generate 100 compounds for each target of all methods using one A6000 GPU. Other methods required tens of minutes or hours to complete this task, while TamGen was able to accomplish the task in an average time of just 9 seconds. This makes TamGen 85, 154, 213 and 394 times faster than ResGen, TargetDiff, Pocket2Mol and 3D-AR, respectively.

Collectively, our results suggest that TamGen is both effective and efficient in generating novel compounds. This positions TamGen as a valuable asset for quickly identifying hit compounds for downstream development.

2.3 TamGen designs novel inhibitors targeting Tuberculosis ClpP protease

We next employed TamGen to design small-molecule inhibitors against ClpP. As mentioned, ClpP plays essential roles in maintaining bacterial homeostasis, rendering it a promising antibiotic target.

Apart from the previously identified Bortezomib, a peptidomimetic compound that targets the human 26S proteasome and exhibits inhibitory activity against bacterial ClpP [50, 51], there are currently no documented advanced antibiotic ClpP inhibitors. Therefore, we leverage TamGen to generate compounds targeting ClpP in *Mycobacterium tuberculosis* (Mtb), a pathogenic bacteria in urgent need for novel drug candidates.

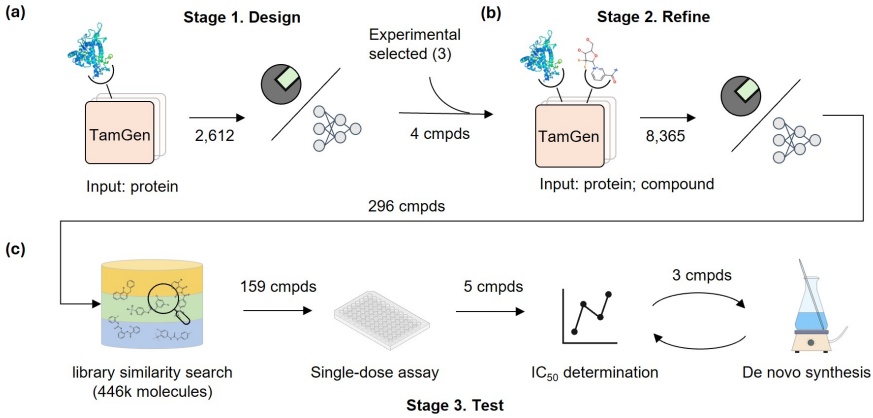


Fig. 3 Illustration of the Design-Refine-Test pipeline for Tuberculosis drug generation. (a) The Design stage. **(b)** The Refine stage. **(c)** The Test stage

We adopted a Design-Refine-Test pipeline driven by TamGen to identify potential ClpP inhibitors (Fig. 3). During the Design stage (Fig. 3a), utilizing the binding pocket of ClpP derived from protein structures (PDB ID 5DZK, and a ClpP-Bortezomib cocrystal structure (unpublished)), TamGen generated 2,612 unique compounds.

These compounds were then screened using molecular docking and Ligandformer, an AI model for phenotypic activity prediction [52] (see Methods for details). At this stage, we eliminated the compounds with worse docking scores compared to Bortezomib and inactive compounds predicted by Ligandformer. Peptidomimetic compounds were also excluded due to their suboptimal ADME properties (which is a known drawback of Bortezomib [53]). Finally, we identified 4 seeding compounds (green squares in Fig. 4a and Fig. S6) for the following Refine stage.

In the Refine stage, TamGen was applied to generate compounds conditioned on both the target protein and seeding compounds (Fig. 3b). Here, in addition to the 4 representative compounds generated by TamGen, we

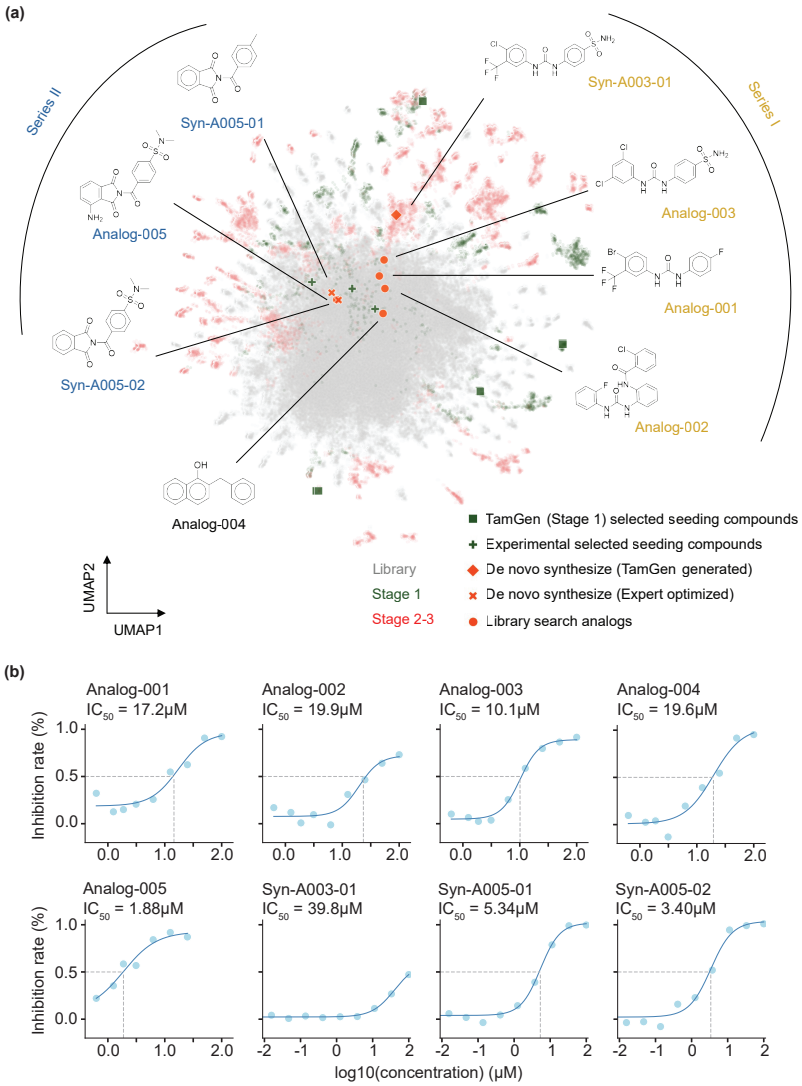


Fig. 4 Visualization and experimental validation on designed compounds. (a) UMAP visualization of library compounds and key compounds identified from the Design-Refine-Test pipeline with TamGen. Gray (background): 100K compounds sampled from library. Green (background): 2,612 compounds generated at Stage 1. Red (background): 8,365 compounds generated at Stage 2. Square and plus markers in green: seeding compounds used for Stage 2 generation. Circle, cross, and diamond markers in orange red: compounds subjected to IC_{50} determinations, stratified into 3 clusters based on molecular scaffold groups. **(b)** Dose-response assays for eight compounds with DMSO as a control. See methods for details of curve fitting and IC_{50} determination.

included 3 compounds with weak inhibitory activities identified from previous experiments (IC_{50} in 100 µM - 200 µM against Mtb ClpP. Fig. S6). Conditioned on the ClpP and these 7 seeding compounds, we generated 8,635 unique compounds using TamGen, and screened the compounds following the same procedure as in the Design stage. Finally, 296 of these generated compounds were selected for the Test (biological assay) stage.

We proceeded to compare the generated compounds with molecules from existing chemical libraries. Using UMAP visualization (Fig. 4a, Methods), we observe that compounds generated by TamGen are distinguishable from those in compound libraries. This indicates that TamGen is capable of exploring untapped chemical spaces when generating potential compounds conditioned on ClpP. Moreover, the compounds generated in the Refine stage showed superior docking scores and more dispersed patterns (an indicative of molecular diversity) compared to those from the Design stage (Fig. S7). This improvement shows that a Design-Refine generation approach can effectively enhance the desired properties of the candidate pool.

2.4 TamGen-driven drug design yields effective inhibitors against Tuberculosis ClpP protease

To expedite the validation process and enhance the efficiency during the Test stage, we first sought commercially available compounds structurally akin to those generated by TamGen (Fig. 3c). From a 446k commercial compound library, we successfully pinpointed 159 analogues with Maximum Common Substructure (MCS) similarity scores exceeding 0.55 in comparison to any of the 296 selected TamGen compounds. Five of these analogue compounds displayed significant inhibitory effects in the ClpP1P2 peptidase activity assay, with Bortezomib serving as a positive control (Fig. S8). Subsequent dose-response experiments revealed IC_{50} values below 20 µM for all five compounds, with Analog-005 standing out with an IC_{50} of 1.9 µM(Fig. 4b). Notably, none of these compounds have been previously documented as ClpP inhibitors, implying that TamGen may have revealed novel candidates for the treatment of Tuberculosis.

To explore the structure-activity relationship (SAR) and expand the pool of hit compounds, we synthesized three novel compounds absent from the commercial library. Considering that Analog-003 exhibited the strongest inhibitory effect in the peptidase activity assay (48% of Bortezomib, Fig. S8), we first synthesized its corresponding source compound generated by TamGen, referred to as Syn-A003-01 (Fig. 4a). Both compounds, along with Analog-001 and Analog-002, share a diphenylurea core (Series I in Fig. 4a), representing a novel scaffold for ClpP inhibitors. Interestingly, single-dose assay showed that replacing trifluoromethyl with chlorine greatly improved inhibitory activity of the compound (Fig. 4b). We reason that the replacement may have altered the charge distribution of the adjacent urine group in the compound, thereby influencing its hydrogen bonding effects. In addition, substituting sulfonamide

group with fluorine also moderately improved the activity. Secondly, we synthesized two derivatives of Analog-005, the compound with the most favorable IC_{50} (Fig. 4a, SeriesII). Similar inhibition efficiency was observed in these two derivatives and Analog-005 (Fig. 4b). This result suggests a marginal contribution to the overall activity from the modified groups and provides the starting point for further modifications. Collectively, out of the eight compounds generated or inspired by TamGen, seven displayed noteworthy IC_{50} values. The high confirmation rate of TamGen-driven drug design also highlights an alternative application of generative models, specifically employing the newly generated molecules as anchors for a more effective and efficient library search. This approach allows us to alleviate the cost in screening process and surmount the challenges posed by the validation and application of novel molecule synthesis in generative methods.

2.5 Structural insights on the mechanisms of compound binding

To investigate the inhibitor binding mechanism, we analyzed the docking poses of two representative compounds, Syn-A003-01 (from Series I) and Analog-005 (from Series II). These two compounds were docked to ClpP structure (PDB ID: 5DZK, see Methods for details) (Fig. 5). For comparison, the binding pose of Bortezomib, derived from an unpublished cocrystal structure, was also aligned into the same crystal structure of ClpP. Similar to Bortezomib, both Analog-005 and Syn-A003-01 maintain multiple hydrogen bonding interactions with ClpP1 (a subunit of ClpP). Meanwhile, the docked pose of Analog-005 suggests that the carbonyl carbon possibly forms a covalent bond with the catalytic residue Ser98, as indicated by both the chemical mechanism and docked complex structural model. This is in accordance with the binding pose of Bortezomib, providing plausible explanation of Analog-005's strong inhibitory activity. Interestingly, the complex structures also reveal that the sulfonamide groups of Analog-005 and Syn-A003-01 extend towards a deep pocket formed by residues Glu101, Phe102, Met150 and Asn154, a feature not observed for Bortezomib. The sulfonamide group may contribute to the binding to ClpP.

Altogether, through the Design-Refine-Test process powered by TamGen, we identified compounds that interact with ClpP protein in distinct modes from that of Bortezomib, thereby unveiling novel mechanisms for future ClpP inhibitor discovery. These compounds possess benzenesulfonamide and diphenylurea groups as scaffolds, which are completely different from the peptidomimetic Bortezomib, providing a possible solution to improve bioavailability and molecular stability of ClpP inhibitors. To sum up, the novelty and strong inhibitory efficacy of these compounds show potential for further development. The success of generating ClpP inhibitory compounds underscores the immense promise of TamGen in designing novel drug candidates and addressing drug-resistant Tuberculosis, implying its broad applications in drug design to treat other diseases.

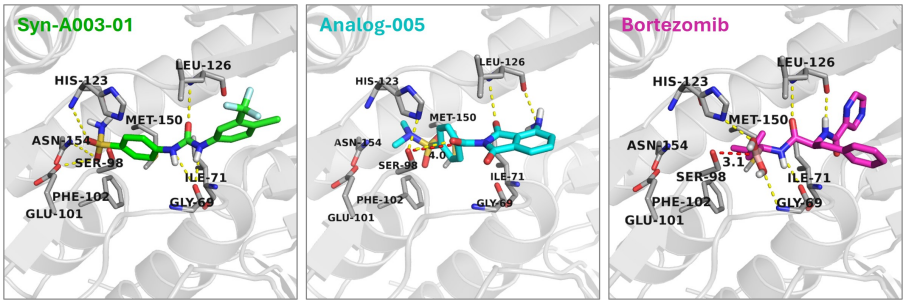


Fig. 5 Proposed binding modes of Syn-A003-01, Analog-005, and Bortezomib against ClpP. ClpP complex 5DZK is presented in grey cartoon. Syn-A003-01, Analog-005, and the reference compound Bortezomib are shown in green, cyan, and magenta sticks, respectively. The yellow dashed lines indicate hydrogen bonds. The red dashed lines with numbers denote distances between atoms.

3 Discussion and conclusions

Designing compounds that have high binding affinity to given pathogenic protein targets can speed up drug discovery process. It has been highly desirable to generate compounds based on target information and many efforts have been made to develop generative AI models to solve this challenging problem. However, few attempts have demonstrated success in real-world application. Here, we present the method, TamGen, not only achieved state-of-the-art performance in benchmark testing, but also discovered several compounds with high inhibition activities against ClpP protease of Mtb, the causative pathogen of infectious tuberculosis disease.

The success of TamGen is attributed to two major factors: (1) Chemical knowledge information embedded in the pre-trained compound decoder model, which enables the generation of high quality compounds that follow chemistry rules to possess properties for drug developments. With an ablation study, we show that pre-training is essential for producing plausible chemical compounds (Fig. S9). (2) An effective binding pocket representation that correlates to chemical compound decoding. The information of target protein binding sites is used to direct compound generation. Furthermore, TamGen can be applied to refine hit compounds reported in the literature or identified in previous rounds to generate better compounds for given targets. These designs overcome the data scarcity caused by shortage of high quality drug-target complex structures, which are usually required to learn the interactions between drug compounds and protein targets. Testing results show that TamGen is capable of generating compounds with high diversity and drug likeliness properties, increasing chances of hitting compounds that can be synthesized and further developed into drugs. This is supported by the successful design of strong inhibitor compounds against Mtb ClpP target. In the ClpP inhibitor generation case, we adopted the Design-Refine-Test workflow to iteratively improve the generated compounds. The Refine stage can be repeated multiple times by

including inhibitors discovered in previous steps, so that TamGen can help further optimize the compounds and increase the chance of generating stronger inhibitors.

The pre-training of compound decoder using chemical compound information in the similar manner as GPT models is a core component of TamGen. This strategy helps overcome the data scarcity issue partially, yet, the generative AI model such as TamGen can still benefit from a larger training dataset composed of high quality target-ligand complex structures.Also, a pre-trained protein structure encoder can be applied to describe target pocket geometry information, which is currently represented using amino acid positions. Such a pre-trained model or other advanced representations for the pocket may improve generated compound qualities [54]. This is particularly important to improve the binding affinity, because the interaction information are embedded in complex structures. TamGen can be further improved to predict the compound properties, such as binding affinity, compound stability, synthesizability, and drug properties including ADME/T. As presented in this work, these properties were assessed by experts in medicinal chemistry using docking analysis and phenotypic prediction. As more 3D complex structural data along with the binding affinity or inhibition activities information become available for model training, TamGen can predict properties and rank generated compounds. Such automation will further accelerate the compound generation and facilitate experimental testing.

Generative AI models, such as TamGen, contribute to the drug discovery not only by speeding up the process, but also enable the exploration in larger chemical space beyond available compound libraries. It is expected that the information will accumulate at an accelerating pace, because the novel compounds generated by AI models will enrich the chemical knowledge once they are validated experimentally. These add-on information will in turn enhance future generative AI models. Furthermore, TamGen has demonstrated the capability of generating diverse compounds based on both binding pocket and seeding compounds. This capability enables compound refinement by providing candidates centered around the seeding compounds for follow-up research. The capability of TamGen is demonstrated in the TB drug design as an application. The same protocol can be immediately applied to design compounds for other target proteins, unleashing its power in facilitating drug discovery in general.

4 Methods

4.1 Details of TamGen

We describe the details about how to process the 3D structure input, the architectures of the protein encoder, the chemical language model, the contextual encoder and the training objective functions.

Preliminaries: Let $\mathbf{a} = (a_1, a_2, \cdots, a_N)$ and $\mathbf{r} = (r_1, r_2, \cdots, r_N)$ denote the amino acids and their 3D coordinates of a binding pocket respectively, where

N is the sequence length and $r_i \in \mathbb{R}^3$ is the centroid of amino acid i (i is an index to label the amino acids around the binding site). a_i is a one-hot vector like $(\cdots, 0, 0, 1, 0, \cdots)$, where the vector length is 20 (the number of possible amino acid types) and the only 1 locates at the position corresponding to the amino acid type. A binding pocket is denoted as $\mathbf{x} = (\mathbf{a}, \mathbf{r})$ and $[N] = \{1, 2, \cdots, N\}$. Let $\mathbf{y} = (y_1, y_2, \cdots, y_M)$ denote the SMILES string of the corresponding ligand/compound with a length M. Our goal is to learn a mapping from $\mathbf{x} = (\mathbf{a}, \mathbf{r})$ to $\mathbf{y}$.

Processing 3D input: The amino acid a_i $\forall i \in [N]$ is mapped to d-dimensional vectors via an embedding layer E_a. Following our previous exploration on modeling the 3D coordinates [55], the coordinate $r_i (i \in [N])$ is mapped to a d-dimensional vector via a linear mapping. Considering we can rotate and translate a binding pocket while its spatial semantic information should be preserved, we apply data augmentation to the coordinates. That is, in the input layer, for any $i \in [N]$,

$$h_i^{(0)} = E_a a_i + E_r \rho \left(r_i - \frac{1}{N} \sum_{j=1}^{N} r_j \right), \tag{1}$$

where (i) E_a and E_r are learnable matrices, and they are optimized during model training; (ii) ρ denotes a random roto-translation operation, and before using ρ, we center the coordinates to the origin. Thus we process the discrete input $\mathbf{x}$ into N continuous hidden representations $h_i^{(0)}$.

Protein encoder: The encoder stacks L identical blocks. The output of the l-th block, i.e., $h_i^{(l)}$, is fed into the $(l+1)$-th layer for further processing and obtain $h_i^{(l+1)}$ for any $i \in [N]$ and $l \in \{0\} \cup [L-1]$. Each block consists of an attention layer and an `FFN` layer, which is a two-layer feed-forward network as that in the original Transformer [23]. To model the spatial distances of amino acids, we propose a new type of distance-aware attention. Mathematically,

$$\begin{aligned} \tilde{h}_i^{(l+1)} &= \sum_{j=1}^{N} \alpha_j (W_v h_j^{(l)}), \\ \alpha_j &= \frac{\exp \hat{\alpha}_j}{\sum_{k=1}^{N} \exp \hat{\alpha}_k}, \\ \hat{\alpha}_j &= \exp\left(-\frac{\|r_i - r_j\|^2}{\tau} \right) (h_i^{(l)\top} W h_j^{(l)}), \end{aligned} \tag{2}$$

where W and W_v are parameters to be optimized, and τ is the temperature hyperparameter to control. After that, $\tilde{h}_i^{(l+1)}$ is processed by an FFN layer and obtain

$$h_i^{(l+1)} = \texttt{FFN}(\tilde{h}_i^{(l+1)}). \tag{3}$$

The output from the last block, i.e., $h_i^{(L)}$ $\forall i \in [N]$, is the eventual representations of $\mathbf{x}$ from the encoder.

The contextual encoder: To facilitate diverse generation, we follow the VAE framework and use a random variable z to control the diverse generation for the same input. Given a protein binding pocket $\mathbf{x}$, a compound $\mathbf{y}$ is sampled according to the distribution $p(\mathbf{y}|\mathbf{x}, z;\ \Theta)$. The contextual encoder (i.e., the VAE encoder) models the posterior distribution of z given a binding pocket $\mathbf{x}$ and the corresponding ligand $\mathbf{y}$. The input of VAE encoder is defined as follows:

$$h_i^{(0)} = \begin{cases} E_a a_i + E_r \rho \left(r_i - \frac{1}{N} \sum_{j=1}^{N} r_j \right), & i \leq N \\ E_y y_{i-N}, \quad i > N, \end{cases} \tag{4}$$

where E_y is the embedding of the SMILES. The VAE encoder follows the architecture of standard Transformer encoder [23], which uses the vanilla self-attention layer rather than the distance-aware version due to the non-availability of the 3D ligand information. The output from the last block, i.e., $h_i^{(L)}\ \forall i \in [N]$, is mapped to the mean μ_i and covariance matrix Σ_i of position i via linear mapping, which can be used for constructing $q(z|\mathbf{x}, \mathbf{y})$, by assuming $q(z|\mathbf{x}, \mathbf{y})$ is Gaussian. The ligand representations, i.e., $h_j^{(L)}\ j > N$, are not used to construct $q(z|\mathbf{x}, \mathbf{y})$.

Chemical language model: The chemical language model is exactly the same as that in [23], which consists of the self-attention layer and the FFN layer. We pre-train the decoder on $10M$ compounds randomly selected from PubChem (denoted as $\mathcal{D}_0$) using the following objective function:

$$\min - \sum_{y \in \mathcal{D}_0} \frac{1}{M_y} \sum_{i=1}^{M_y} \log P(y_i | y_{i-1}, y_{i-2}, \cdots, y_1), \tag{5}$$

where M_y is the length of y. The chemical language model is pre-trained on eight V100 GPUs for 200k steps.

After pre-training the chemical language model, the cross-attention module is introduced to the compound decoder as shown in Fig. 1(c) (top panel). It takes all $h_i^{(L)}$ as inputs. Under the VAE variant, during training and compound refinement, the inputs are $h_i^{(L)} + z_i'$, where z_i' is sampled from the distribution $q(z|\mathbf{x}, \mathbf{y})$ introduced above. During inference, the inputs are $h_i^{(L)} + z_i$ where z_i is randomly sampled from $\mathcal{N}(0, I)$.

Training: The training objective is to minimize the following function:

$$\min_{\Theta, q} \frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{D}} - \log P(\mathbf{y}|\mathbf{x}, z;\ \Theta) + \beta \mathcal{D}_{\text{kl}} \left(q(z|\mathbf{x}, \mathbf{y}) \| p(z) \right). \tag{6}$$

In Eqn.(6), $\mathcal{D}$ is the training corpus, a collection of (pocket, SMILES) pairs; z in $\log P(\cdots)$ is sampled from $q(z|\mathbf{x}, \mathbf{y})$; β is a hyperparameter; $p(z)$ denotes the standard Gaussian distribution; $\mathcal{D}_{\text{kl}}$ denotes the KL divergence; Θ denotes the parameter

Implementation details For the results in Section 2.2, for fair comparison with the previous methods like Pocket2Mol [14], Targetdiff [12], we use the same data as them. The data is filtered from CrossDocked [40] and there are about $100k$ target-ligand pairs. For inference, the z is sampled from multivariant standard Gaussian distribution. Both the pocket encoder and VAE encoder have 4 layers with hidden dimension 256. The decoder has 12 layers with hidden dimension 768. We use Adam optimizer [56] with initial learning 3×10^{-5}. In the context of generating the compound database for Tuberculosis (TB), the current methodology incorporates an augmented dataset that includes the CrossDocked database and the Protein Data Bank (PDB), cumulatively accounting for approximately 300,000 protein-ligand pairs. To elaborate, this process involved the extraction of pocket-ligand pairs from about 72,000 PDB files. A pocket is defined on the basis of spatial proximity criteria: if any atom of an amino acid is less than 10Å away from any atom of the ligand, the corresponding amino acid is taken as part of the pocket.

4.2 The phenotype screening predictor Ligandformer

We utilized an adapted version of the Graph Neural Network (GNN) model as proposed in [52] to predict potential phenotypic activity. Compared with traditional GNNs, our model is designed such that the output from one layer is propagated to all subsequent layers for enhanced processing. We implemented a 5-layer architecture. Our phenotypic predictor was trained using a dataset of 18,886 samples, which are gathered from a variety of sources including ChEMBL, published datasets, and academic literature as compiled by [57]. At the inference stage, we interpreted an output value exceeding 0.69 (a threshold determined based on validation performance) as indicative of a positive sample.

4.3 Baselines and evaluations

4.3.1 Baselines

We mainly compare our method with the following baselines:

1. 3D-AR [38], a representative deep learning baseline that uses a graph neural network to encode the 3D pocket information and direct generates the 3D conformation of candidate drugs. The atom type and coordinates are generated sequentially. 3D-AR does not explicitly generate the position of the next, by use MCMC for generation.
2. Pocket2Mol [14] is an improved version of 3D-AR, which has specific modules to predict atom type, coordinate positions and bond type.
3. ResGen [39] is also an autoregressive method of generating compounds in 3D space directly. Compared with Pocket2Mol, ResGen uses residue-level encoding while Pocket2Mol uses atomic-level encoding.
4. TargetDiff [12] utilizes diffusion models to generate compounds. Compared with the previous method, all atom types and coordinates are

generated simultaneously, and iteratively refined until obtaining a stable conformation.

4.3.2 TamGen without pre-training

To assess the impact of pre-training, we introduce a TamGen version without pre-training, in which the compound generator is initialized randomly. We observed overfitting when a 12-layer chemical language model was used in the non pre-trained version. Upon evaluating layers 4, 6, 8, and 12 based on their validation performance, we discovered that a model with 4 layers yielded the most optimal results.

4.3.3 Mean Reciprocal Rank (MRR)

Mean Reciprocal Rank (MRR) calculation [58] is a widely used method to evaluate a method across different metrics. To elaborate, denote the rank of a method on metric i as r_i. The MRR for a particular method is hence defined as $\frac{1}{N}\sum_{i=1}^{N}\frac{1}{r_i}$, where N represents the total number of evaluation metrics being considered.

4.3.4 Fused rings

In this work, *fused rings* denote a structural element in compounds where two or more ring structures share at least one common bond. The size of the largest group of these "fused" rings within a molecule is denoted as the number of fused rings. In Fig. 2(d), from left to right, the number of fused rings of the four compounds are 2, 5, 4 and 4 respectively.

4.4 Experimental details

4.4.1 Peptidase activity assay

ClpP1P2 complex in Mtb can catalyse the hydrolysis of small peptides. Following previous protocols, we measure the in vitro inhibition of ClpP peptidase activity by monitoring the cleavage of fluorogenic peptide Ac-Pro-Lys-Met-AMC [59–61].

0.4 µL of candidate inhibitors, Bortezomib, or DMSO control are added into a black flat bottom 384-well plate by Echo®20 Liquid Handler and mixed with 20 µL enzyme buffer (The final ClpP1P2 dimer concentration is 50nM; reaction buffers: PIPES 30mM (pH 7.5), NaCl: 200mM and 0.005% Tween20). The solution is pre-incubated at room temperature for 2 hours. Then, 20 µL substrate buffer with Ac-Pro-Lys-Met-AMC is added (final concentration of Ac-Pro-Lys-Met-AMC is 10 µM; reaction buffer is the same with the above). Fluorescence (Ex/Em: 380/ 440 nm) is recorded for 120 min at 37℃.

4.4.2 Single-dose response measurement

Inhibition rates of compounds were determined by Relative Fluorescence Units (RFU) compared with Bortezomib control [62, 63] and DMSO control, which

is defined as follows:

$$\text{Inhibition Rate} = \frac{\text{RFU(test)} - \text{RFU(DMSO)}}{\text{RFU(bortezomib)} - \text{RFU(DMSO)}} \times 100\%. \tag{7}$$

In this case, fluorescence of DMSO is seen as none inhibition (0%), and fluorescence of Bortezomib is seen as completed inhibition (100%). Compounds with inhibition rates more than 20 % at 20 µM are considered as hits.

4.4.3 Dose-response assay and IC_{50} determination

To determine IC_{50}, candidate inhibitors are assayed at 9 or 10 gradient concentrations. A series of candidate inhibitor, Bortezomib, or DMSO dilutions is prepared starting from a maximum concentration of 100 µM, with each subsequent concentration being half or one third of the previous one (2-fold or 3-fold dilution gradient). IC_{50} is determined by the change of recorded fluorescence (as RFU) and gradient dilution of inhibitors concentration. Non-linear fit (log(inhibitor) vs. normalized response) is used for IC_{50} curve fitting.

4.5 Compound generation in Design and Refine stages for ClpP

4.5.1 Compound generation

Given a complex crystal structure with a protein receptor and a ligand, the center of the ligand is denoted as c. For each residue i of a protein, if its centroid p_i satisfies the condition $\|c - p_i\| \leq \tau$, i.e., within a distance cutoff τ from the ligand center c, then residue i is included in the pocket, where the distance cutoff τ is pre-defined.

In the case of ClpP complex, we first designed compounds based on published complex structure (PDB 5DZK) and our co-crystalized Bortezomib-ClpP structure. We took two values of τ to be 10 Å and 15 Å. Multiple binding sites can be extracted. We used beam search with beam size 20 to generate compounds. The β of the VAE was set to be 0.1 or 1. We initialized compound generation with 20 unique random seeds, ranging from 1 to 20. After removing duplicate and invalid generated compounds, we obtained 2.6k unique compounds.

During the following Refine stage, in addition to the binding pocket information, we included guiding information encoded in 4 representative compounds and 3 experimentally discovered compounds exhibiting weak inhibition activities. The parameter τ was set to 10 Å, 12 Å, and 15 Å. We used beam search with beam sizes of 4, 10, and 20 for compound generation. The β parameter of the VAE was set to 0.1 or 1. We initiated compound generation with 100 unique random seeds, ranging from 1 to 100. After removing duplicates and invalid compounds, we obtained a total of 8.4k unique compounds.

4.5.2 UMAP visualization

Compounds are converted to 1024-dimensional vectors with function `GetMorganFingerprintAsBitVect` from `rdkit`. UMAP transformation [64] is performed with parameters: `n_neighbors=20, min_dist=0.7, metric=sokal_michener`.

4.6 Ligand docking to protein target

The SMILES of generated compounds were converted to 3D structures with Open Babel program. Subsequently, AutoDock Tools was employed to add hydrogens and assign the Gasteiger charge to both the converted 3D compounds and the RCSB downloaded protein 5DZK before the docking process. The 5DZK ligand-centered maps were defined by the program AutoGrid and grid box was generated with definitions of $20 \times 20 \times 20$ points and 1 Å spacing. Molecular docking was performed with AutoDock Vina program with default settings. The predicted binding poses were visualized using the PyMol program.

Acknowledgments. We thank Dr. Nathan Baker, Dr. Christopher M. Bishop, Dr. Sheng Ding, Dr. Marwin Segler, Dr. Ryota Tomioka and Dr. Rumin Zhang for their insightful discussions and feedback.

References

[1] Wu, K., Xia, Y., Fan, Y., Deng, P., Liu, H., Wu, L., Xie, S., Wang, T., Qin, T., Liu, T.-Y.: Tailoring Molecules for Protein Pockets: a Transformer-based Generative Solution for Structured-based Drug Design (2022)

[2] Schneider, G., Fechner, U.: Computer-based de novo design of drug-like molecules. Nature Reviews Drug Discovery **4**(8), 649–663 (2005)

[3] Wang, M., Wang, Z., Sun, H., Wang, J., Shen, C., Weng, G., Chai, X., Li, H., Cao, D., Hou, T.: Deep learning approaches for de novo drug design: An overview. Current Opinion in Structural Biology **72**, 135–144 (2022)

[4] Liu, G., Catacutan, D.B., Rathod, K., Swanson, K., Jin, W., Mohammed, J.C., Chiappino-Pepe, A., Syed, S.A., Fragis, M., Rachwalski, K., Magolan, J., Surette, M.G., Coombes, B.K., Jaakkola, T., Barzilay, R., Collins, J.J., Stokes, J.M.: Deep learning-guided discovery of an antibiotic targeting acinetobacter baumannii. Nature Chemical Biology **19**(11), 1342–1350 (2023). https://doi.org/10.1038/s41589-023-01349-8

[5] Stokes, J.M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N.M., MacNair, C.R., French, S., Carfrae, L.A., Bloom-Ackermann, Z., Tran, V.M., Chiappino-Pepe, A., Badran, A.H., Andrews, I.W., Chory, E.J., Church, G.M., Brown, E.D., Jaakkola, T.S., Barzilay, R., Collins,

J.J.: A deep learning approach to antibiotic discovery. Cell **180**(4), 688–70213 (2020)

[6] Wong, F., Zheng, E.J., Valeri, J.A., Donghia, N.M., Anahtar, M.N., Omori, S., Li, A., Cubillos-Ruiz, A., Krishnan, A., Jin, W., Manson, A.L., Friedrichs, J., Helbig, R., Hajian, B., Fiejtek, D.K., Wagner, F.F., Soutter, H.H., Earl, A.M., Stokes, J.M., Renner, L.D., Collins, J.J.: Discovery of a structural class of antibiotics with explainable deep learning. Nature (2023). https://doi.org/10.1038/s41586-023-06887-8

[7] Stanley, M., Segler, M.: Fake it until you make it? generative de novo design and virtual screening of synthesizable molecules. Current Opinion in Structural Biology **82**, 102658 (2023). https://doi.org/10.1016/j.sbi.2023.102658

[8] Corsello, S.M., Bittker, J.A., Liu, Z., Gould, J., McCarren, P., Hirschman, J.E., Johnston, S.E., Vrcic, A., Wong, B., Khan, M., Asiedu, J., Narayan, R., Mader, C.C., Subramanian, A., Golub, T.R.: The drug repurposing hub: a next-generation drug library and information resource. Nature Medicine **23**(4), 405–408 (2017). https://doi.org/10.1038/nm.4306

[9] Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B., Zaslavsky, L., Zhang, J., Bolton, E.E.: PubChem 2023 update. Nucleic Acids Research **51**(D1), 1373–1380 (2022) https://arxiv.org/abs/https://academic.oup.com/nar/article-pdf/51/D1/D1373/48441598/gkac956.pdf. https://doi.org/10.1093/nar/gkac956

[10] Irwin, J.J., Shoichet, B.K.: ZINC–a free database of commercially available compounds for virtual screening. J. Chem. Inf. Model. **45**(1), 177–182 (2005)

[11] Reymond, J.-L.: The chemical space project. Accounts of Chemical Research **48**(3), 722–730 (2015). https://doi.org/10.1021/ar500432k

[12] Guan, J., Qian, W.W., Peng, X., Su, Y., Peng, J., Ma, J.: 3d equivariant diffusion for target-aware molecule generation and affinity prediction. The Eleventh International Conference on Learning Representations (2023)

[13] Zhang, O., Zhang, J., Jin, J., Zhang, X., Hu, R., Shen, C., Cao, H., Du, H., Kang, Y., Deng, Y., Liu, F., Chen, G., Hsieh, C.-Y., Hou, T.: Resgen is a pocket-aware 3d molecular generation model based on parallel multiscale modelling. Nature Machine Intelligence **5**, 1020–1030 (2023). https://doi.org/10.1038/s42256-023-00712-7. Accessed 2023-12-05

[14] Peng, X., Luo, S., Guan, J., Xie, Q., Peng, J., Ma, J.: Pocket2mol: Efficient molecular sampling based on 3d protein pockets. International Conference

on Machine Learning (2022)

[15] Chenthamarakshan, V., Hoffman, S.C., Owen, C.D., Lukacik, P., Strain-Damerell, C., Fearon, D., Malla, T.R., Tumber, A., Schofield, C.J., Duyvesteyn, H.M.E., Dejnirattisai, W., Carrique, L., Walter, T.S., Screaton, G.R., Matviiuk, T., Mojsilovic, A., Crain, J., Walsh, M.A., Stuart, D.I., Das, P.: Accelerating drug target inhibitor discovery with a deep generative foundation model. Science Advances **9**(25), 7865 (2023) https://arxiv.org/abs/https://www.science.org/doi/pdf/10.1126/sciadv.adg7865. https://doi.org/10.1126/sciadv.adg7865

[16] Choung, O.-H., Vianello, R., Segler, M., Stiefl, N., Jiménez-Luna, J.: Extracting medicinal chemistry intuition via preference machine learning. Nature Communications **14**(1), 6651 (2023). https://doi.org/10.1038/s41467-023-42242-1

[17] Sanchez-Lengeling, B., Outeiral, C., Guimaraes, G.L., Aspuru-Guzik, A.: Optimizing distributions over molecular space. An Objective-Reinforced Generative Adversarial Network for Inverse-design Chemistry (ORGANIC) (2023). https://chemrxiv.org/engage/chemrxiv/article-details/60c73d91702a9beea7189bc2

[18] Segler, M.H.S., Kogej, T., Tyrchan, C., Waller, M.P.: Generating focused molecule libraries for drug discovery with recurrent neural networks. ACS Central Science **4**(1), 120–131 (2018). https://doi.org/10.1021/acscentsci.7b00512. PMID: 29392184

[19] Prykhodko, O., Johansson, S.V., Kotsias, P.-C., Arús-Pous, J., Bjerrum, E.J., Engkvist, O., Chen, H.: A de novo molecular generation method using latent vector based generative adversarial network. J. Cheminform. **11**(1), 74 (2019)

[20] Skalic, M., Jiménez, J., Sabbadin, D., De Fabritiis, G.: Shape-based generative modeling for de novo drug design. Journal of Chemical Information and Modeling **59**(3), 1205–1214 (2019). https://doi.org/10.1021/acs.jcim.8b00706. PMID: 30762364

[21] Schneider, P., Walters, W.P., Plowright, A.T., Sieroka, N., Listgarten, J., Goodnow Jr, R.A., Fisher, J., Jansen, J.M., Duca, J.S., Rush, T.S., *et al.*: Rethinking drug design in the artificial intelligence era. Nature Reviews Drug Discovery **19**(5), 353–364 (2020)

[22] OpenAI: GPT-4 Technical Report (2023)

[23] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems, 5998–6008 (2017)

[24] OpenAI: GPT-4V(ision) System Card (2023). https://cdn.openai.com/papers/GPTV_System_Card.pdf

[25] Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. ICML'23 (2023)

[26] AI4Science, M.R., Quantum, M.A.: The Impact of Large Language Models on Scientific Discovery: a Preliminary Study using GPT-4 (2023)

[27] Weininger, D.: Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. Journal of Chemical Information and Computer Sciences **28**(1), 31–36 (1988). https://doi.org/10.1021/ci00057a005

[28] Organization, W.H.: Fact sheets of Tuberculosis from WHO (2023). https://www.who.int/news-room/fact-sheets/detail/tuberculosis

[29] Dartois, V.A., Rubin, E.J.: Anti-tuberculosis treatment strategies and drug development: challenges and priorities. Nature Reviews Microbiology **20**(11), 685–701 (2022)

[30] Organization, W.H.: Global tuberculosis report 2023 (2023). https://www.who.int/publications/i/item/9789240083851

[31] Waller, N.J., Cheung, C.-Y., Cook, G.M., McNeil, M.B.: The evolution of antibiotic resistance is associated with collateral drug phenotypes in mycobacterium tuberculosis. Nature Communications **14**(1), 1517 (2023)

[32] d'Andrea, F.B., Poulton, N.C., Froom, R., Tam, K., Campbell, E.A., Rock, J.M.: The essential ¡i¿m. tuberculosis¡/i¿ clp protease is functionally asymmetric in vivo. Science Advances **8**(18), 7943 (2022) https://arxiv.org/abs/https://www.science.org/doi/pdf/10.1126/sciadv.abn7943. https://doi.org/10.1126/sciadv.abn7943

[33] Culp, E., Wright, G.D.: Bacterial proteases, untapped antimicrobial drug targets. The Journal of Antibiotics **70**(4), 366–377 (2017). https://doi.org/10.1038/ja.2016.138

[34] Maia, E.H.B., Assis, L.C., De Oliveira, T.A., Da Silva, A.M., Taranto, A.G.: Structure-based virtual screening: from classical to artificial intelligence. Frontiers in chemistry **8**, 343 (2020)

[35] Benaroudj, N., Raynal, B., Miot, M., Ortiz-Lombardia, M.: Assembly and proteolytic processing of mycobacterial clpp1 and clpp2. BMC Biochemistry **12**(1), 61 (2011). https://doi.org/10.1186/1471-2091-12-61

[36] Kingma, D.P., Welling, M.: Auto-encoding variational bayes. International Conference on Learning Representations (2014)

[37] Masuda, T., Ragoza, M., Koes, D.R.: Generating 3d molecular structures conditional on a receptor binding site with deep generative models. arXiv preprint arXiv:2010.14442 (2020)

[38] Luo, S., Guan, J., Ma, J., Peng, J.: A 3d generative model for structure-based drug design. Advances in Neural Information Processing Systems **34** (2021)

[39] Zhang, O., Zhang, J., Jin, J., Zhang, X., Hu, R., Shen, C., Cao, H., Du, H., Kang, Y., Deng, Y., Liu, F., Chen, G., Hsieh, C.-Y., Hou, T.: Resgen is a pocket-aware 3d molecular generation model based on parallel multiscale modelling. Nature Machine Intelligence **5**(9), 1020–1030 (2023). https://doi.org/10.1038/s42256-023-00712-7

[40] Francoeur, P.G., Masuda, T., Sunseri, J., Jia, A., Iovanisci, R.B., Snyder, I., Koes, D.R.: Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. Journal of Chemical Information and Modeling **60**(9), 4200–4215 (2020)

[41] Trott, O., Olson, A.J.: 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)

[42] Bickerton, G.R., Paolini, G.V., Besnard, J., Muresan, S., Hopkins, A.L.: Quantifying the chemical beauty of drugs. Nature chemistry **4**(2), 90–98 (2012)

[43] Lipinski, C.A., Lombardo, F., Dominy, B.W., Feeney, P.J.: Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Advanced Drug Delivery Reviews **23**(1), 3–25 (1997). https://doi.org/10.1016/S0169-409X(96)00423-1. In Vitro Models for Selection of Development Candidates

[44] Ertl, P., Schuffenhauer, A.: Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. Journal of Cheminformatics **1**(1), 8 (2009). https://doi.org/10.1186/1758-2946-1-8

[45] Piccaro, G., Poce, G., Biava, M., Giannoni, F., Fattorini, L.: Activity of lipophilic and hydrophilic drugs against dormant and replicating mycobacterium tuberculosis. The Journal of Antibiotics **68**(11), 711–714 (2015). https://doi.org/10.1038/ja.2015.52

[46] Skoraczyński, G., Kitlas, M., Miasojedow, B., Gambin, A.: Critical assessment of synthetic accessibility scores in computer-assisted synthesis planning. Journal of Cheminformatics **15**(1), 6 (2023). https://doi.org/10.1186/s13321-023-00678-z

[47] Ertl, P., Schuffenhauer, A.: Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. Journal of cheminformatics **1**(1), 1–11 (2009)

[48] Peng, X., Guan, J., Liu, Q., Ma, J.: Moldiff: Addressing the atom-bond inconsistency problem in 3d molecule diffusion generation. ICML'23 (2023)

[49] Ritchie, T.J., Macdonald, S.J.F.: The impact of aromatic ring count on compound developability – are too many aromatic rings a liability in drug design? Drug Discovery Today **14**(21), 1011–1020 (2009). https://doi.org/10.1016/j.drudis.2009.07.014

[50] Moreira, W., Ngan, G.J., Low, J.L., Poulsen, A., Chia, B.C., Ang, M.J., Yap, A., Fulwood, J., Lakshmanan, U., Lim, J., *et al.*: Target mechanism-based whole-cell screening identifies bortezomib as an inhibitor of caseinolytic protease in mycobacteria. MBio **6**(3), 10–1128 (2015)

[51] Moreira, W., Santhanakrishnan, S., Dymock, B.W., Dick, T.: Bortezomib warhead-switch confers dual activity against mycobacterial caseinolytic protease and proteasome and selectivity against human proteasome. Frontiers in Microbiology **8**, 746 (2017)

[52] Guo, J., Liu, Q., Guo, H., Lu, X.: Ligandformer: A graph neural network for predicting compound property with robust interpretation. arXiv preprint arXiv:2202.10873 (2022)

[53] Coghi, P.S., Zhu, Y., Xie, H., Hosmane, N.S., Zhang, Y.: Organoboron compounds: Effective antibacterial and antiparasitic agents. Molecules **26**(11), 3309 (2021)

[54] Luo, S., Chen, T., Xu, Y., Zheng, S., Liu, T.-Y., Wang, L., He, D.: One transformer can understand both 2d and 3d molecular data. The Eleventh International Conference on Learning Representations (2023)

[55] Zhu, J., Xia, Y., Liu, C., Wu, L., Xie, S., Wang, Y., Wang, T., Qin, T., Zhou, W., Li, H., Liu, H., Liu, T.: Direct molecular conformation generation. Transactions on Machine Learning Research (2022)

[56] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations, ICLR 2015, San

Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)

[57] Lane, T., Russo, D.P., Zorn, K.M., Clark, A.M., Korotcov, A., Tkachenko, V., Reynolds, R.C., Perryman, A.L., Freundlich, J.S., Ekins, S.: Comparing and validating machine learning models for mycobacterium tuberculosis drug discovery. Mol. Pharm. **15**(10), 4346–4360 (2018)

[58] Radev, D.R., Qi, H., Wu, H., Fan, W.: Evaluating web-based question answering systems. Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02) (2002)

[59] Akopian, T., Kandror, O., Tsu, C., Lai, J.H., Wu, W., Liu, Y., Zhao, P., Park, A., Wolf, L., Dick, L.R., Rubin, E.J., Bachovchin, W., Goldberg, A.L.: Cleavage specificity of mycobacterium tuberculosis ClpP1P2 protease and identification of novel peptide substrates and boronate inhibitors with anti-bacterial activity. J. Biol. Chem. **290**(17), 11008–11020 (2015)

[60] Fraga, H., Rodriguez, B., Bardera, A., Cid, C., Akopian, T., Kandror, O., Park, A., Colmenarejo, G., Lelievre, J., Goldberg, A.: Development of high throughput screening methods for inhibitors of ClpC1P1P2 from mycobacteria tuberculosis. Anal. Biochem. **567**, 30–37 (2019)

[61] Li, M., Kandror, O., Akopian, T., Dharkar, P., Wlodawer, A., Maurizi, M.R., Goldberg, A.L.: Structure and functional properties of the active form of the proteolytic complex, ClpP1P2, from mycobacterium tuberculosis. J. Biol. Chem. **291**(14), 7465–7476 (2016)

[62] Hu, G., Lin, G., Wang, M., Dick, L., Xu, R.-M., Nathan, C., Li, H.: Structure of the mycobacterium tuberculosis proteasome and mechanism of inhibition by a peptidyl boronate. Mol. Microbiol. **59**(5), 1417–1428 (2006)

[63] Lin, G., Tsu, C., Dick, L., Zhou, X.K., Nathan, C.: Distinct specificities of mycobacterium tuberculosis and mammalian proteasomes for n-acetyl tripeptide substrates. J. Biol. Chem. **283**(49), 34423–34431 (2008)

[64] McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018)

[65] Preuer, K., Renz, P., Unterthiner, T., Hochreiter, S., Klambauer, G.: Fréchet chemnet distance: A metric for generative models for molecules in drug discovery. Journal of Chemical Information and Modeling **58**(9), 1736–1741 (2018). https://doi.org/10.1021/acs.jcim.8b00234. PMID: 30118593

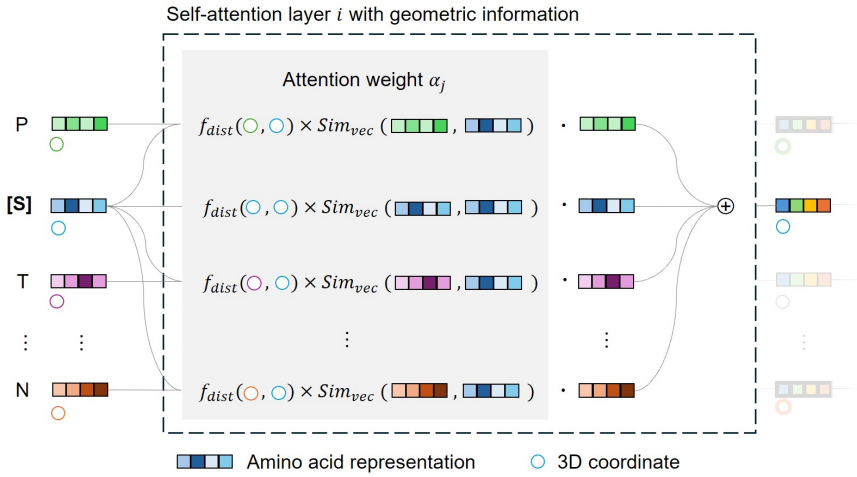


Fig. S1 Details of the self-attention mechanism with geometric information used in the protein encoder. For each amino acid representation in layer i, the attention weight α's is calculated as the product of the amino acid representation similarity and negative geometric distances between pairs of amino acids (i.e., $\exp(-\text{distances}^2/\tau)$ where τ is a hyperparameter). The output of layer i is then derived from the sum of the α's multiplied by the amino acid representation.

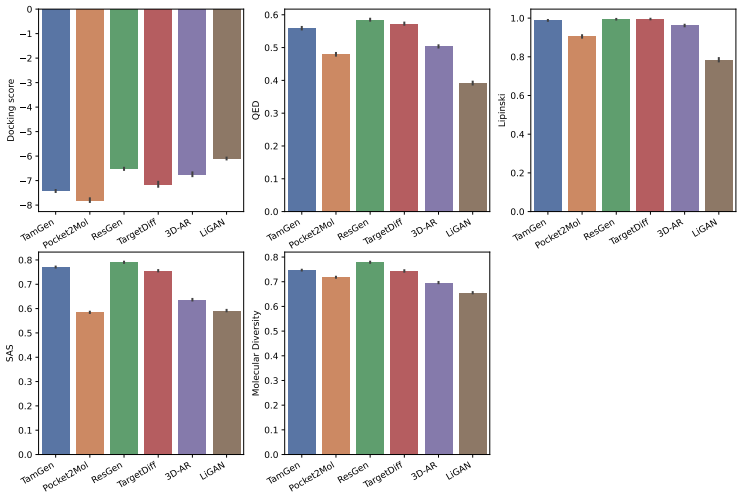


Fig. S2 Docking scores, QED, Lipinski, SAS, and Molecular Diversity of various generative drug design methods in relation to the CrossDocked2020 task. Error bar, 95% confidence interval.

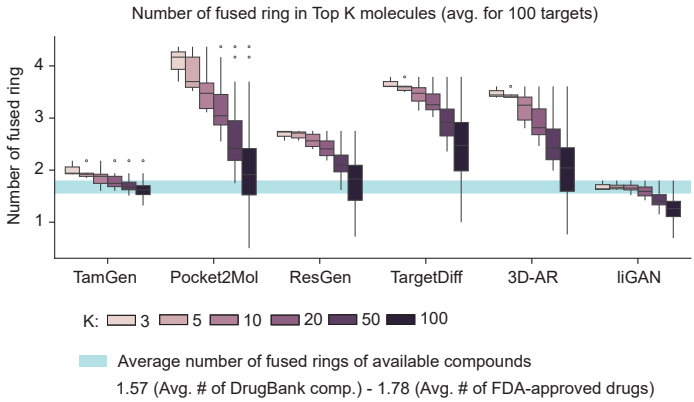


Fig. S3 The distribution of fused ring numbers in compounds generated by different methods. K represents the number of compounds having top-K docking scores against each target protein. Center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points, outliers.

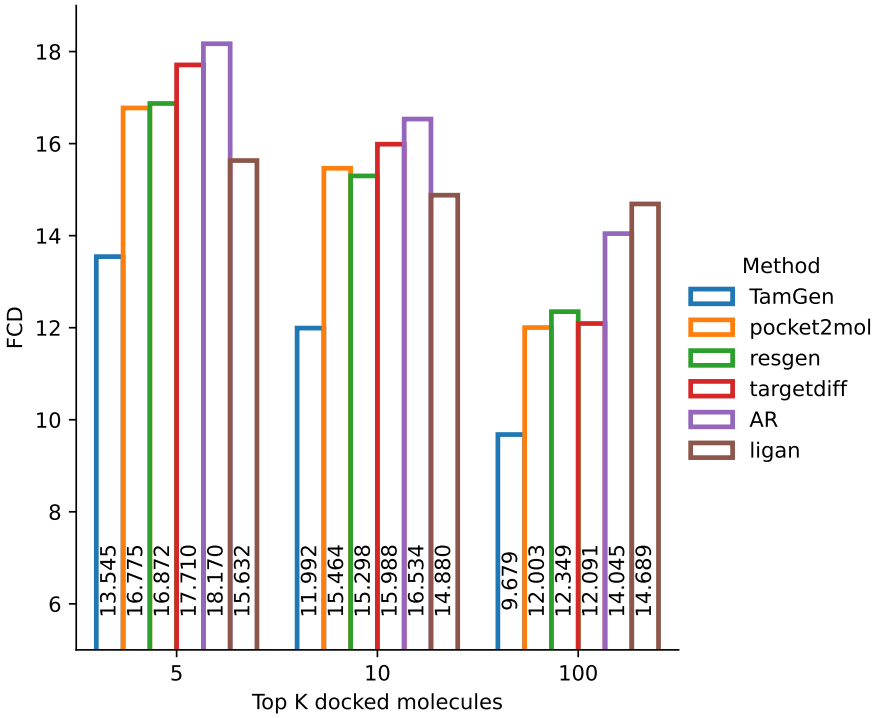


Fig. S4 The Fréchet ChemNet Distance (FCD) similarity [65] scores between FDA-approved drugs and compounds produced by different methods. FCD is a metric that quantifies the distributional dissimilarities between two compound sets, referred to as group A and group B. In this context, group A comprises all FDA-approved drugs, while group B includes compounds generated through various methods. A lower FCD score indicates a closer distribution of the generated compounds to the FDA approved drugs, signifying their similarity. TamGen demonstrates the capability to generate compounds that are most akin to FDA-approved drugs, as evidenced by the lowest FCD scores.

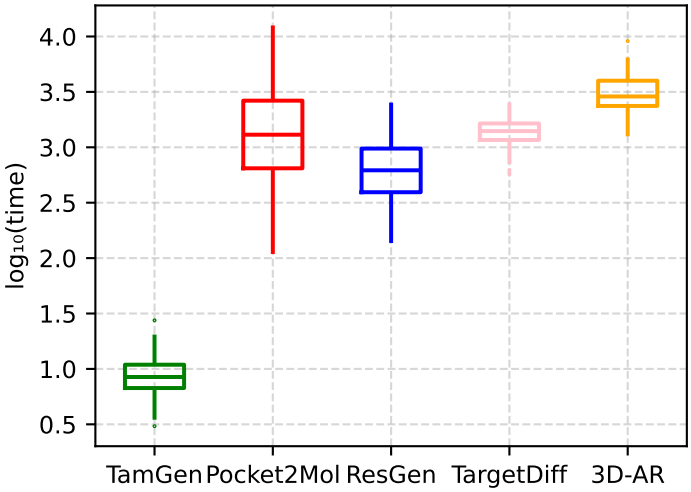


Fig. S5 TamGen significantly outperforms alternate methods on running time. The y-axis is scaled using a logarithm base 10.

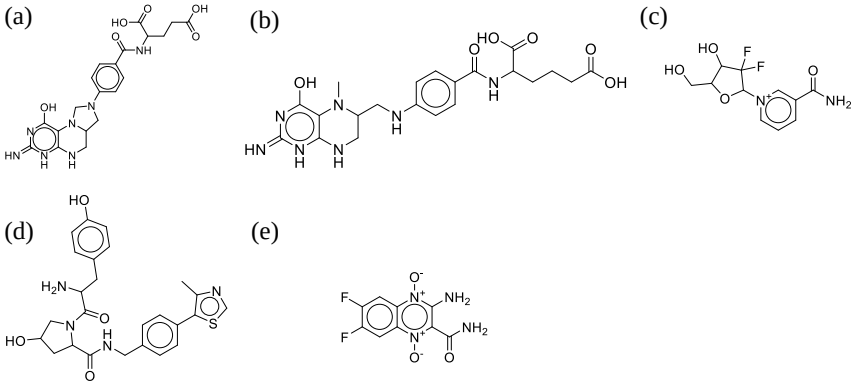


Fig. S6 Seeding compounds for Stage 2 generation. (a-d) The four seeding compounds selected from the first round; **(e)**: One example of the experimental selected compound.

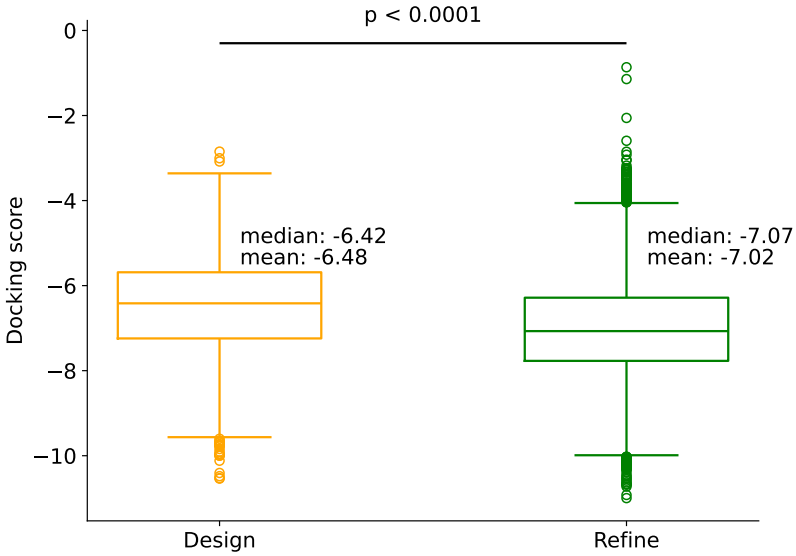


Fig. S7 Distribution of docking scores for generated compounds against ClpP. Center line, median; box limits, upper and lower quartiles. p-value is calculated with Mann–Whitney U test (`scipy.stats.mannwhitneyu`).

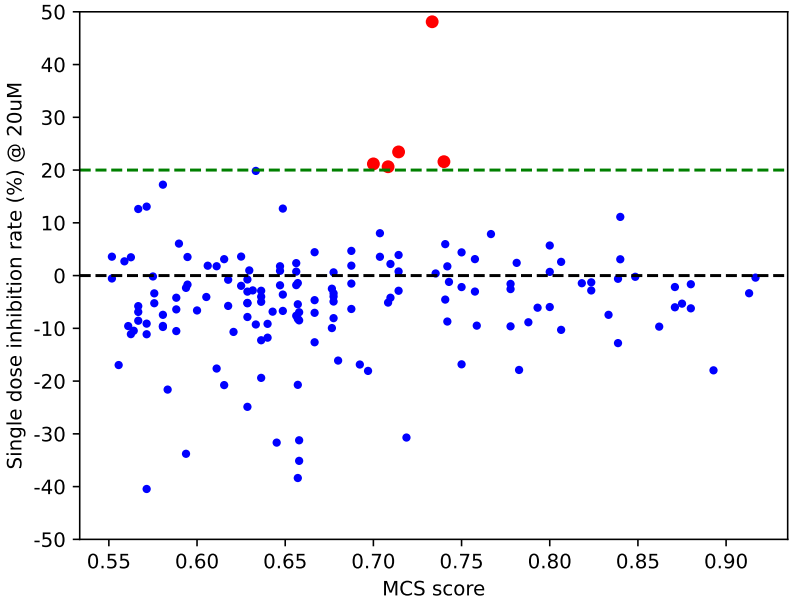
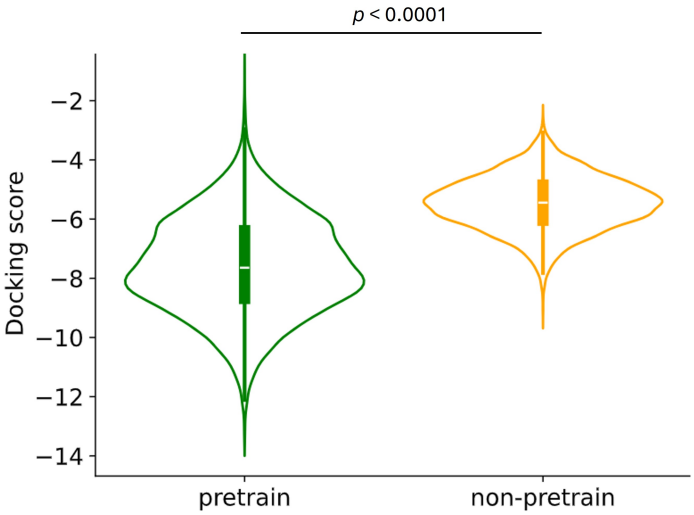
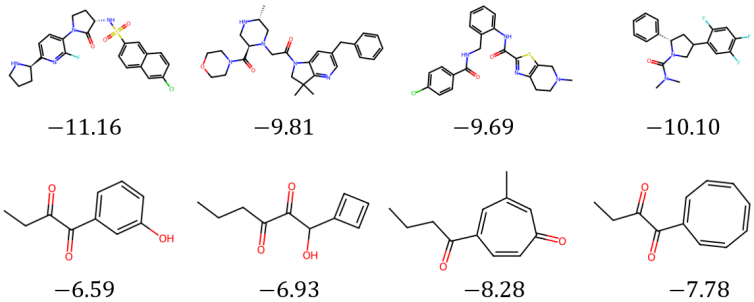


Fig. S8 Inhibition rate of the 159 library search analogs relative to Bortezomib. All compounds were evaluated at the concentration of 20 µM. The dashed line indicates the threshold for analog selection. x-axis: Maximum Common Substructure (MCS) similarity scores. See Methods for details.



(a) The violin plot illustrates the docking scores of pretrained and non-pretrained compound decoder. Pretrained decoder shows a significant improvement compared to non-pretrained decoder. p-value is calculated with Mann–Whitney U test (`scipy.stats.mannwhitneyu`).



(b) Case study of the generated compounds. The top/bottom rows are the compounds generated by pre-trained / non-pretrained compound generators respectively. The corresponding docking scores for each compound are displayed under their respective structures. Each column corresponds to the same target. The compounds are visualized using RDKit.

Fig. S9 Ablation study indicates that pre-training is essential for molecule generation of the compound decoder.

Model \ Metric	Vina Dock (↓)	QED (↑)	SAS (↑)	Diversity (↑)	LogP∈ [0, 5]	Lipinski (↑)
liGAN*	-6.099	0.392	0.592	0.655	55.0%	78.4%
3D-AR	-6.746	0.503	0.637	0.698	55.0%	96.2%
Pocket2Mol	-7.152	0.573	0.756	0.741	75.5%	99.5%
TargetDiff	-7.802	0.480	0.585	0.717	65.2%	90.5%
ResGen	-6.326	0.567	0.767	0.756	78.0%	96.5%
TamGen	-7.475	0.559	0.771	0.747	87.9%	98.8%

Table S1 Compilation of performance statistics for all methods across various evaluation metrics.

ID	PubChem	Commercial library source	IC_{50} (μM)
Analog-001	2810424	Maybridge Screening Collection	17.2
Analog-002	45503904	Life Chemicals HTS Compound Collection	19.9
Analog-003	2813477	Maybridge Screening Collection	10.1
Analog-004	160268	reframeDB	19.6
Analog-005	4851126	Selleck PFZ	1.9

Table S2 Resources of the analogue compounds. The index of the compounds, PubChem CID, Commercial library source and IC_{50} values are summarized.