nature computational science

Article

https://doi.org/10.1038/s43588-025-00850-5

Pharmacophore-oriented 3D molecular generation toward efficient feature-customized drug discovery

Received: 15 December 2024

Accepted: 15 July 2025

Published online: 26 August 2025

Check for updates

Jian Peng[1,3], Jun-Lin Yu[1,3], Zeng-Bao Yang[1,3], Yi-Ting Chen[1,3], Si-Qi Wei[1], Fan-Bo Meng[1], Yao-Geng Wang[1], Xiao-Tian Huang[1] & Guo-Bo Li[1,2]

Molecular generation is a cutting-edge technology with the potential to revolutionize intelligent drug discovery. However, currently reported ligand-based or structure-based molecular generation methods remain unpractical for real-world drug discovery. Here we propose an explicit pharmacophore-oriented 3D molecular generation method, termed PhoreGen. PhoreGen employs asynchronous perturbations and updates on both atomic and bond information, coupled with a message-passing mechanism that incorporates prior knowledge of ligand–pharmacophore mapping during the diffusion–denoising process. Evaluations revealed that PhoreGen efficiently generates 3D molecules well aligned with pharmacophores, maintaining good chemical reasonability, diversity, drug-likeness and binding affinity and, importantly, produces feature-customized molecules at high frequency. By using PhoreGen, we successfully identified new bicyclic boronate inhibitors of evolved metallo-β-lactamase and serine-β-lactamases, which potentiate meropenem against clinically isolated superbugs. Moreover, we identified inhibitors of metallo-nicotinamidases, emerging targets for insecticides. This work explores an explicitly constrained mode for molecular generation and demonstrates its potential in feature-customized drug discovery.

Drug discovery is resource intensive, costly and high risk, particularly in its early stages, where the challenge lies in identifying potential drug candidates from a vast chemical space[1,2]. This has fueled continuous innovation in drug discovery technologies with the goal of enhancing efficiency and reducing costs in the drug discovery process[3,4]. Recently, rapid advancements in deep generative models have led to substantial breakthroughs in the development of new technologies for drug discovery, such as molecular generation[5–7]. The main advantage of molecular generation is its ability to efficiently navigate chemical space and create novel chemical entities with potential pharmacological activity. This technology has the potential to accelerate drug discovery and usher in a new era of intelligence and automated drug design.

Current molecular generation methods can be broadly categorized into two main categories: ligand-based molecular generation (LBMG)[8–10] and structure-based molecular generation (SBMG)[11–17]. LBMG methods were developed earlier, such as GENTRL[18] and MolGAN[19], which usually utilize variational autoencoders or generative adversarial networks to model a chemical space containing potential active molecules and then perform interpolation within this space to generate new molecules. The issue with such methods is the

[1]Key Laboratory of Drug-Targeting and Drug Delivery System of the Education Ministry and Sichuan Province, Department of Medicinal Chemistry, West China School of Pharmacy, Sichuan University, Chengdu, China. [2]Children's Medicine Key Laboratory of Sichuan Province, Chengdu, China. [3]These authors contributed equally: Jian Peng, Jun-Lin Yu, Zeng-Bao Yang, Yi-Ting Chen. e-mail: liguobo@scu.edu.cn

lack of target structure information, making it difficult to generate target-specific molecules[20]. SBMG methods can explicitly incorporate the principles of target–ligand complementarity, facilitating the direct generation of three-dimensional (3D) molecules within the target's binding pocket. Many SBMG methods, such as AR[11], GraphBP[21] and Pocket2Mol[12], employed an autoregressive strategy for 3D molecular generation, which are generally capable of producing molecules with high chemical validity. Inspired by point cloud generation in computer vision[22–24], several SBMG methods were established recently using diffusion models, such as TargetDiff[13], DiffSBDD[25], DecompDiff[26] and IRDiff[27]. These models leverage both local and global information from the target to generate entire molecules, achieving improved performance in generating target-specific molecules[28]. Even so, current SBMG methods remain unpractical for real-world drug discovery and are particularly unsuitable for feature-customized application scenes such as covalent drug discovery[29]. Therefore, it is necessary to leap out of the current modes of molecular generation and develop more explicitly oriented approaches for 3D molecular generation.

Pharmacophores, as abstractions of essential chemical interaction patterns, can not only represent 3D structural information of both ligands and targets but also inherently encapsulate the principles of protein–ligand complementarity[30]. Currently, there are a few studies that utilize pharmacophores to assist in molecular generation. DEVELOP[31] employs 3D pharmacophores as constraints to assist structural optimization, encompassing tasks such as linker design and R-group modification. PGMG[32] employs pharmacophores as inputs to guide non-3D molecular generation by establishing latent variables to solve many-to-many mapping relations between pharmacophores and molecules. While these methods showcase the potential of the pharmacophore to enhance generation quality and offer effective guidance during the generation process, pharmacophore-based molecular generation (PBMG) methods that enable the direct de novo design of 3D molecules that are well aligned with the pharmacophore model are still lacking.

Here, we propose PhoreGen, a pharmacophore-oriented conditional diffusion model for 3D molecular generation (Fig. 1), which is enriched with chemical and pharmacophore mapping knowledge and powered by hierarchical data learning. In PhoreGen, we introduced a specific heterogeneous geometric graph to represent 3D ligand–pharmacophore pairs, which can uniquely capture the pharmacophore features including internal orientation. Prior knowledge of ligand–pharmacophore mapping is integrated into the network, guiding the molecular generation process through pairwise direction-matching encodings within the message-passing mechanism. The diffusion–denoising process uses the bond-first noise approach to coordinate atomic and bond information, largely mitigating the generation of unrealistic local structures and improper bonds that are common in methods relying solely on atomic information. The auxiliary module for molecular size prediction is based on geometric constraints from pharmacophore features and shape restrictions from exclusion spheres. In addition, we complied over one million ligand–pharmacophore pairs from energetically favorable 3D ligand structures, experimental protein–ligand complex structures and cross-docking predicted complex structures, for model training. Notably, we reinforced the low-frequency pharmacophore features (for instance, covalent and metal-coordination features) in ligand–pharmacophore pairs aiming to exert the unique advantage of PBMG in feature-customized drug design as well as in unlocking underexplored chemical spaces.

The comprehensive evaluations demonstrate its ability in generating 3D molecules that accurately match the given pharmacophore models, while maintaining reasonable 3D chemical structures, good drug-likeness properties and high binding potential. It also showed the ability to efficiently generate promising molecules with the required covalent or metal-coordination features at a high frequency, showcasing a high level of feature-customized molecular design that no existing molecular generation method can achieve. By using PhoreGen, we successfully identified inhibitors for evolved metallo-β-lactamases (MBLs) and serine-β-lactamases (SBLs), key drivers for antimicrobial resistance[33–35], and identified the covalent inhibitors for metallo-nicotinamidases, emerging targets for insecticides[36].

Results

Pharmacophore-oriented 3D molecular generative framework

PhoreGen is a pharmacophore-oriented conditional diffusion model designed to generate entire 3D molecules that are precisely aligned with the pharmacophore constraints (Fig. 1a). It iteratively denoises random noise guided by the interaction patterns and the steric constraints encoded by the pharmacophore features ($\mathcal{P}$), effectively transforming noise into valid 3D molecular structures. Specifically, PhoreGen follows a diffusion process ($q(M_t|M_{t-1})$) where Gaussian noise is progressively added to real molecular structures (M_0), ultimately leading to a state of 'pure noise' (M_T) (Fig. 1b). Conversely, the generative process utilizes a neural network (parameterized by θ, denoting the set of all learnable parameters) to learn the noise-removal steps conditioned on the pharmacophore information, enabling the recovery of the original molecule (M_0) through the distribution $p_\theta(M_{t-1}|M_t, \mathcal{P})$. By incorporating additional bond information into the diffusion–denoising workflow, PhoreGen enhances the chemical plausibility of the generated molecules.

PhoreGen acts through an E(3)-equivariant graph neural network[37] (Fig. 1b). It takes the perturbed molecule (M_t) at time step t and pharmacophore features ($\mathcal{P}'$) as input, which are encoded as a heterogeneous graph ($\mathcal{G}^t = \{\mathcal{G}_L^t, \mathcal{G}_K^t\}$) by a Transformer network. Here, $\mathcal{G}_L$ represents the ligand structure information as a fully connected graph, explicitly considering the chemical knowledge, such as bonds and intramolecule interactions. $\mathcal{G}_K$ is a k-nearest neighbor graph, standing for the correlations between the ligand atoms and pharmacophore features. Importantly, the pharmacophore context ($\mathcal{P}'$) remains fixed as the conditional information, ensuring that the generated molecules fit with the specified pharmacophore constraints. The encoded graph $\mathcal{G}^t$ is then fed into the cross-attention layer to recognize the underlying relationships of ligand–pharmacophore pairs. The update module utilizes pairwise direction-matching encodings to explicitly incorporate the unique knowledge of ligand–pharmacophore mapping (Fig. 1c), with the aim to enhance the mapping degree of generated molecules with the given pharmacophore constraints. Finally, the equivalent kernel predicts the noise to recover the perturbed molecules. In addition, a molecule size prediction module is employed to predict the number of atoms to be generated (Fig. 1d), by comprehensively considering the geometric and shape constraints rooted in the pharmacophore model.

For the generative process, we first use the trained molecule size prediction module to determine the atom number range based on the pharmacophore model, thus ensuring the ligand diversity and molecule size control. Next, pure noise data M_T is randomly sampled from the standard Gaussian distribution $N(0, I)$ and iteratively denoised toward the final molecule M_0 utilizing the trained conditional distribution $p_\theta(M_{t-1}|M_t, \mathcal{P})$. To further ensure the structural validity of the molecule, we introduced a guidance strategy during sampling to enhance spatial consistency between the generated molecules and the pharmacophore models while maintaining realistic geometric structure.

To enable PhoreGen to fully learn the intrinsic relations between 3D ligand structures and pharmacophore features, we implemented a two-phase training scheme with three custom-compiled datasets: LigPhore (containing 2,350,797 ligand–pharmacophore pairs derived from 3D ligands), CpxPhore (12,581 pairs from experimental protein–ligand complex structures) and DockPhore (76,203 pairs from docking-yielded complex structures) training sets (Supplementary Table 1);

https://doi.org/10.1038/s43588-025-00850-5

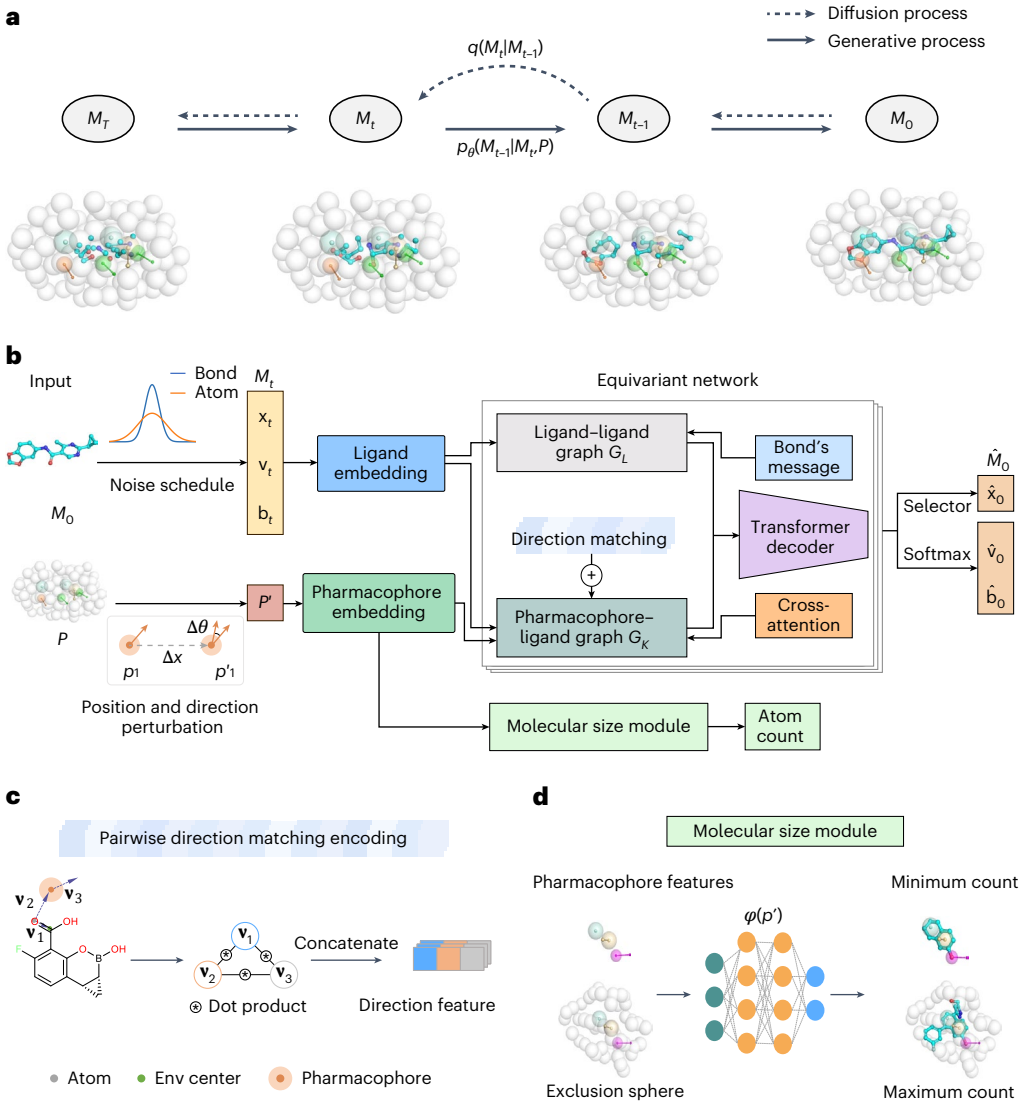


Fig. 1 | The overall architecture of PhoreGen. a, A diagram of diffusion and generative process in PhoreGen. The forward diffusion process starts with a real molecule M_0 and gradually adds noise to it until it becomes a purely random noise M_T. The generative process aims to reverse the forward diffusion by learning to remove noise from the noise distribution with a parameterized network. **b**, The workflow of PhoreGen. PhoreGen employs an E(3)-equivariant graph neural network, encoding the bond-first perturbed molecule M_t at time step t and pharmacophore P' with slightly perturbed position and direction into a heterogeneous graph $G = \{G_L, G_L\}$. The graph is processed through bond's message, cross-attention and a direction-matching module, which collectively ensure molecular plausibility and enhance the efficiency of ligand–pharmacophore mapping. The transformer decoder then takes the processed graph representation to output the denoised molecule M_0. **c**, The direction-matching encoding illustration. We define three key directional vectors: $\mathbf{v}_1$ captures local molecular geometry, $\mathbf{v}_2$ aligns with the pharmacophore feature and $\mathbf{v}_3$ is the pharmacophore feature orientation. These vectors undergo dot products to form a unified feature D_{ij}. 'Env center' refers to the coordinates of the center of all neighboring atoms within a 3 Å radius of a given atom, used to capture the local geometric environment of the ligand. **d**, The molecule size prediction module. The minimum atom count is predicted by a neural network using pharmacophore features alone, while maximum count considers both pharmacophore and exclusion-sphere features. During generation, the atom count is sampled from this predicted range.

for them, a total of 12 types of pharmacophore features, including four covalent features, were considered (see 'Dataset preparation' section and Supplementary Tables 2 and 3). The initial training phase uses the LigPhore dataset, which contains well-matched ligand–pharmacophore pairs, allowing the model to learn generalizable mapping patterns across a wide chemical and pharmacophoric space. The subsequent refinement phase employs CpxPhore and DockPhore, both of which contain imperfectly matched ligand–pharmacophore pairs. This enables the model to adapt to biased mapping patterns and recognize induced-fit effects in ligand–target interactions. In addition, we implement an asynchronous noise schedule that introduces more noise to the bonds than to the atoms, ensuring appropriate perturbation during the diffusion process.

As revealed by ablation analysis, removing the pretraining phase or the second-stage refinement training resulted in notable declines in the performance of PhoreGen, including reduced 3D conformational validity, decreased alignment with pharmacophore models and lower drug-likeness (Supplementary Table 4). Omitting bond information during the diffusion–denoising process substantially impaired the model's ability to generate molecules with chemically and energetically plausible conformations (Supplementary Table 4). Since pharmacophore models inherently incorporate a tolerance range for ligand–pharmacophore mapping, we introduced additional perturbations to the feature centers and orientations to enrich the input data and alleviate the extreme constraints imposed by these models, ultimately enhancing the model's transferability.

https://doi.org/10.1038/s43588-025-00850-5

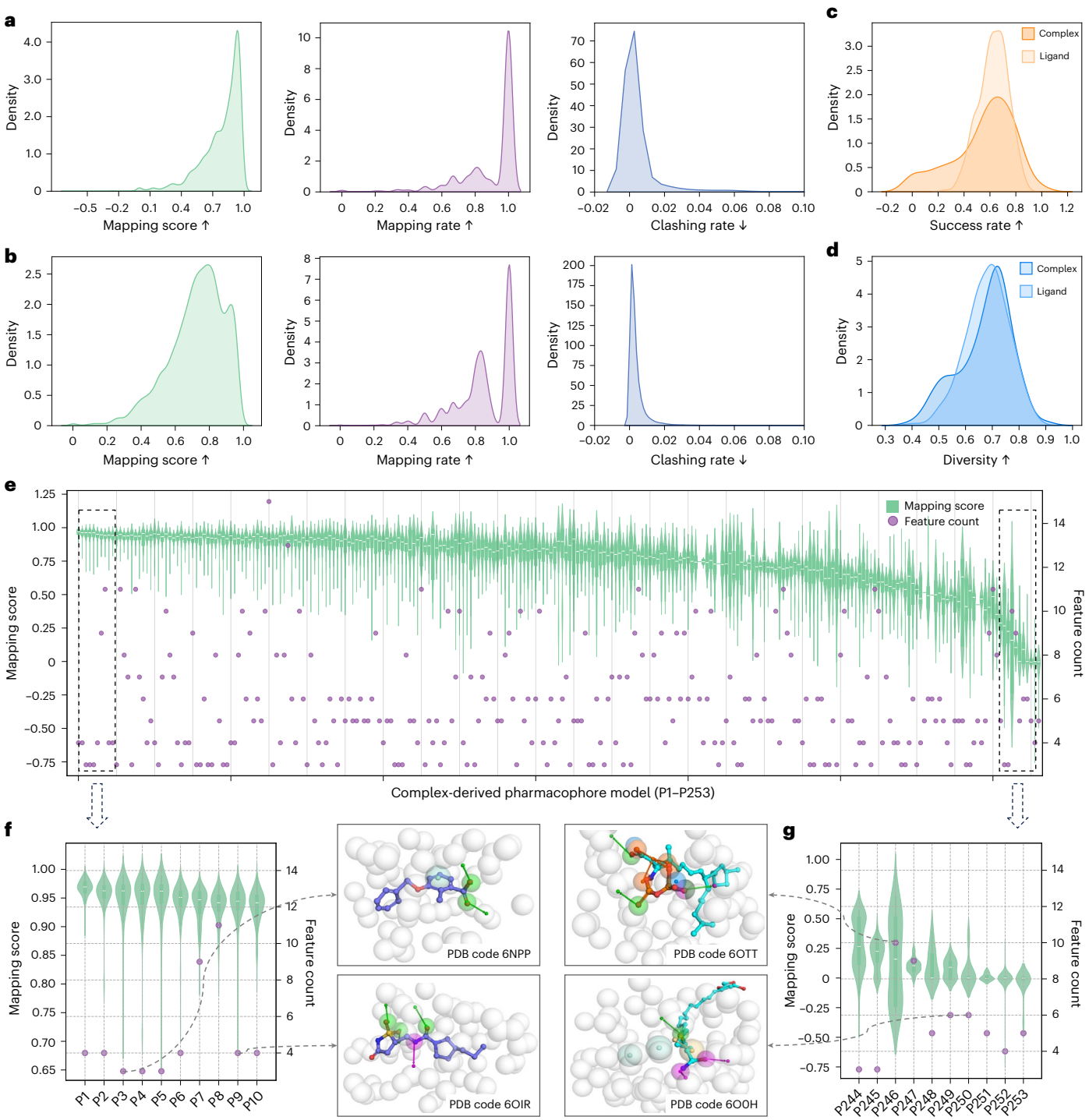


Fig. 2 | PhoreGen generates 3D molecules well aligned with complex- and ligand-derived pharmacophore models. a,**b**, The distributions of mapping scores, feature mapping rates and exclusion-sphere clashing rates of the molecules generated for complex-derived (**a**) and ligand-derived (**b**) pharmacophore models. **c**, PoseBusters success rate distribution for the molecules generated from both complex-derived and ligand-derived pharmacophore models, reflecting the ability of PhoreGen to produce realistic poses. **d**, Diversity distribution of molecules generated from both complex-derived and ligand-derived pharmacophore models, illustrating the chemical diversity of the generated molecules. The upward arrows indicate metrics where higher values are preferred, while downward arrows indicate metrics where lower values are preferred. **e**, The relationships between mapping scores of the generated molecules and feature counts of the complex-derived pharmacophore models. **f**,**g**, Detailed views for the ten complex-derived pharmacophore models, showcasing the generated molecules with the highest ($n = 100$) (**f**) and lowest ($n = 100, 46, 26, 56, 10, 11, 8, 4, 62$ and 13) (**g**) mapping scores. Examples include high-scoring or low-scoring molecules generated for selected pharmacophore models. Carbon atoms are displayed as slate blue or cyan, and nitrogen atoms, oxygen atoms and sulfur atoms are displayed as blue, red and goldenrod, respectively. The box represents the interquartile range, the line inside the box is the median, the whiskers extend to ±1.5× interquartile range, and dots outside the whiskers are outliers.

https://doi.org/10.1038/s43588-025-00850-5

Table 1 | Performance comparison of PhoreGen and the baseline models across various metrics[a]

Method	Validity (↑)	QED (↑)	Synthetic accessibility (↑)	Lipinski (↑)	log*P*	Diversity (↑)	Time (s, ↓)	Docking score (↓)	High affinity (↑)	IFP similarity (↑)	PoseBusters (↑)
Test set	–	0.54	0.75	4.10	0.89	–	–	−8.73	–	–	–
AR	0.84	0.47	0.71	4.94	1.34	0.65	1,277	−5.79	21.4%	0.19	0.45
Pocket2Mol	0.98	0.53	0.83	4.99	0.94	0.90	1,906	−4.64	3.6%	0.08	0.91
DeepICL	0.99	0.62	0.34	4.98	1.28	0.83	29	−7.02	19.6%	0.40	0.35
TargetDiff[b]	0.19	0.25	0.55	3.66	3.14	0.37	19,995	−7.03	27.8%	0.39	0.29
DiffSBDD	0.68	0.51	0.61	4.66	1.32	0.77	85	−7.17	22.5%	0.56	0.56
PMDM	0.92	0.56	0.57	4.73	2.77	0.66	3097	−7.96	42.4%	0.11	0.77
DrugFlow	0.72	0.56	0.70	4.63	1.78	0.68	277	−7.96	37.6%	0.63	0.81
ShEPhERD	0.22	0.58	0.68	4.81	1.32	0.64	3,876	−6.71	31.2%	0.64	0.88
PGMG	–	0.57	0.80	4.79	3.71	0.65	0.66	−6.59	22.0%	–	–
PhoreGen	0.91	0.61	0.71	4.68	1.52	0.65	919	−8.03	36.7%	0.66	0.74

[a]Validity, the proportion of 3D molecules passing the sanitization process in RDKit, with values ranging from 0 to 1. QED, with ranging from 0 to 1. Synthetic accessibility, evaluating the ease of chemical synthesis for a molecule, with values ranging from 0 to 1. Lipinski, the average count of Lipinski's rules satisfied, ranging from 0 to 5. log*P*, the lipid–water partition coefficient, indicating the molecule's lipophilicity. Diversity, the structural variability, measured by the average Tanimoto dissimilarity, with values ranging from 0 to 1. Time, the average computation time required to generate 100 valid molecules. Docking score, a metric from AutoDock Vina estimating the binding affinity, with lower scores indicating stronger binding. High affinity, the percentage of generated molecules achieving better binding scores than the reference ligand. IFP similarity, the cosine similarity of interaction fingerprints, ranging from 0 to 1. PoseBusters, the success rate of generated molecules passing chemical and physical conformation checks, with values ranging from 0 to 1. An upward (or downward) arrow next to each metric indicates that higher (or lower) values represent better performance [b]We observed substantial differences in the validity and diversity of TargetDiff compared with the reported values in other studies, probably due to the use of the CpxPhore test set in our analysis, whereas other papers reported results based on sampling proteins from the CrossDocked2020 test set. Each model generated 100 molecules per target and the values for each metric in the table are means.

Generating 3D molecules well mapped with pharmacophores

In this section, we evaluated the capability of PhoreGen to generate 3D molecules that match the specified pharmacophore models. We used 269 complex-derived and 300 ligand-derived pharmacophore models, none of which was included in the training sets. Performance was evaluated using three metrics: mapping score, feature mapping rate and exclusion-sphere clashing rate (see 'Evaluation metrics' section).

Given the complex-derived pharmacophore models, PhoreGen-generated 3D molecules that matched well, with an average mapping scoring of 0.784, an average feature mapping rate of 0.894 and a low clash rate of 0.006 with the exclusion spheres (Fig. 2a). For ligand-derived models, the generated molecules also exhibited good mapping, albeit with slightly lower mapping scores (average 0.725) and mapping rates (average 0.847) (Fig. 2b). Using PoseBusters[38], PhoreGen achieved successful conformational accuracy rates of 0.552 for complex-derived and 0.626 for ligand-derived models (Fig. 2c). The generated molecules also showed considerable chemical diversity, with diversity scores of 0.673 for complex-derived and 0.682 for ligand-derived models (Fig. 2d), highlighting the ability of PhoreGen to explore a wide chemical space while minimizing structural redundancy.

Next, we examined the relationships between the quality of the generated molecules (based on the mapping score and PoseBusters validity) and key aspects of pharmacophore models (including feature count, feature composition and number of exclusion spheres). Figure 2e shows the mapping scores and feature counts for the molecules generated using the complex-derived pharmacophore models. When pharmacophore models had a balanced feature count, feature composition and alignment with common drug–target interactions (Extended Data Fig. 1 and Supplementary Tables 5–12), PhoreGen generally produced well-mapped molecules with reasonable chemical structures and energy profiles (Fig. 2f and Supplementary Tables 5–12). In contrast, for models with spatially dense, homogenized and/or distant features, PhoreGen generated fewer valid 3D molecules with low diversity or suboptimal quality (Fig. 2g and Supplementary Fig. 1). These models usually correspond to molecules with specific chemical scaffolds within a limited chemical space. This could be improved by optimizing pharmacophore models or training on specifically customized datasets. A similar trend was observed for the ligand-derived models, where molecule quality was closely linked to the characteristics of the pharmacophore models (Supplementary Fig. 2 and Tables 13–21).

Unlike other diffusion models that determine the molecular size of generated molecules based on empirical distributions, PhoreGen employs a neural network to predict molecular size directly from pharmacophore features and exclusion spheres (Fig. 1d). We examined how the count of pharmacophore features, the number of exclusion spheres and cavity size (the average distance between pharmacophore features and exclusion spheres) relate to the molecular size and chemical diversity of generated molecules. For the complex-derived models, the atom count increases with the number of pharmacophore features and exclusion spheres (Extended Data Fig. 2a,b), a trend also observed for ligand-derived models (Extended Data Fig. 2c,d). The atom counts of generated molecules exhibit a strong positive correlation ($r > 0.85$) with cavity sizes of both complex-derived and ligand-derived models (Extended Data Fig. 3a,b), indicating that the neural network effectively controls molecular size in generation by capturing spatial information. In terms of molecular diversity, as the number of pharmacophore features and exclusion spheres increases, the diversity of the generated molecules decreases (Supplementary Fig. 3). The number of pharmacophore features has a relatively greater impact on diversity than exclusion spheres or cavity size (Supplementary Fig. 4), highlighting the importance of pharmacophore features in mapping chemical space. These results also suggest that a balance between pharmacophore features, exclusion spheres and molecular diversity is key to improving success rates in practical applications.

De novo drug design

We next assessed the performance of PhoreGen in de novo drug design by comparing it with nine baseline methods, including seven 3D structure-based generation models—AR[11], Pocket2Mol[12], DeepICL[15], TargetDiff[13], DiffSBDD[25], PMDM[14] and DrugFlow[39]—as well as ShePhERD[40] (a 3D molecular generation model that jointly considers 3D molecular shapes, electrostatic surfaces and pharmacophores) and PGMG[32] (a pharmacophore-based SMILES generation model). We selected ten protein targets that were excluded from the training set, with low sequence identity (average 37%) and low 3D pharmacophore similarity to the training set (Supplementary Fig. 5), but clear protein–ligand

interaction patterns (Supplementary Fig. 6). For each target, 100 molecules were generated using different models and key metrics such as sampling validity, quantitative estimate of drug-likeness (QED), synthetic accessibility, docking score, interaction fingerprint (IFP) similarity and conformation quality were evaluated (see 'Evaluation metrics' section).

PhoreGen achieved an average sampling validity of 91% for the ten targets, markedly surpassing the diffusion-based models, TargetDiff and ShEPhERD (Table 1). It performed similarly to baseline methods in QED, synthetic accessibility, Lipinski and log*P* metrics, indicating its capability to generate drug-like and synthetically accessible molecules. In terms of structural diversity, PhoreGen was comparable to AR, PMDM, DrugFlow and ShEPhERD, but slightly behind Pocket2Mol, DeepICL and DiffSBDD (Table 1), partly owing to its stricter pharmacophore feature constraints. Notably, PhoreGen effectively avoids generating molecules with undesirable traits, such as low molecular weights or sizes that deviate substantially from target-adapted compounds, overcoming common limitations in many SBMG methods (Supplementary Fig. 7). This advantage can be partly attributed to the molecular size prediction module (Fig. 1d). In addition, PhoreGen has acceptable computational efficiency, requiring an average of only 919 s to generate 100 valid molecules across the 10 targets (Table 1).

Subsequently, we conducted molecular docking analyses using AutoDock Vina 1.2.2 (ref. 41) for the molecules generated by the above-described models with each target, ensuring a fair comparison. PhoreGen achieved an average docking score of −8.03 kcal mol^{-1}, outperforming the nine baseline methods in generating molecules with potential high binding affinity (Table 1 and Supplementary Fig. 8). By analyzing the IFPs with each target, we observed that PhoreGen-generated molecules had a high average IFP similarity of 0.66, surpassing other baseline models, including AR (0.19), Pocket2Mol (0.08), DeepICL (0.40), TargetDiff (0.39), PMDM (0.11) and DiffSBDD (0.56) (Table 1 and Supplementary Fig. 9). The ShEPhERD model, which incorporates directional pharmacophore features (defined differently from PhoreGen), along with molecular shapes and electrostatic potential surfaces, achieved a similarly high IFP similarity of 0.63. These results, along with an analysis of the top-ranked molecules (Supplementary Table 22), indicate that PhoreGen-generated molecules align closely with common interaction patterns, suggesting a higher likelihood of being active molecules. In addition, analysis of IFP similarity versus molecular diversity (Supplementary Fig. 10) showed that high IFP similarity does not compromise molecular diversity.

We then assessed the molecular conformation of the generated molecules using PoseBusters[38]. PhoreGen achieved an overall success rate of 0.74, substantially outperforming the 3D generative models such as AR (0.45), DeepICL (0.35), TargetDiff (0.29) and DiffSBDD (0.56) (Table 1). Although Pocket2Mol achieved a high success rate of 0.91, it generates many small-sized molecules (Supplementary Fig. 7) that may have limited biological activity, as evidenced by their low binding affinity and IFP similarity (Table 1 and Supplementary Figs. 8 and 9). PMDM, DrugFlow and ShEPhERD also exhibited good success rates of 0.77, 0.81 and 0.88, respectively. Meanwhile, we analyzed the distributions of ring substructures in the generated molecules, finding that PhoreGen produced fewer low-frequency ring structures (for example, three-membered rings at 3.3%, four-membered rings at 1.4%, eight-membered rings at 0.3%, nine-membered rings at 0.1% and polycyclic rings at 4.9%) (Supplementary Table 23). The results highlight the ability of PhoreGen to generate appropriately sized 3D molecules with chemically and energetically reasonable ring structures.

Furthermore, we expanded the evaluation to the CrossDocked2020 test set (100 targets)[42] and found that PhoreGen consistently performed well, especially in binding affinity and IFP similarity, remaining competitive with baseline models (Supplementary Table 24). However, its performance was suboptimal for a few targets with uncommon or complex pharmacophore models, requiring more structurally specific molecules. In addition, we noticed differences in the chemical and pharmacophoric spaces, as well as in the distribution of ligand–pharmacophore pairs, between the CrossDocked2020 test set and our training set (Supplementary Fig. 11). Performance on such targets could be enhanced by incorporating more diverse ligand–pharmacophore pairs in the training set.

Feature-customized drug design

In target-centered drug discovery, designing molecules with specific pharmacophore features (such as anchoring features, covalent groups, metal-binding motifs and so on) is often essential. Here, we evaluate the capability of PhoreGenin feature-customized drug discovery, with a focus on covalent and metalloenzyme drug design.

For covalent drug discovery, we defined four covalent features based on chemotypes and residues involved in covalent reactions (mainly from the CovalentInDB 2.0 database[43]; Supplementary Table 3). We trained PhoreGen with 674,507 ligand-derived ligand–pharmacophore pairs representing prereaction states and 635 complex-derived pairs for postreaction states. We evaluated the performance of PhoreGen on 10 targets (excluded from the training set) with covalent reactive sites, generating 100 molecules per target. The results showed 76% of the generated molecules (on average across the 10 tested targets) contained suitable covalent motifs, with an average mapping score of 0.64 (Supplementary Tables 25 and 26). Notably, most motifs represented the postreaction state (Fig. 3a), reflecting the impact of refinement training with complex-derived ligand–pharmacophore pairs that correspond to postreaction states.

For metalloenzyme drug design, we used 1,659,215 ligand-based and 1,686 complex-based ligand–pharmacophore pairs, each with metal-coordination features, to train PhoreGen. On the ten selected metalloenzyme targets (excluded from the training set), over 95% of generated molecules contained reasonable metal-binding motifs, with high pharmacophore mapping scores (Supplementary Tables 27 and 28). PhoreGen also can generate chemically reasonable and diverse metal-binding motifs, even for metalloenzymes (for instance, PA and MAP; Fig. 3b) with multicoordination features or multiple metal ions. In contrast, the baseline SBMG models struggle to generate molecules with covalent or metal-binding features for these targets. These results highlight the potential of PhoreGen in covalent and metalloenzyme drug design, showcasing the unique advantage of pharmacophore-oriented 3D molecular generation for feature-customized drug discovery.

Design of dual MBL/SBL inhibitors

Resistance to β-lactam antibiotics is an increasingly growing clinical problem, substantially driven by the spread of evolved MBLs and SBLs that efficiently hydrolyze β-lactams[33–35]. Currently, only a few dual MBL/SBL inhibitors are in clinical trials, with QPX7728 showing potent, broad-spectrum inhibition against various MBL/SBL isoforms at nanomolar level[44]. We used PhoreGen to generate novel molecules based on pharmacophore models derived from crystal structures of QPX7728 in complex with MBL/SBL enzymes (Supplementary Fig. 12). Of the 300 generated molecules (Supplementary Figs. 13–15), we selected two new bicyclic boronates, **CB1** and **CB2**, for synthesis (Fig. 4a). Both compounds displayed broad-spectrum inhibition against clinically relevant SBL and MBL enzymes (Fig. 4b). **CB1** manifested nanomolar inhibitory activity against SBL KPC-2 (half-maximum inhibitory concentration (IC_{50} of 0.031 μM) and OXA-48 (IC_{50} of 0.16 μM), and **CB2** showed IC_{50} values of 0.11 μM and 0.013 μM against KPC-2 and OXA-48, respectively. Both compounds were potent inhibitors of NDM-1 and VIM-2, with moderate activity against IMP-1 (Fig. 4b). The OXA-48:**CB1** complex structure (Protein Data Bank (PDB) code 9KSA; Supplementary Table 29) revealed that **CB1** forms a covalent

https://doi.org/10.1038/s43588-025-00850-5

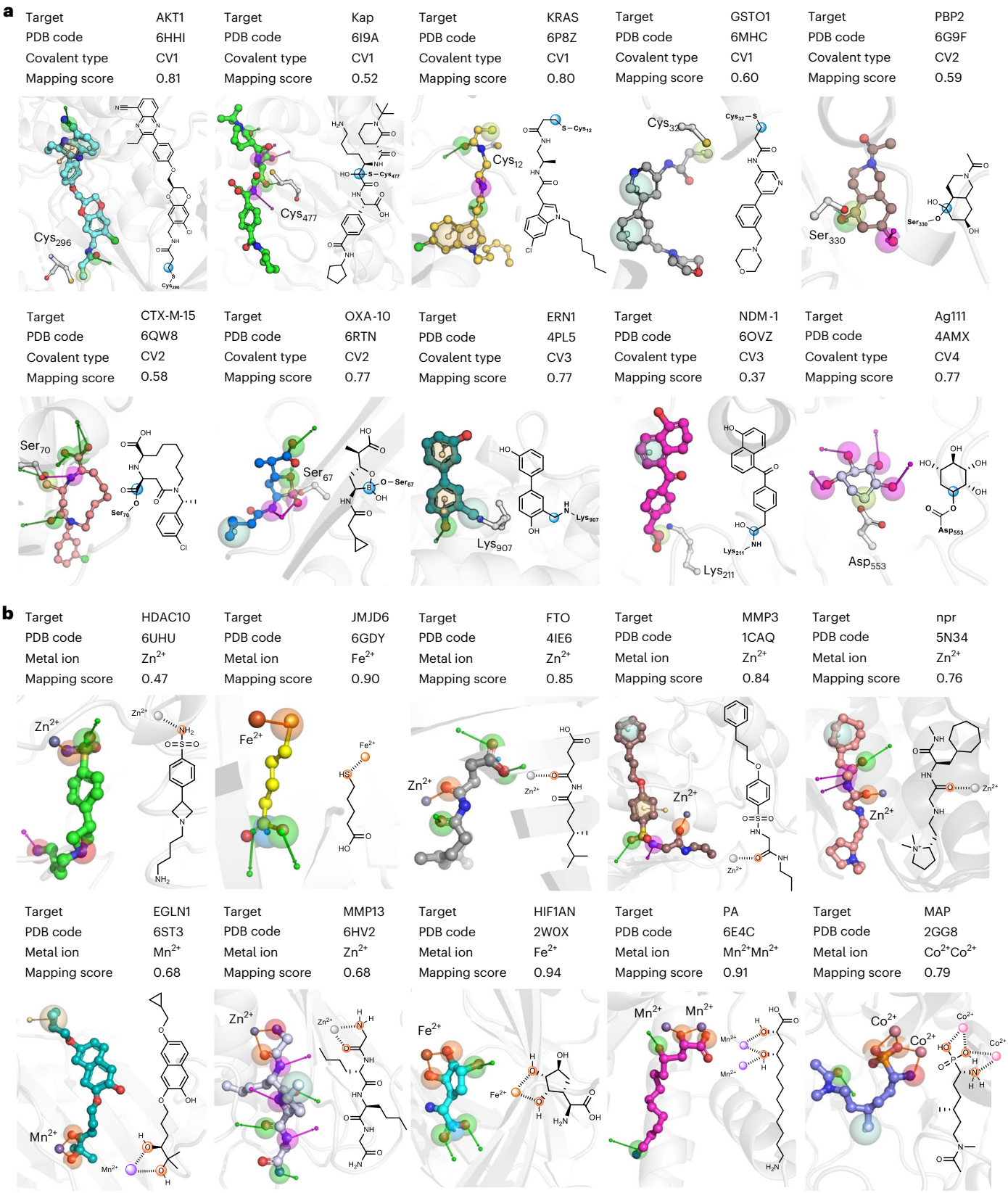
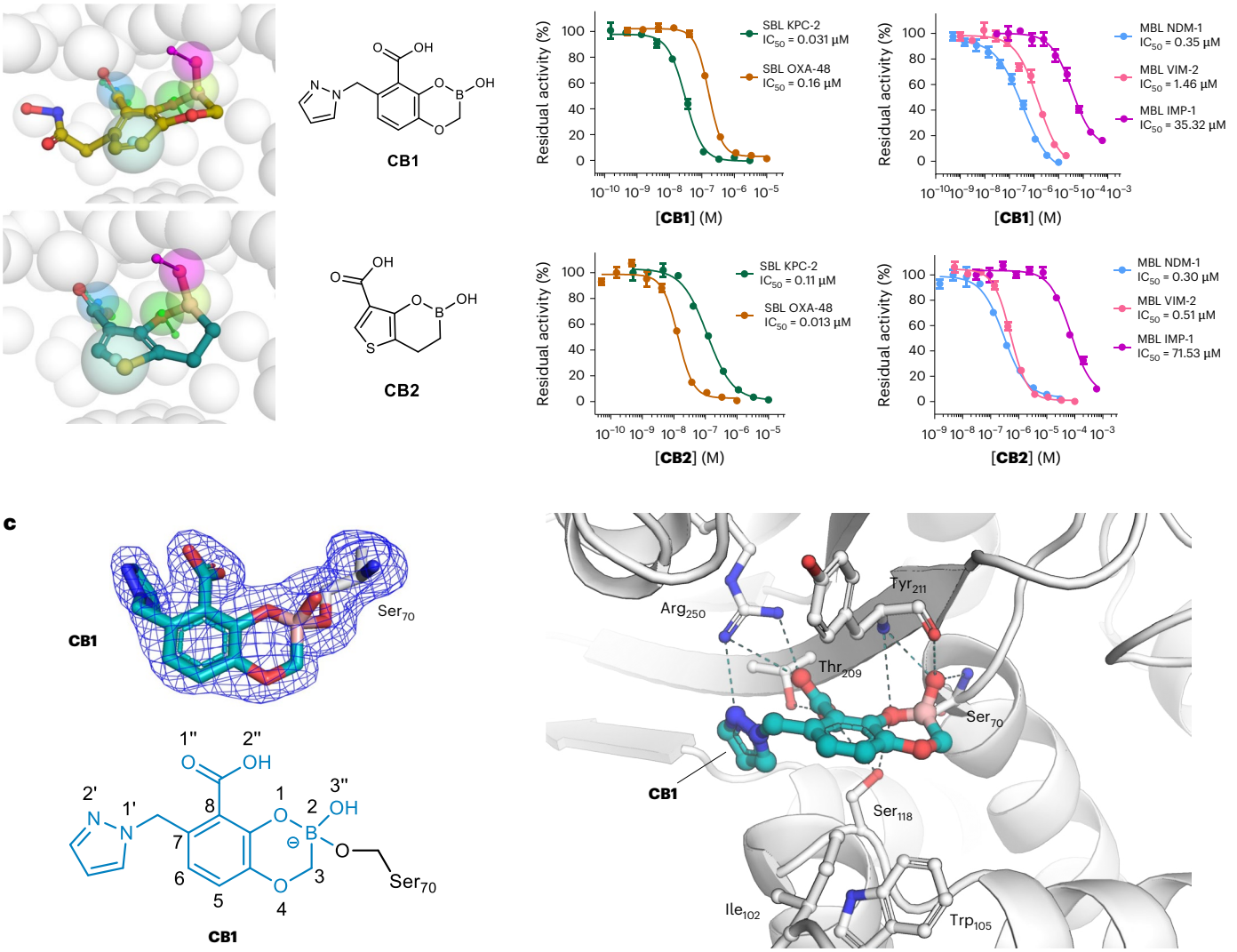


Fig. 3 | PhoreGen shows great potential in covalent and metalloenzyme drug design. a, Views of representative 3D molecules generated for the pharmacophore models across ten targets with a covalent reactive site, indicating the ability of PhoreGen in generating molecules with suitable covalent motifs. The covalent atoms of the generated molecules are highlighted as blue spheres. **b**, Views of representative 3D molecules generated for the pharmacophore models across ten metalloenzyme targets, showing that PhoreGen can generate molecules with metal-binding groups, highlighted as orange spheres. Boron atoms, nitrogen atoms, oxygen atoms, phosphorus atoms, sulfur atoms and chlorine atoms are pink, blue, red, dark orange, goldenrod and lime green, respectively, while the other colors are carbon atoms.

https://doi.org/10.1038/s43588-025-00850-5



d

Compound	Meropenem MIC (μg ml^{-1}) ± 10 μM compound						
	Kp140912 (*bla*$_{KPC-2}$)	Kp090344 (*bla*$_{KPC-2}$)	Kp140508 (*bla*$_{KPC-2}$)	Kp090571 (*bla*$_{NDM-1}$)	Kp140509 (*bla*$_{NDM-1}$)	Kp090567 (*bla*$_{NDM-1}$)	AbS88090704 (*bla*$_{OXA-23}$)
None	128	>128	8	64	64	32	64
CB1	32	128	0.5	4	1	1	64
CB2	0.25	4	0.25	2	2	4	16

Fig. 4 | Application of PhoreGen in discovering novel dual MBL/SBL inhibitors. **a**, The generated 3D conformers well mapped with the pharmacophore model. Considering the feasibility of chemical synthesis, we selected **CB1** for synthesis, which shares the same core boronate scaffold and an isosteric side chain with the generated molecule no. 157 (Supplementary Figs. 13–15). **b**, IC_{50} curves of CB1/CB2 against clinically relevant SBL (KPC-2 and OXA-48) and MBL enzymes (NDM-1, VIM-2 and IMP-1). Data are presented as mean values ± s.e.m. from $n = 3$ biological replicates, with error bars representing the s.e.m. **c**, The view of a crystal structure of OXA-48 in complex with **CB1** (PDB code 9KSA) reveals that the inhibitor binds to make a covalent bond with Ser_{70}, and form ionic interactions with Arg_{250} and hydrogen bonds with Ser_{70} ($O3''_{CB1}$–$N1_{Ser70}$ distance of 2.9 Å), Ser_{118} ($O1_{CB1}$–$O2_{Ser118}$ distance of 3.0 Å, $O2''_{CB1}$–$O2_{Ser118}$ distance of 3.3 Å), Thr_{209} ($O2''_{CB1}$–$O2_{Thr209}$ distance of 2.7 Å), Try_{211} ($O3''_{CB1}$–$O1_{Try211}$ distance of 2.7 Å, $O3''_{CB1}$–$N1_{Try211}$ distance of 2.9 Å) and Arg_{250} ($O1''_{CB1}$–$N3_{Arg250}$ distance of 2.8 Å, $O1''_{CB1}$–$N4_{Arg250}$ distance of 3.5 Å, $N2'_{CB1}$–$N4_{Arg250}$ distance of 3.2 Å). The clear mF_o–DF_c electron density (OMIT maps) around **CB1** (blue mesh, contoured to 3σ) calculated from the final refined model, indicates that **CB1** is confidently modeled. The r.m.s.d. value of the predicted binding mode of **CB1** core scaffold with that from crystal structures is 0.39 Å. **d**, The MICs of meropenem with and without **CB1** and **CB2** against SBL-producing and MBL-producing bacteria, including *Klebsiella pneumoniae* (Kp) and *Acinetobacter baumannii* (Ab), demonstrate their potential in reversing carbapenem resistance. Boron atoms, nitrogen atoms, oxygen atoms and sulfur atoms are pink, blue, red and goldenrod, respectively, while the other colors are carbon atoms.

https://doi.org/10.1038/s43588-025-00850-5

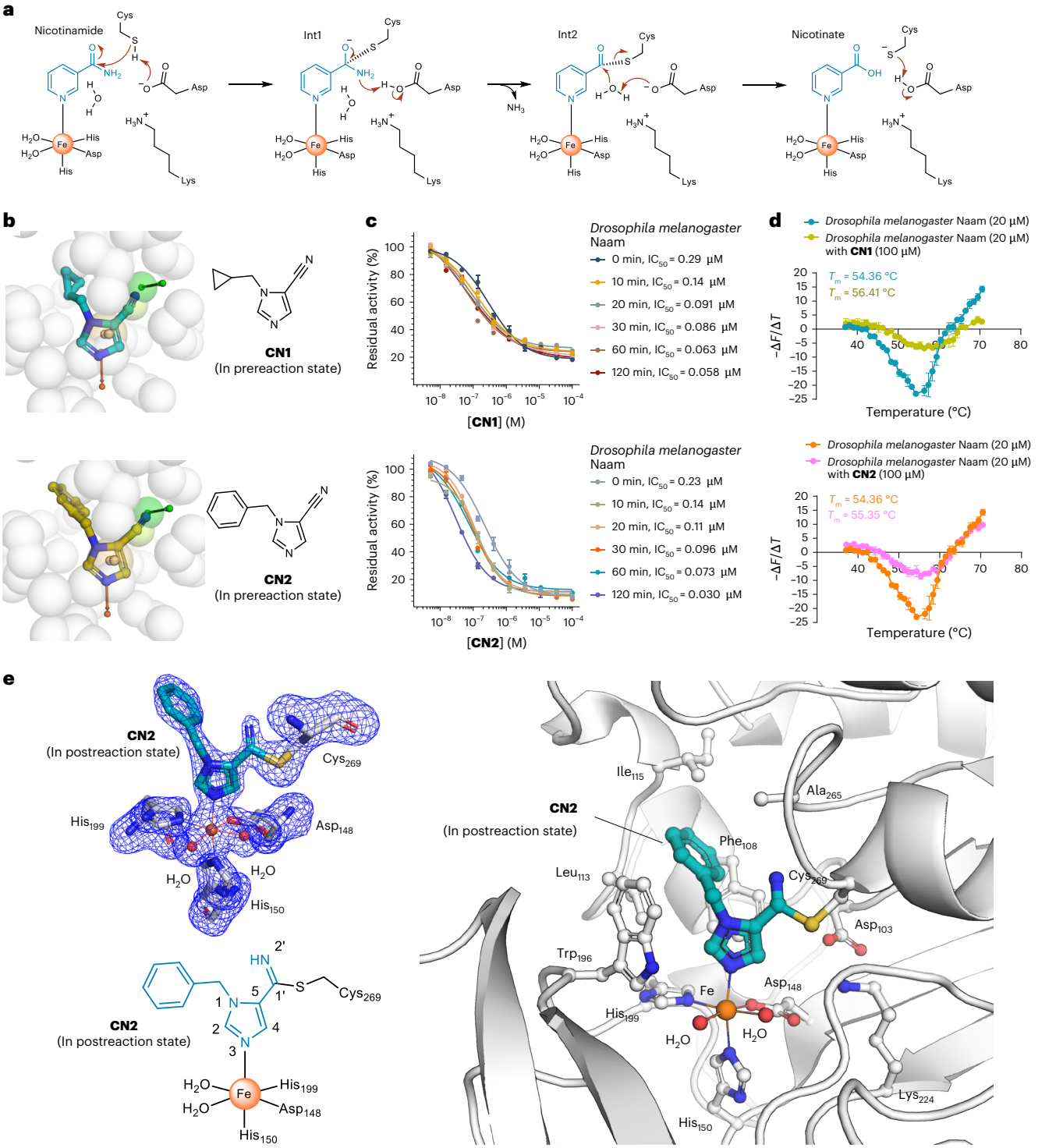


Fig. 5 | Application of PhoreGen leads to identification of the covalent inhibitors of Naam. a, The proposed chemical mechanism of Naam-catalyzed transformation of nicotinamide into nicotinate, which is the basis for establishing the pharmacophore models for molecular generation. Int1 and Int2 refer to two intermediates in the catalytic pathway. **b**, Of the 300 molecules generated by PhoreGen (Supplementary Figs. 17–19), compounds **CN1** and **CN2** were selected for synthesis, with the cyano group serving as the motif to covalently react with the catalytic cysteine. **c**, The IC_{50} curves of compounds **CN1** and **CN2** inhibiting *Drosophila melanogaster* Naam (15 nM) show that both compounds exhibit time-dependent inhibition to the enzyme, probably due to the reaction of the cyano group with the catalytically active cysteine. Data are presented as mean values ± s.e.m. from $n = 3$ biological replicates, with error bars representing the s.e.m. **d**, The melting curves (first derivative of dissociation) of *Drosophila melanogaster* Naam (20 μM) in the presence or absence of **CN1** (100 μM) or **CN2** (100 μM), revealing that both compounds bind to thermodynamically stabilize the enzyme. Data are presented as mean values ± s.e.m. from $n = 3$ biological replicates, with error bars representing the s.e.m. **e**, The view of a crystal structure of *Drosophila melanogaster* Naam in complex with **CN2** (PDB code 9U8M) reveals that **CN2** makes a coordination bond with Fe^{2+} (N3–Fe^{2+} distance of 2.1 Å) and forms a covalent bond with Cys_{269} (C1′–S_{Cys269} distance of 1.8 Å). **CN2** is observed in the postreaction state after being attacked by the nucleophilic Cys_{269}. The clear mF_o-DF_c electron density (OMIT maps) around **CN2** and Fe^{2+}-coordinated residues (blue mesh, contoured to 3σ) calculated from the final refined model, indicates that **CN2** and Fe^{2+}-coordinated residues are confidently modeled. The r.m.s.d. value of the predicted binding mode of **CN2** compared with the crystal structures is 0.21 Å. Nitrogen atoms, oxygen atoms and sulfur atoms are blue, red and goldenrod, respectively, while the other colors are carbon atoms.

bond with the catalytic Ser_{70} and makes ionic interactions with Arg_{250}, as well as multiple hydrogen bonds with Ser_{70}, Ser_{118}, Thr_{209}, Tyr_{211} and Arg_{250} (Fig. 4c). This is consistent with the mode observed for QPX7728 (Supplementary Fig. 16) and aligns well with the generated binding mode. Cellular assays revealed that both compounds enhanced the efficacy of meropenem against resistant Gram-negative bacteria, reducing minimum inhibitory concentrations (MICs) by up to 512-fold (Fig. 4d). These chemotypes, particularly **CB2**, represent promising leads for combating carbapenem resistance.

Design of covalent Naam inhibitors

Nicotinamidase (Naam), an Fe^{2+}-dependent enzyme, converts nicotinamide to nicotinic acid and has been identified as a target for neurotoxic insecticides[36]. Currently, there are almost no reported small-molecule inhibitors targeting Naam in pests. Based on Naam's catalytic mechanism[45] (Fig. 5a), we created a pharmacophore model that contain a covalent bond feature, a metal-coordination feature, an aromatic ring and a hydrogen-bond acceptor (Fig. 5b). From 300 molecules generated by PhoreGen (Supplementary Figs. 17–19), we identified 1*H*-imidazole-5-carbonitrile compounds, **CN1** and **CN2** (Fig. 5b), which displayed time-dependent inhibition of *Drosophila melanogaster* Naam, with IC_{50} values of 0.058 and 0.030 μM, respectively, after 120 minutes of incubation (Fig. 5c). Both compounds exhibited the ability to thermodynamically stabilize *Drosophila melanogaster* Naam (ΔT_m of 2.05 °C for **CN1** and 0.99 °C for **CN2;** Fig. 5d). Liquid chromatography–tandem mass spectrometry (LC–MS/MS) analysis confirmed that **CN1** and **CN2** form a covalent bond with the catalytic residue Cys_{269} of *Drosophila melanogaster* Naam (Extended Data Fig. 4). Crystallography of the **CN2**–*Drosophila melanogaster* Naam complex (PDB code 9U8M; Supplementary Table 29) revealed that **CN2** forms a coordination bond with Fe^{2+} via a typical octahedron geometry and a covalent bond with Cys_{269} (Fig. 5e), consistent with the LC–MS/MS results. Superimposing the predicted and co-crystal binding modes showed a high consistency (root mean squared deviation (r.m.s.d. of 0.21 Å). **CN1** and **CN2** also exhibited time-dependent activity against Naam enzymes in the pests *Myzus persicae* and *Bemisia tabaci*, while they thermodynamically stabilize both enzymes (Extended Data Fig. 5). These covalent inhibitors show potential for developing new insecticides targeting Naam.

Discussion

Designing chemical entities based on a specific mode of action is a key strategy in modern drug discovery, with many FDA-approved small-molecule drugs being derivatives or mimics of natural products, endogenous ligands or known drugs. However, there is a lack of intelligent methods for feature-customized drug discovery. This study explores the potential of the pure PBMG approach and develops PhoreGen, a practical tool toward intelligent drug discovery. The results show that PhoreGen merges the strengths of both LBMG and SBMG methods, making it applicable to pharmacophore models derived from ligands, proteins, complexes and catalytic mechanisms. PhoreGen generates molecules with a higher probability of bioactivity, as shown by re-docking analyses and case studies. The ShEPhERD[40] model also produces promising results by incorporating directional pharmacophore features and learning additional molecular shapes and electrostatics. While PharmacoBridge[46] uses pharmacophores to generate 3D molecules for protein targets, its applicability is limited by the lack of covalent, metal-binding and exclusion-sphere features. In contrast, PhoreGen can efficiently generate feature-customized molecules, such as covalent motifs, metal-binding groups and distinctive molecules (for example, boronates), showcasing its unique potential. The flexibility of the pharmacophore model, especially with ensemble pharmacophore models, improves molecule quality, diversity and success rate. Compared with fixed substructures/scaffold constraints, PBMG offers greater flexibility, allowing for more diverse chemical scaffolds and customization. These advantages make PhoreGen a practical and superior tool for molecular generation over SBMG and LBMG methods.

PhoreGen leverages specific embeddings and perturbations of pharmacophore features to handle mapping deviations between ligands and pharmacophores. By embedding pharmacophore–ligand mapping principles into the message-passing process, the model effectively captures key relationships between pharmacophores and 3D molecular conformations, enhancing sampling efficiency. It also integrates chemical knowledge, including bond information, into ligand generation, improving the quality of simultaneous atom and bond generation. In addition, PhoreGen utilizes spatial information from pharmacophore features and exclusion spheres to control molecular size, reducing the likelihood of producing misaligned-size compounds. Its efficiency is further boosted by high-quality datasets of ligand–pharmacophore pairs and the hierarchical training strategy. The potential of the model can be expanded by updating the dataset to explore metallo-drugs, radionuclide drugs and feature-rich compounds such as sugar analogs. Furthermore, receptor-based pharmacophore modeling[47,48] could enable 'reference-free' molecule generation, directly deriving pharmacophore features from target structures. As a typical knowledge-driven and data-driven model, PhoreGen lays a strong foundation for artificial intelligence-enabled drug discovery.

PhoreGen has enabled the discovery of new bicyclic boronates (especially **CB2**) as dual MBL–SBL inhibitors, establishing new starting points for tackling carbapenem resistance. It also identified covalent inhibitors for Naam, supporting the development of novel pesticides. These cases demonstrate that PhoreGen is effective for both the 'old' targets with well-defined pharmacophore features and 'novel' targets guided by catalytic or unique features. PhoreGen still has some limitations, including an imperfect balance between feature compromise matching and precise chemical conformation, challenges in generating reasonable compounds for uncommon pharmacophore models and a lack of consideration for synthetic feasibility. Overall, this work provides a new valuable mode and successful cases for advancing intelligent drug discovery.

Methods

Definitions and notations

This work provides the first PBMG method for generating entire 3D molecules that align with a given pharmacophore model (Fig. 1a). A pharmacophore model can be represented as $\mathcal{P} = \left\{\left(x_P^{(i)}, v_P^{(i)}\right)\right\}_{i=1}^{N_P}$, where N_P denotes the number of pharmacophore features, $x_P^{(i)} \in \mathbb{R}^3$ represents the 3D coordinate of the pharmacophore point, and $v_P^{(i)} \in \mathbb{R}^k$ stands for the pharmacophore features (for example, pharmacophore type, tolerance range, normal direction and the feature number k). The molecule structure can be more comprehensively described as $\mathcal{M} = \left\{\left(x_M^{(i)}, v_M^{(i)}, b_M^{(i)}\right)\right\}_{i=1}^{N_M}$, where N_M indicates the number of atoms, $x_M^{(i)} \in \mathbb{R}^3$ and $v_M^{(i)} \in \mathbb{R}^l$ denote the atom position and atom type, respectively, and $b_M^{(i)} \in \mathbb{R}^{N_M}$ defines the chemical bond. Consequently, for the ligand molecule $M = [\mathrm{x}, \mathrm{v}, \mathrm{b}]$ in a matrix format, where $\mathrm{x} \in \mathbb{R}^{N_M \times 3}$ represents the atom positions, $\mathrm{v} \in \mathbb{R}^{N_M \times l}$ denotes the atom types, $\mathrm{b} \in \mathbb{R}^{N_M \times N_M}$ represents the chemical bonds and $[,,]$ refers to concatenation operation, the PBMG task in our study can be formalized as the conditional distribution $p(M|\mathcal{P})$.

Details of PhoreGen

Diffusion process. In a typical diffusion model[49], the diffusion process iteratively adds noise to the data according to a predefined noise schedule, while the generative process gradually removes the noise using a neural network until the original data are reconstructed (Fig. 1a). We model the continuous atom coordinates using a Gaussian distribution, while the discrete atom types and bond types are modeled with

two separate categorical distributions. The noisy molecule (M_t) at each time step t ($t = 0, 1, \ldots T$) is then sampled from the distribution $q(M_t|M_{t-1})$, where M_0 represents the real molecule

$$q(M_t|M_{t-1}) = \mathcal{N}\left(\mathrm{x}_t; \sqrt{1-\beta_t}\mathrm{x}_{t-1}, \beta_t I\right) \cdot \mathcal{C}\left(\mathrm{v}_t|(1-\beta_t)\mathrm{v}_{t-1} + \beta_t \mathbb{1}_k\right) \cdot \mathcal{C}\left(\mathrm{b}_t|(1-\beta_t)\mathrm{b}_{t-1} + \beta_t \mathbb{1}_{k'}\right). \quad (1)$$

Here, β_t is the predefined noise schedule, I is the identity matrix and $\mathbb{1}_k$ represents the one-hot vector where the kth element is one and all other elements are zeros. Notably, the noise scheduler β_t is set differently for atom coordinates, atom types and bond types.

By setting $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, the diffusion process is desired to sample the data at any arbitrary time step t through a closed-form formulation via a reparameterization trick. Specifically, this allows us to express the noisy molecular distribution $q(M_t|M_0)$ as

$$q(\mathrm{x}_t|\mathrm{x}_0) = \mathcal{N}\left(\mathrm{x}_t; \sqrt{\bar{\alpha}_t}\mathrm{x}_0, (1-\bar{\alpha}_t) I\right) \quad (2)$$

$$q(\mathrm{v}_t|\mathrm{v}_0) = \mathcal{C}\left(\mathrm{v}_t|\bar{\alpha}_t \mathrm{v}_0 + (1-\bar{\alpha}_t)\mathbb{1}_k\right) \quad (3)$$

$$q(\mathrm{b}_t|\mathrm{b}_0) = \mathcal{C}\left(\mathrm{b}_t|\bar{\alpha}_t \mathrm{b}_0 + (1-\bar{\alpha}_t)\mathbb{1}_{k'}\right). \quad (4)$$

The parameter $\bar{\alpha}_t$ represents the information retained from the original data at step t, determined by the noise level β_t. As t approaches infinity, the atom coordinates $q(\mathrm{x}_t|\mathrm{x}_0)$ tend toward to a standard Gaussian distribution, while the atom types $q(\mathrm{v}_t|\mathrm{v}_0)$ and bond types $q(\mathrm{b}_t|\mathrm{b}_0)$ converge to a specific noise category. Specifically, atom types are gradually perturbed into an extra category introduced to represent undefined atom types. Similarly, bond types are perturbed toward the 'none' type, where bonds are treated as nonexistent. These distributions serve as priors for the reverse generative process.

Generative process. In the generative process, we reverse the noise introduced during the diffusion process to reconstruct the ground-truth molecule M_0. By leveraging Bayes theorem, we can compute the normal posterior distribution $q(M_{t-1}|M_t, M_0, \mathcal{P})$ from equation (1) and equations (2–4) as follows:

$$q(\mathrm{x}_{t-1}|\mathrm{x}_t, \mathrm{x}_0) = \mathcal{N}\left(\mathrm{x}_{t-1}; \tilde{\mu}_t(\mathrm{x}_t, \mathrm{x}_0), \tilde{\beta}_t I\right) \quad (5)$$

$$q(\mathrm{v}_{t-1}|\mathrm{v}_t, \mathrm{v}_0) = \mathcal{C}\left(\mathrm{v}_{t-1}|\tilde{c}_t(\mathrm{v}_t, \mathrm{v}_0)\right) \quad (6)$$

$$q(\mathrm{b}_{t-1}|\mathrm{b}_t, \mathrm{b}_0) = \mathcal{C}\left(\mathrm{b}_{t-1}|\tilde{c}'_t(\mathrm{b}_t, \mathrm{b}_0)\right) \quad (7)$$

$$\tilde{\mu}_t(\mathrm{x}_t, \mathrm{x}_0) = \frac{\sqrt{\bar{\alpha}_t - 1}\beta_t}{1-\bar{\alpha}_t}\mathrm{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_t - 1)}{1-\bar{\alpha}_t}\mathrm{x}_t \quad (8)$$

$$\tilde{\beta}_t = \frac{1-\bar{\alpha}_t - 1}{1-\bar{\alpha}_t}\beta_t \quad (9)$$

$$\tilde{c}_t(\mathrm{v}_t, \mathrm{v}_0) \propto [\alpha_t \mathrm{v}_t + (1-\alpha_t)\mathbb{1}_k] \times [\bar{\alpha}_{t-1}\mathrm{v}_0 + (1-\bar{\alpha}_{t-1})\mathbb{1}_k] \quad (10)$$

$$\tilde{c}'_t(\mathrm{b}_t, \mathrm{b}_0) \propto [\alpha_t \mathrm{b}_t + (1-\alpha_t)\mathbb{1}_{k'}] \times [\bar{\alpha}_{t-1}\mathrm{b}_0 + (1-\bar{\alpha}_{t-1})\mathbb{1}_{k'}]. \quad (11)$$

The generative transition distribution with learnable parameters θ is then defined as

$$p_\theta(M_{t-1}|M_t, \mathcal{P}) = \mathcal{N}\left(\mathrm{x}_{t-1}; \mu_\theta(M_t, t, \mathcal{P}), \sigma_t^2 I\right) \cdot \mathcal{C}\left(\mathrm{v}_{t-1}|c_\theta(M_t, t, \mathcal{P})\right) \cdot \mathcal{C}\left(\mathrm{b}_{t-1}|c'_\theta(M_t, t, \mathcal{P})\right) \quad (12)$$

In this formulation, the μ_θ represents the mean of atom coordinates predicted by neural networks and σ_t^2 is the predefined variance. Similarly, the categorical distribution c_θ and c'_θ are also predicted by neural networks, determining the probabilities of atom and bond types. The overall training objective of the model is to optimize the variational lower bound (VLB), minimizing the Kullback–Leibler (KL) divergence between the posterior distribution $q(M_{t-1}|M_t, M_0, \mathcal{P})$ and the predicted prior $p_\theta(M_{t-1}|M_t, \mathcal{P})$

$$\mathcal{L}_{\mathrm{VLB}} = -D_{\mathrm{KL}}\left(q(M_{t-1}|M_t, M_0, \mathcal{P}) || p_\theta(M_{t-1})M_t, \mathcal{P})\right). \quad (13)$$

Bond-first noise schedule. As mentioned above, the ground-truth molecule is perturbed by disturbing both the atom and bond information. However, it is notable that bond types exhibit strong dependencies on both atom distances and atom types. If bonds are subjected to the same noise schedule as atoms, the system may introduce inconsistencies. Specifically, as more noise is added, bonded atoms may shift to positions that violate realistic chemical bond lengths, which might mislead the neural network toward problematic local structure constraints. Herein, we adopt a bond-first noise schedule[50], perturbing the bond types more aggressively than atom coordinates and atom types. This allows bonds to be disconnected appropriately when atoms are too far apart, pushing bond types toward the none-type early in the diffusion process. Meanwhile, the noise applied to atoms is introduced more smoothly, preserving their spatial relationships for a longer duration. As bonds undergo increased diffusion, atoms are eventually perturbed toward their Gaussian priors. In this way, the chemical correlations between the atoms and bonds in the noisy molecule might be more reasonable.

SE(3)-equivariant molecular denoising module. We introduce a transformer-based SE(3)-equivariant graph neural network that simultaneously considers both node-level and edge-level hidden states to better capture the dynamics of molecular diffusion during the denoising process. Given the input perturbed molecule M_t and pharmacophore model P' at time step t, the ligand embedding and pharmacophore embedding layers are utilized to feature the ligand graph $\mathcal{G}_L$ and pharmacophore graph $\mathcal{G}_P$, respectively. With the two graph embeddings, we then build a k-nearest neighbor graph $\mathcal{G}_K$ to describe the interactions between pharmacophore points and ligand atoms. The update process for the $\mathcal{G}_K$ is defined as

$$\Delta h_{K,i} = \sum_{j \in \mathcal{N}_K(i)} \phi_K(h_i, h_j, ||\mathrm{x}_i - \mathrm{x}_j||, E_{ij}, D_{ij}, t), \quad (14)$$

where h is the hidden state of the pharmacophore or ligand atoms, E_{ij} is the edge feature between node i and its neighboring node $j \in \mathcal{N}_K(i)$ and D_{ij} denotes the direction-matching feature between the pharmacophore and ligand atom (detailed in the next section). A transformer-based neural network ϕ_K is exploited to compute the update for hidden state $h_{K,i}$, ensuring that both the spatial and directional relationships between pharmacophore features and ligand atoms are properly captured.

The updates for the fully connected molecular graph $\mathcal{G}_\mathrm{L}$ that models the molecular structure and intraligand atomic interactions is computed as

$$\Delta h_{L,i} = \sum_{j \in \mathcal{N}_L(i)} \phi_L(h_i, h_{j'}, e_{ij'}, t), \quad (15)$$

where e represents the hidden state of the bond embedding between atoms. The neural network ϕ_L, same as the architecture ϕ_K, takes these features as inputs and computes the hidden state update $\Delta h_{L,i}$ for $\mathcal{G}_L$.

Then, we leverage two transformer-based neural networks, $\phi_{x,K}$ and $\phi_{x,L}$, to obtain the coordinates updates from both graph $\mathcal{G}_K$ and $\mathcal{G}_L$. Notably, $\phi_{x,K}$ implements a cross-attention mechanism to fuse

the updated pharmacophore and ligand feature representations, along with their corresponding spatial coordinates. Here, messages between nodes m_{ji} are obtained from a multilayer perceptron (MLP) ϕ_d.

$$\Delta \mathrm{x}_{K,i} = \sum_{j \in \mathcal{N}_K(i)} (\mathrm{x}_j - \mathrm{x}_i) \phi_{x,K}(h_i, h_j, \|\mathrm{x}_i - \mathrm{x}_j\|, E_{ij}, D_{ij}, t) \tag{16}$$

$$\Delta \mathrm{x}_{L,i} = \sum_{j' \in \mathcal{N}_L(i)} (\mathrm{x}_{j'} - \mathrm{x}_i) \phi_{x,\mathrm{L}}(h_i, h_{j'}, \|\mathrm{x}_{j'} - \mathrm{x}_i\|, \mathrm{m}_{j'i}, t) \tag{17}$$

$$m_{ji} = \phi_d(\|\mathrm{x}_j - \mathrm{x}_i\|, e_{ji}). \tag{18}$$

Then, the model aggregates all the modification for the hidden states and the node coordinates and refresh the heterogeneous graph $\mathcal{G} = \{\mathcal{G}_K, \mathcal{G}_L\}$, sequentially updating bond types, atom features and atom coordinates in each layer. Formally, we first update the bond types with a transformer ϕ_e as

$$e_{ji} = e_{ji} + \sum_{k \in \mathcal{N}_L(j) \setminus \{i\}} \phi_e(h_i, h_j, h_k m_{kj}, m_{ji}, t). \tag{19}$$

The atom features and coordinates are updated as

$$h_i = h_i + \phi_h(\Delta h_{K,i} + \Delta h_{L,i}, t) \tag{20}$$

$$\mathrm{x}_i = \mathrm{x}_i + (\Delta \mathrm{x}_{K,i} + \Delta \mathrm{x}_{L,i}) \cdot \mathbb{1}_{\mathrm{mol}}, \tag{21}$$

where ϕ_h is a single-layer linear, and $\mathbb{1}_{\mathrm{mol}}$ represents a selector that ensures only the coordinates of ligand atoms are updated while the pharmacophore coordinates remain fixed.

After multiple iterations of the update process, we can directly obtain the final coordinates $\hat{\mathrm{x}}_0$. For the prediction of atomic types $\hat{\mathrm{v}}_0$ and bond types $\hat{\mathrm{b}}_0$, we apply two MLPs followed by a Softmax activation function to the final ligand hidden states h_i and e_{ij} as

$$\hat{\mathrm{v}}_i = \mathrm{Softmax}(\mathrm{MLP}(h_i)) \tag{22}$$

$$\hat{\mathrm{b}}_{ij} = \mathrm{Softmax}(\mathrm{MLP}(e_{ij} + e_{ji})). \tag{23}$$

Direction matching module. For ligand–pharmacophore mapping, five pharmacophore features (including hydrogen-bond acceptor, hydrogen-bond donor, metal coordination, aromatic ring and halogen bonding) involve direction matching. Here, we introduce directional vectors between ligand atoms and pharmacophore features during the Equivariant Graph Neural Network (EGNN) message-passing process (Fig. 1c). Specifically, for any atom of the ligand, we define two key directional vectors. The first vector, $\mathbf{v}_1$, points from the atom to the center of neighboring atoms within 3 Å radius, which is designed to capture the local geometric configuration of the ligand. The second vector, $\mathbf{v}_2$, points from the atom to the associated pharmacophore feature. These two vectors are further complemented by a third vector, $\mathbf{v}_3$, which represents the pharmacophore's orientation.

To efficiently utilize the geometric information encoded by the three vectors in the training process, we computed the dot products of each pair of vectors, which are concatenated as a unified feature representation D_{ij}. It is then embedded into the networks ϕ_K and $\phi_{x,K}$ to update both the atom features and coordinates.

Molecular size module. In one-shot molecular generation tasks, it is essential to determine the number of atoms before the diffusion and denoising processes. In PhoreGen, we incorporate a module specifically designed to predict the atom count based on the pharmacophore models, ensuring that the generated molecules maintain appropriate steric and shape constraints while exhibiting chemical diversity.

We exploit a neural network ϕ_c to encode the pharmacophore's feature information, which includes both the pharmacophore descriptors h_p and their 3D coordinates x_p.

$$z_{\mathrm{count}} = \phi_c(h_p, x_p). \tag{24}$$

From this encoded representation z_{count}, two separate MLPs are employed: one predicts the upper bound of the atom count N_{upper} based on the full feature z_{count}, while the second, which excludes exclusion volume constraints, predicts the lower bound N_{lower} to ensure the generated molecule has the fundamental scaffold that matches the pharmacophore features (Fig. 1d).

During the sampling phase, the predicted upper and lower bounds are used to define a normal distribution from which the final atom count is sampled. This approach allows the model to account for molecular size diversity while ensuring that the generated ligands conform to the pharmacophore constraints. The distribution is constructed as

$$N_{\mathrm{atoms}} \sim \mathcal{N}(\mu = \frac{N_{\mathrm{upper}} + N_{\mathrm{lower}}}{2}, \sigma^2) \tag{25}$$

where ~ denotes 'distributed as' and σ is the standard deviation for the user-defined hyperparameter.

Loss function. PhoreGen can be trained by optimizing a composite loss function that consists of four primary terms: losses for atom coordinates, atom types, bond types and atom counts. For the atom coordinates loss, we can minimize the mean squared error (MSE) between the predicted denoised coordinates $\hat{x}_0$ and the ground-truth coordinates x_0. This loss can be written as

$$\mathcal{L}_{\mathrm{coord}}^{t-1} = \frac{1}{2\tilde{\beta}_t^2} \|\tilde{\mu}_t(x_t, x_0) - \mu_\theta(M_t, t, \mathcal{P})\|^2 = \gamma_t \|x_0 - \hat{x}_0\|^2, \tag{26}$$

where γ_t is a time-dependent weight parameter based on the noise schedule, while in practice, we usually set $\gamma_t = 1$.

The atom type and bond type prediction are treated as two categorical classification tasks. The cross-entropy loss for both can be formulated as

$$\mathcal{L}_{\mathrm{atom}}^{t-1} = \sum_k \mathcal{C}(v_t, v_0)_k \log \frac{\mathcal{C}(v_t, v_0)_k}{\mathcal{C}(v_t, \hat{v}_0)_k} \tag{27}$$

$$\mathcal{L}_{\mathrm{bond}}^{t-1} = \sum_k \mathcal{C}(b_t, b_0)_k \log \frac{\mathcal{C}(b_t, b_0)_k}{\mathcal{C}(b_t, \hat{b}_0)_k}. \tag{28}$$

For the atom count loss, we utilize a quality-driven prediction interval loss inspired by recent advancements in prediction interval methodologies[51]. Specifically, the atom count loss is defined to ensure that the predicted upper and lower bounds contain the true number of atoms in the molecule, while also minimizing the width of the predicted interval. The loss is calculated as

$$k_h = \mathrm{relu}(\mathrm{sign}(N_{\mathrm{upper}} - N_{\mathrm{true}})) \cdot \mathrm{relu}(\mathrm{sign}(N_{\mathrm{true}} - N_{\mathrm{lower}})) \tag{29}$$

$$k_s = \mathrm{sigmoid}(N_{\mathrm{upper}} - N_{\mathrm{true}}) \cdot \mathrm{sigmoid}(N_{\mathrm{true}} - N_{\mathrm{lower}}) \tag{30}$$

$$\mathrm{MPIW}_c = \mathrm{sum}((N_{\mathrm{upper}} - N_{\mathrm{lower}}) \cdot k_h / \mathrm{sum}(k_h) \tag{31}$$

$$\mathcal{L}_{\mathrm{count}}^{t-1} = \mathrm{MPIW}_c + \lambda(\mathrm{relu}((1-\alpha) - \mathrm{sum}(k_s))^2), \tag{32}$$

where λ controls the penalty for misclassification and Mean Prediction Interval Width for captured points (MPIW_c; referring to the average

interval width for data points successfully covered by the predicted intervals) controls the interval width to ensure it remains narrow.

The final loss function is a weighted sum of these individual losses

$$\mathcal{L}^{t-1} = \mathcal{L}^{t-1}_{\text{coord}} + \lambda_1 \mathcal{L}^{t-1}_{\text{atom}} + \lambda_2 \mathcal{L}^{t-1}_{\text{bond}} + \mathcal{L}^{t-1}_{\text{count}}, \quad (33)$$

where λ_1 and λ_2 are the weights assigned to each loss term. These weights are tuned to balance the different loss components effectively.

Guidance for structural validity. To ensure that the generated molecules maintain structural validity and conform to the pharmacophore constraints, we apply a guidance strategy during the sampling phase, inspired by the classifier guidance framework used in diffusion models[52]. This strategy introduces additional drift terms to adjust the sampling to satisfy two key structural requirements: (1) ensuring spatial consistency by aligning the molecule's center with the pharmacophore center (excluding the exclusion volume) and (2) constraining bond lengths between atoms capable of forming bonds to remain within realistic chemical limits. Specifically, the guidance is introduced via the gradient of the log-probability of the specific constraint condition y. Mathematically, this is expressed as

$$\nabla_{x_t} \log P(x_t|\, y) = \nabla_{x_t} \log P(x_t) + \nabla_{x_t} \log P(\, y|x_t), \quad (34)$$

where x is the generated atom coordinates. The first term represents the original score function for sampling, while the second term, $\nabla_{x_t} \log P(y|x_t)$, adds a drift term that adjusts the sampling based on the constraint condition y.

To guarantee spatial consistency, the generated molecule's center ρ_{mol} is aligned with the center of the pharmacophore ρ_{phore}, excluding the exclusion volumes. This alignment is ensured through the following center drift function:

$$\text{drift}_{\text{center}} = -\nabla_{x_t} \|\rho_{\text{mol}} - \rho_{\text{phore}}\|^2. \quad (35)$$

To maintain chemical validity, we enforce a bond length constraint that ensures connected atoms have distances within a physically plausible range. This constraint is imposed via

$$\text{drift}_{\text{bond}} = -\nabla_{x_t} \sum_{(i,j)\in\varepsilon} \max(0, d_{ij} - d_{\max}) + \max(0, d_{\min} - d_{ij}), \quad (36)$$

where ε is the set of bonded atom pairs, $d_{\min}$ and $d_{\max}$ are hyperparameters approximating the reasonable bond length range, which we set to 1.0 and 3.0, respectively.

Model details. PhoreGen consists of two primary learnable components: the atom count prediction network and the denoising network. The atom count prediction network employs two MLPs with ReLU and sigmoid activation functions. These MLPs predict the lower and upper bounds of the molecule's atom count, respectively, based on the pharmacophore features and exclusion-sphere information. The denoising network is composed of three transformer-based modules for atom updates, coordinate updates and bond updates. Each module aggregates graph information through graph attention mechanisms, where the key, value and query vectors are computed via two-layer MLPs equipped with LayerNorm and ReLU activation. These three modules collectively form a single layer, and the network stacks six such layers with a hidden dimension of 128 per layer.

Training. PhoreGen is trained using the Adam optimizer with a learning rate of 1×10^{-4}, betas of (0.9, 0.999), and a weight decay of 1×10^{-12}. A plateau learning rate scheduler was applied, with a decay factor of 0.9 and a patience setting of 20 epochs. The batch size for training is set to 4. The training process is conducted on eight NVIDIA GTX 4090 GPUs. The pretraining phase lasted for 60 epochs over 6 days, followed by fine-tuning for 500 epochs across 3 days.

Dataset preparation

To enable the model to learn the essence of 3D ligand–pharmacophore mapping, we constructed three datasets, LigPhore, CpxPhore and DockPhore, to facilitate two-phase training. In LigPhore, the pharmacophore models were derived from energetically favorable ligand conformations, generated by considering both pharmacophore feature diversity and ligand structural diversity. LigPhore finally contains a broader range of perfectly matched ligand–pharmacophore pairs. The complex-derived pharmacophores (that is, CpxPhore and DockPhore datasets) are extracted from experimentally validated crystal complex structures and cross-docking- generated complexes, respectively; these pharmacophores reflect real but biased ligand–pharmacophore mapping scenarios, constrained by the specific binding contexts of the source complexes. The t-SNE analysis of the chemical and pharmacophoric feature space distributions for ligand-derived and complex-derived pharmacophore models is presented in Supplementary Fig. 11. This analysis reveals that ligand-derived and complex-derived pharmacophores exhibit both similarity and difference within the feature space; by comparison, the ligand-derived pharmacophores show a broader distribution of feature diversity.

Specifically, based on our recently developed first LigPhore dataset, which was used to train a model for 3D ligand–pharmacophore mapping[53], we updated the LigPhore dataset with the following improvements, aiming to better facilitate practical pharmacophore-oriented 3D molecular generation. First, we modified our previously developed anchor pharmacophore tool, AncPhore[54], to generate 12 pharmacophore features for each ligand conformation, including metal coordination (MB), hydrogen-bond acceptor (HA), hydrogen-bond donor (HD), positively charged feature (PO), negatively charged feature (NE), aromatic ring (AR), hydrophobic feature (HY), halogen bonding feature (HB), four covalent features (CV1, CV2, CV3 and CV4) and exclusion spheres (Supplementary Table 2). Then, we proportionately sampled 499,672 representative ligands from the ligand groups clustered by structural similarity. To improve feature balance, we sampled 60,000 ligands for generating pharmacophore models containing the underrepresented features, including PO, NE, XB, CV1, CV2, CV3 and CV4. As an exploratory effort, we also included 60,000 boron-containing compounds. Ultimately, the updated LigPhore dataset comprises 2,398,776 3D ligand–pharmacophore pairs.

The CpxPhore dataset contains real-world 3D ligand–pharmacophore pairs, which were derived from experimentally determined complex structures in PDBbind v2020 (refs. 55,56) using our modified AncPhore program[54]. The samples with a ligand atom count more than 60 are excluded to guarantee the data quality, resulting in 13,585 ligand–pharmacophore pairs in the CpxPhore dataset. The DockPhore dataset comprises near real-world 3D ligand–pharmacophore pairs, which were derived from the CrossDocked2020 dataset[42] by the following procedure. Briefly, we analyzed the 22.5 million docked protein–ligand pairs, and filtered out docked conformations with r.m.s.d. greater than 1 Å, and split the training and test sets to ensure that the sequence similarity was less than 30%. We next used the modified AncPhore[54] to generate pharmacophore models, and further filtered out ligand–pharmacophore pairs containing fewer than 4 or more than 78 ligand atoms, as well as pairs with fewer than 3 or more than 15 pharmacophore features. This eventually led to 84,767 samples in the DockPhore dataset. Finally, the LigPhore, CpxPhore and DockPhore datasets were separately split into corresponding training set, validation set and test set (Supplementary Table 1). All the pharmacophore models in the test set have no overlap or high pharmacophore similarity with those in the training set (Supplementary Fig. 5).

Evaluation metrics

To comprehensively evaluate the quality of the PhoreGen-generated molecules, we use the following metrics:

(1) The mapping score (S_{mapping}) evaluates the alignment between the 3D ligand (L) and the pharmacophore model (P), considering both feature matching and exclusion-sphere collision. It is calculated using an in situ max-matching approach (equations (37–40))

$$S_{\text{mapping}} = \frac{V_{\text{overlap}}}{V_{\text{ref}}} - S_{\text{clashing}} \tag{37}$$

$$V_{\text{overlap}} = \sum_{i=1}^{n} C_i W_i \lambda_i f(\theta) \exp\left(\frac{-d_{i,L-i,P}^2}{\sigma_{i,L} + \sigma_{i,P}}\right) \tag{38}$$

$$S_{\text{clashing}} = \text{MIN}\left(\frac{V_{\text{overlapEX}}}{\epsilon}, 1\right) \tag{39}$$

$$f(\theta) = \begin{cases} \cos(\theta - \theta_0) & \text{for HA, HD and MB} \\ |\cos(\theta)| & \text{for AR} \\ 1 & \text{for other features} \end{cases}, \tag{40}$$

where V_{overlap} is the total overlap volume between the ligand and reference pharmacophore features, computed as the sum of individual feature overlap volumes. This calculation incorporates scaling factors (C_i), basic weights (W_i), chemical group weights (λ_i), directional differences ($f(\theta)$), tolerance ranges ($\sigma_{i,L}, \sigma_{i,P}$) and the distance between matched pharmacophore pairs ($d_{i,L-i,P}$). The directional parameter θ_0 is set to 0 when the number of root atoms (that is, the number of neighboring heavy atoms associated with a specific pharmacophore feature) equals 1, or $\frac{\pi}{3}$ when it is greater than 1. V_{ref} represents the total volume of the reference pharmacophore features. S_{clashing} represents the collision score of the ligand with the exclusion spheres, $V_{\text{overlapEX}}$ refers to the sum of ligand atom volumes overlapping with reference exclusion spheres, and ϵ refers to the maximum tolerance for ligand collisions with exclusion spheres, set to 500.

(2) The feature mapping rate (R_{mapping}) evaluates the percentage of matched pharmacophore feature pairs, as computed by equation (41)

$$R_{\text{mapping}} = \frac{n}{N_{\text{ref}}}, \tag{41}$$

where n is the number of pharmacophore pairs matched within their respective tolerance ranges and N_{ref} is the total number of pharmacophore features in the reference model. Here, two pharmacophore features are considered as a matched pair if their spatial separation falls within the combined tolerance range.

(3) The exclusion-sphere clashing rate (S_{clashing}) quantifies the extent of collisions between the ligand and exclusion spheres (that is, regions that can represent protein residues), thereby reflecting the degree of steric clashes between the generated molecules and protein atoms. The clashing rate is determined by the clashing score S_{clashing} and calculated using equation (39).

(4) PoseBusters examines the chemical and physical validity of a molecule's 3D conformation by evaluating a comprehensive set of criteria, including sanitizability, full atom connectivity, valid bond lengths and angles, absence of internal steric clashes, flat aromatic rings and double bonds, low internal energy (using the universal force field), correct valence and kekulizability, thereby ensuring adherence to realistic structural and geometric constraints.

(5) Diversity accesses the structural variability of the generated molecules by calculating their average pairwise Tanimoto dissimilarity score.

(6) Validity refers to the proportion of generated 3D molecules that pass the sanitization process in RDKit, ensuring that all atomic connections are intact and bond valences are accurate.

(7) QED provides an estimate of the drug-likeness of a molecule by integrating several molecular properties into a single score.

(8) SA indicates the molecular synthetic accessibility.

(9) Lipinski checks how many of Lipinski's rules a molecule follows.

(10) log$P$ measures the lipid–water partition coefficient of a molecule, providing insight into the lipophilicity of the molecule.

(11) Time is the average time to generate 100 valid molecules for each pharmacophore or target.

(12) The docking score evaluates the binding affinity between a generated molecule and its target pocket using the AutoDock Vina re-docking process.

(13) High affinity assesses the percentage of generated molecules that show better binding affinities, as indicated by docking scores, than the reference molecule for a given protein pocket.

(14) IFP similarity calculates the cosine similarity of interaction fingerprints between the generated molecule and the reference molecule, reflecting the degree of similarity in their binding interactions.

Baselines

PhoreGen was compared with seven representative SBMG models: AR[11], Pocket2Mol[12], DeepICL[15], TargetDiff[13], DiffSBDD[25], PMDM[14] and DrugFlow[39], alongside two additional models, ShEPhERD[40] and PGMG[32]. AR, Pocket2Mol and DeepICL are autoregressive methods that employ graph neural networks to sequentially generate molecules atom by atom within the protein pocket; notably, DeepICL leverages noncovalent interactions (for instance, hydrophobic, hydrogen bonding, salt bridge and π–π stacking) to guide molecular generation. TargetDiff, DiffSBDD and PMDM are diffusion-based models that generates 3D molecules through an iterative denoising process. DrugFlow integrates continuous flow matching with discrete Markov bridges, offering a hybrid approach to molecular generation. ShEPhERD employs a diffusion framework to jointly model molecular graphs, shapes, electrostatic surfaces and pharmacophores, enabling the conditional generation of 3D molecules. PGMG uses pharmacophores as inputs to guide molecular generation (in SMILES format). This comprehensive set of baseline models allows for a thorough evaluation of PhoreGen across different generative strategies and conditioning paradigms.

Chemical synthesis

The synthetic routes, methods and characterizations (NMR, HRMS and HPLC purity) for compounds **CB1**, **CB2**, **CN1** and **CN2** are given in Supplementary Notes and Supplementary Figs. 20–29.

Biological assays for MBLs/SBLs

Constructs, protein expression and purification. The SBL KPC-2 (amino acids 26–289) and OXA-48 (amino acids 25–265) and MBL NDM-1 (amino acids 1–270), VIM-2 (amino acids 27–266) and IMP-1 (amino acids 19–246) enzymes with N-terminal His-tags were expressed in *E. coli* Transetta (DE3) cells at 37 °C using LB medium with 50 mg ml^{-1} chloramphenicol and 50 mg ml^{-1} kanamycin (200 mM sorbitol and 5 mM betaine for KPC-2)[57–59]. When the OD_{600} reached 0.5–0.6, the temperature was lowered to 16–20 °C, and 0.3–0.5 mM isopropyl-β-D-thiogalactoside was added to induce protein expression for 18–20 h. Cells were collected by centrifugation (1,753*g*, 30 min), resuspended in lysis buffer (20 mM Tris–HCl, pH 8.0, 250 mM NaCl)

and lysed with an ultrahigh-pressure homogenizer. The cell lysates were centrifuged at 15,777*g* for 30 min and the supernatant was then purified using nickel-ion affinity chromatography. The His-tags of NDM-1, VIM-2, IMP-1, KPC-2 and OXA-48 proteins were further removed by TEV protease cleavage (1:100 w/w) at 4 °C overnight and captured on Ni-NTA resin (Roche). Proteins were then loaded onto a Superdex S75 column equilibrated in the corresponding buffers and concentrated for crystallization. The protein concentrations were determined using a NanoDrop 2000 spectrophotometer (Thermo Scientic). All the proteins were stored at −80 °C.

Activity assays for MBLs/SBLs. Activity assays were performed in 96-well plates using the FC-5 fluorogenic substrate[60]. The assay conditions include enzymes (1 nM KPC-2, 25 nM OXA-48, 0.2 nM NDM-1, 0.2 nM VIM-2 or 0.2 nM IMP-1), 5 mM FC-5 and appropriate concentrations of **CB1** or **CB2** in threefold dilutions. The enzymes were pre-incubated with compounds in assay buffers supplemented with 0.01% Triton at room temperature for 10 min. The reactions were initiated by adding FC-5 and monitored by detecting the fluorescence changes (λ_{ex} = 380 nm and λ_{ex} = 460 nm) using a microplate reader. The assays were performed in triplicates. The data were then fitted using a four-parameter logistic equation in GraphPad Prism 8.02 to determine IC_{50} values.

X-ray crystallography. The OXA-48:**CB1** crystal structure (PDB code 9KSA) was obtained by co-crystallization experiments. Freshly purified OXA-48 protein (10 mg ml^{-1}) in crystallization buffer (20 mM HEPES, pH 7.5, 100 mM NaCl) was incubated with 3.58 mM **CB1** for 30 min at 4 °C before setting up crystallization drops. Crystallization drops were prepared by mixing the OXA-48:**CB1** solution and reservoir solution in a 1:1 ratio. The OXA-48:**CB1** crystals appeared after 2–3 weeks under the conditions of 5% 1-butanol, 0.1 M HEPES pH 7.5 and 15% PEG8000. Crystals were cryoprotected with ~20% (v/v) glycerol and flash-cooled in liquid nitrogen. X-ray diffraction data were collected at the Shanghai Synchrotron Radiation Facility BL19U1 beamline[61]. Data collection and refinement statistics are provided in Supplementary Table 29.

Biological assays for Naam

Plasmid construction and protein expression. Full-length cDNAs encoding *Drosophila melanogaster* Naam (NP_001262738; amino acids 1–357), *Myzus persicae* Naam (XP_022168570.1; amino acids 1–333), *Bemisia tabaci* Naam (XP_018907173; amino acids 1–349) and the auxiliary enzyme glutamate dehydrogenase (GDH, WP_003420866; amino acids 1–421) were subcloned into the pET-28TEV-pccdB expression plasmid. All recombinant cDNAs contain a C-terminal 6×His tag. The plasmids were synthesized by General Biosystems (Anhui) Co. Ltd, and their sequences were confirmed by DNA sequencing. The proteins were induced in *E. coli* BL21(DE3) by adding 0.3 mM isopropyl-β-D-thiogalactopyranoside when the cell density reached an optical density of 0.6 and then overexpressed for 16 h at 24 °C. After that, the cells were collected and lysed in a buffer (Naam expression buffer contains 20 mM Tris–HCl pH 7.5 and 300 mM NaCl; GDH expression buffer contains 20 mM Na_2HPO_4, 500 mM NaCl, pH 8.0) and removed the lysates by centrifugation at 15,777*g* for 60 min. The supernatants were loaded on a Ni-NTA column equilibrated with expression buffer. By washing with a buffer containing expression buffer and 10–60 mM imidazole, the target proteins were eluted from the affinity resin with a buffer containing expression buffer and 250 mM imidazole. Finally, the target proteins were concentrated and washed with expression buffer at 2,219*g*. The concentrated samples were collected and stored at −80 °C for enzyme activity testing.

Activity assays for Naam enzymes. The activity assays for three Naam enzymes were performed within 100 µl reaction mixtures (pH 7.5, 25 °C), which contain 15 nM *Drosophila melanogaster* Naam (15 nM *Myzus persicae* Naam 30 nM or *Bemisia tabaci* Naam), 1 mM α-ketoglutarate, 0.5 mM NADH, 1.377 µM GDH and 0.5 mM nicotinamide in an ultraviolet-transparent 96-well microplate. The enzymes were initially incubated with a series of threefold concentrations of **CN1** or **CN2** for a duration ranging from 0 to 120 min at 4 °C. Next, the α-ketoglutarate, NADH and glutamate dehydrogenase were added, and the mixtures were then shaken for 3 min at room temperature. Reactions were initiated by addition of nicotinamide, and the microplate reader monitored the decrease absorbance of NADH at 340 nm for a total of 900 s with 15 s intervals. The IC_{50} values were obtained using GraphPad Prism 8.02.

Differential scanning fluorimetry assays. The assays were conducted using a real-time fluorescence quantitative PCR instrument, with reactions carried out in a 20 µl mixture containing 20 µM *D. melanogaster* Naam (10 µM *M. persicae* Naam or 10 µM *B. tabaci* Naam), 50 µM **CN1** or **CN2**, SYPRO orange dye in a buffer of 20 mM Tris–HCl pH 7.5 and 300 mM NaCl. The mixtures were incubated for 1 h at 4 °C, then scanned from 37 °C to 95 °C at a scan rate of 1 °C s^{-1}. The melting temperature (T_m) was determined by fitting the negative rate of change of fluorescence with temperature using GraphPad Prism 8.02.

Analysis of inhibitors covalently bound to Naam. The covalent product of the inhibitor cyano group with the Naam catalytic cysteine residue was detected by LC–MS/MS analysis. The inhibitor **CN1** or **CN2** (120 µM) was incubated with *D. melanogaster* Naam enzyme (25 µM) at 4 °C for 2 h in the buffer (20 mM Tris–HCl pH 7.5 and 300 mM NaCl). The resulting mixture was transferred to a pre-activated ultrafiltration membrane spin column and then washed by centrifugation three times with PBS, twice with TEAB (50 mM), and finally reconstituted with TEAB (50 mM). The solution was resuspended in a 500 mM dithiothreitol aqueous solution (final concentration 10 mM) and incubated at 37 °C for 60 min. For alkylation, 1 M IAA aqueous solution (final concentration 20 mM) was added, followed by incubation for 30 min at 25 °C in the dark. Subsequently, trypsin was added at a ratio of 1:50 for protein digestion. The reaction was incubated overnight and then terminated with 10% TFA. The washing buffers were combined and desalted using a C18 column. After evaporation, the samples were analyzed by LC–MS/MS.

X-ray crystallography. The *D. melanogaster* Naam:**CN2** crystal structure (PDB code 9U8M) was obtained by co-crystallization experiments. Freshly purified *D melanogaster* Naam (15 mg ml^{-1}) in crystallization buffer (20 mM Tris–HCl pH 7.5 and 300 mM NaCl) was incubated with 1.88 mM **CN2** for 60 min at 4 °C before setting up crystallization drops. Crystallization drops were prepared by mixing the *D. melanogaster* Naam:**CN2** solution and reservoir solution in a 1:1 ratio. The crystals appeared after 1–2 weeks under the conditions of 0.1 M HEPES, pH 7.5 and 1.0 M sodium citrate. Crystals were cryoprotected with ~25% (v/v) glycerol and flash-cooled in liquid nitrogen. The data were collected at the Shanghai Synchrotron Radiation Facility BL19U1 beamline[61]. Data collection and refinement statistics are provided in Supplementary Table 29.

Statistics and reproducibility

No statistical method was used to predetermine sample size. All models sampled 100 molecules per pharmacophore model or protein target, using a fixed random seed for consistency. Molecules that failed sampling or did not pass RDKit's sanitization step were excluded due to the requirements of the evaluation metrics.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The LigPhore, CpxPhore and DockPhore datasets for training and evaluation are available via Zenodo at https://zenodo.org/records/15518867 (ref. 62). The PDBBind dataset is available at http://pdbbind.org.cn. The CrossDocked2020 dataset is available at https://bits.csb.pitt.edu/files/crossdock2020/. The pharmacophore models for molecular generation can be created via our web server at https://phoregen.ddtmlab.org. The complex crystal structures (PDB codes 9KSA and 9U8M) are available via PDB at https://www.rcsb.org. Source data are provided with this paper.

Code availability

The source code of PhoreGen is available on our web server at https://phoregen.ddtmlab.org, via GitHub at https://github.com/ppjian19/PhoreGen and via Zenodo at https://zenodo.org/records/15518867 (ref. 62) under an open-source license.

References

1. Berdigaliyev, N. & Aljofan, M. An overview of drug discovery and development. *Future Med. Chem.* **12**, 939–947 (2020).
2. Goodnow, J. R. A. Hit and lead identification: integrated technology-based approaches. *Drug Discov. Today* **3**, 367–375 (2006).
3. Sadybekov, A. V. & Katritch, V. Computational approaches streamlining drug discovery. *Nature* **616**, 673–685 (2023).
4. Catacutan, D. B. et al. Machine learning in preclinical drug discovery. *Nat. Chem. Biol.* **20**, 960–973 (2024).
5. Du, Y. et al. Machine learning-aided generative molecular design. *Nat. Mach. Intell.* **6**, 589–604 (2024).
6. Cheng, Y. et al. Molecular design in drug discovery: a comprehensive review of deep generative models. *Brief. Bioinformatics.* **22**, bbab344 (2021).
7. Bian, Y. & Xie, X. Q. Generative chemistry: drug discovery with deep learning generative models. *J. Mol. Model.* **27**, 71 (2021).
8. Hoogeboom, E. et al. Equivariant diffusion for molecule generation in 3D. In *International Conference on Machine Learning* (eds Chaudhuri, K. et al.) (PMLR, 2022).
9. Xu, M. et al. Geometric latent diffusion models for 3D molecule generation. In *International Conference on Machine Learning* (eds Chaudhuri, K. et al.) (PMLR, 2022).
10. Huang, L. et al. MDM: molecular diffusion model for 3D molecule generation. In *Association for the Advancement of Artificial Intelligence* (AAAI Press, 2022).
11. Luo, S. et al. A 3D generative model for structure-based drug design. In *Conference on Neural Information Processing Systems* (eds Ranzao, M. et al.) (OpenReview.net, 2021).
12. Peng, X. et al. Pocket2Mol: efficient molecular sampling based on 3D protein pockets. In *International Conference on Machine Learning* (eds Chaudhuri, K. et al.) (PMLR, 2022).
13. Guan, J. et al. 3D equivariant diffusion for target-aware molecule generation and affinity prediction. In *International Conference on Learning Representations* (OpenReview.net, 2023).
14. Huang, L. et al. A dual diffusion model enables 3D molecule generation and lead optimization based on target pockets. *Nat. Commun.* **15**, 2657 (2024).
15. Zhung, W., Kim, H. & Kim, W. Y. 3D molecular generative framework for interaction-guided drug design. *Nat. Commun.* **15**, 2688 (2024).
16. Wu, P. et al. Guided diffusion for molecular generation with interaction prompt. *Brief. Bioinformatics* **25**, bbae174 (2024).
17. Lee, J., Zhung, W. & Kim, W. NCIDiff: non-covalent interaction-generative diffusion model for improving reliability of 3D molecule generation inside protein pocket. In *International Conference on Machine Learning* (OpenReview.net, 2024).
18. Zhavoronkov, A. et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* **37**, 1038–1040 (2019).
19. De Cao, N. & Kipf, T. MolGAN: an implicit generative model for small molecular graphs. In *International Conference on Machine Learning* (eds Dy, J. et al.) (PMLR, 2018).
20. Zhang, O. et al. ResGen is a pocket-aware 3D molecular generation model based on parallel multiscale modelling. *Nat. Mach. Intell.* **5**, 1020–1030 (2023).
21. Liu, M. et al. Generating 3D molecules for target protein binding. In *International Conference on Machine Learning* (eds Chaudhuri, K. et al.) (PMLR, 2022).
22. Luo, S. & Hu, W. Diffusion probabilistic models for 3D point cloud generation. In *IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2021).
23. Sun, J. et al. A critical revisit of adversarial robustness in 3D point cloud recognition with diffusion-driven purification. In *International Conference on Machine Learning* (eds Krause, A. et al.) (PMLR, 2023).
24. Lyu, Z. et al. A conditional point diffusion-refinement paradigm for 3D point cloud completion. In *International Conference on Learning Representations* (OpenReview.net, 2022).
25. Schneuing, A. et al. Structure-based drug design with equivariant diffusion models. *Nat. Comput. Sci.* **4**, 899–909 (2024).
26. Guan, J. et al. DecompDiff: diffusion models with decomposed priors for structure-based drug design. In *International Conference on Machine Learning* (eds Krause, A. et al.) (PMLR, 2023).
27. Huang, Z. et al. Interaction-based retrieval-augmented diffusion models for protein-specific 3D molecule generation. In *International Conference on Machine Learning* (OpenReview.net, 2024).
28. Alakhdar, A., Poczos, B. & Washburn, N. Diffusion models in de novo drug design. *J. Chem. Inf. Model.* **64**, 7238–7256 (2024).
29. Boike, L., Henning, N. J. & Nomura, D. K. Advances in covalent drug discovery. *Nat. Rev. Drug Discov.* **21**, 881–898 (2022).
30. Schaller, D. et al. Next generation 3D pharmacophore modeling. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **10**, e1468 (2020).
31. Imrie, F., Hadfield, T. E., Bradley, A. R. & Deane, C. M. Deep generative design with 3D pharmacophoric constraints. *Chem. Sci.* **12**, 14577–14589 (2021).
32. Zhu, H., Zhou, R., Cao, D., Tang, J. & Li, M. A pharmacophore-guided deep learning approach for bioactive molecular generation. *Nat. Commun.* **14**, 6234 (2023).
33. Bush, K. & Bradford, P. A. Interplay between β-lactamases and new β-lactamase inhibitors. *Nat. Rev. Microbiol.* **17**, 295–306 (2019).
34. Yang, Y. et al. Metallo-β-lactamase-mediated antimicrobial resistance and progress in inhibitor discovery. *Trends Microbiol.* **31**, 735–748 (2023).
35. Brem, J. et al. Imitation of β-lactam binding enables broad-spectrum metallo-β-lactamase inhibitors. *Nat. Chem.* **14**, 15–24 (2022).
36. Qiao, X. et al. An insecticide target in mechanoreceptor neurons. *Sci. Adv.* **8**, eabq3132 (2022).
37. Satorras, V. G., Hoogeboom, E. & Welling, M. E(n) equivariant graph neural networks. In *International Conference on Machine Learning* (eds Meila, M. et al.) (PMLR, 2021).
38. Buttenschoen, M., Morris, G. M. & Deane, C. M. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chem. Sci.* **15**, 3130–3139 (2024).
39. Schneuing, A. et al. Multi-domain distribution learning for de novo drug design. In *International Conference on Learning Representations* (OpenReview.net, 2025).
40. Adams, K. et al. ShEPhERD: diffusing shape, electrostatics, and pharmacophores for bioisosteric drug design. In *International Conference on Learning Representations* (OpenReview.net, 2025).

41. Eberhardt, J., Santos-Martins, D., Tillack, A. F. & Forli, S. AutoDock Vina 1.2.0: new docking methods, expanded force field, and Python bindings. *J. Chem. Inf. Model.* **61**, 3891–3898 (2021).
42. Francoeur, P. G. et al. Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. *J. Chem. Inf. Model.* **60**, 4200–4215 (2020).
43. Du, H. et al. CovalentInDB 2.0: an updated comprehensive database for structure-based and ligand-based covalent inhibitor design and screening. *Nucleic Acids Res.* **53**, D1322–D1327 (2025).
44. Hecker, S. J. et al. Discovery of cyclic boronic acid QPX7728, an ultrabroad-spectrum inhibitor of serine and metallo-β-lactamases. *J. Med. Chem.* **63**, 7491–7507 (2020).
45. Fyfe, P. K., Rao, V. A., Zemla, A., Cameron, S. & Hunter, W. N. Specificity and mechanism of *Acinetobacter baumanii* nicotinamidase: implications for activation of the front-line tuberculosis drug pyrazinamide. *Angew. Chem. Int. Ed.* **48**, 9176–9179 (2009).
46. Wang, C. & Rajapakse, J. C. Pharmacophore-guided de novo drug design with diffusion bridge. Preprint at https://doi.org/10.48550/arXiv.2412.19812 (2025).
47. Heider, J. et al. Apo2ph4: a versatile workflow for the generation of receptor-based pharmacophore models for virtual screening. *J. Chem. Inf. Model.* **63**, 101–110 (2023).
48. Seo, S. & Kim, W. Y. PharmacoNet: deep learning-guided pharmacophore modeling for ultra-large-scale virtual screening. *Chem. Sci.* **15**, 19473–19487 (2024).
49. Ho, J., Jain, A. & Abbeel, P. J. A. Denoising diffusion probabilistic models. In *Conference on Neural Information Processing Systems* (eds Larochelle, H. et al.) (Curran Associates Inc, 2020).
50. Peng, X., Guan, J., Liu, Q. & Ma, J. MolDiff: addressing the atom-bond inconsistency problem in 3d molecule diffusion generation. In *International Conference on Machine Learning* (eds Krause, A. et al.) (PMLR, 2023).
51. Pearce, T., Brintrup, A., Zaki, M. & Neely, A. High-quality prediction intervals for deep learning: a distribution-free, ensembled approach. In *International Conference on Machine Learning* (eds Dy, J. et al.) (PMLR, 2018).
52. Dhariwal, P. & Nichol, A. Diffusion models beat GANs on image synthesis. In *Conference on Neural Information Processing Systems* (eds Ranzao, M. et al.) (OpenReview.net, 2021).
53. Yu, J. et al. Knowledge-guided diffusion model for 3D ligand–pharmacophore mapping. *Nat. Commun.* **16**, 2269 (2025).
54. Dai, Q. et al. AncPhore: a versatile tool for anchor pharmacophore steered drug discovery with applications in discovery of new inhibitors targeting metallo-β-lactamases and indoleamine/tryptophan 2,3-dioxygenases. *Acta Pharm. Sin. B* **11**, 1931–1946 (2021).
55. Liu, Z. et al. Forging the basis for developing protein–ligand interaction scoring functions. *Acc. Chem. Res.* **50**, 302–309 (2017).
56. Su, M. et al. Comparative assessment of scoring functions: the CASF-2016 update. *J. Chem. Inf. Model.* **59**, 895–913 (2019).
57. Yan, Y.-H. et al. Discovery of 2-aminothiazole-4-carboxylic acids as broad-spectrum metallo-β-lactamase inhibitors by mimicking carbapenem hydrolysate binding. *J. Med. Chem.* **66**, 13746–13767 (2023).
58. Xiao, Y.-C. et al. Design and enantioselective synthesis of 3-(α-acrylic acid) benzoxaboroles to combat carbapenemase resistance. *Chem. Commun.* **57**, 7709–7712 (2021).
59. Wang, Y.-L. et al. Structure-based development of (1-(3′-mercaptopropanamido) methyl)boronic acid derived broad-spectrum, dual-action inhibitors of metallo- and serine-β-lactamases. *J. Med. Chem.* **62**, 7160–7184 (2019).
60. Van Berkel, S. S. et al. Assay platform for clinically relevant metallo-β-lactamases. *J. Med. Chem.* **56**, 6945–6953 (2013).
61. Xiao, Q. et al. Upgrade of crystallography beamline BL19U1 at the Shanghai Synchrotron Radiation Facility. *J. Appl. Crystallogr.* **57**, 630–637 (2024).
62. Peng, J. PhoreGen source code and data (LigPhore, CpxPhore, and DockPhore dataset and trained weights). *Zenodo* https://doi.org/10.5281/zenodo.15518867 (2025).

Acknowledgements

This work is financially supported by the National Natural Science Foundation of China (grant nos. 82122065, 82473845 and 82073698), the National Key R&D Program of China (grant no. 2023YFF1204901), the Sichuan Science and Technology Program (grant no. 2025YFHZ0085), the Foundation for Innovative Research Groups of the Natural Science Foundation of Sichuan Province (grant no. 2024NSFTD0026) and the Basic Research Foundation of Sichuan University (grant no. 2023SCUH0073). We thank the staff at beamline BL19U1 of the Shanghai Synchrotron Radiation Facility, National Facility for Protein Science (Shanghai, China), for their great support. We also thank the members of the Mass Spectrometry Platform (R. Wang and X. Wu) for proteomic techniques and data interpretation.

Author contributions

J.P., J.-L.Y., Z.-B.Y. and Y.-T.C. contributed equally to this work. G.-B.L. conceived, planned and supervised this study. J.P. and J.-L.Y. designed and trained the model supervised by G.-B.L.; J.P., J.-L.Y. and Y.-G.W. performed model validations. Z.-B.Y. and Y.-T.C. performed chemical synthesis. S.-Q.W., F.-B.M. and Y.-T.C. conducted protein production and bioactivity testing. J.P., J.-L.Y., Z.-B.Y., Y.-T.C. and G.-B.L. wrote the manuscript. G.-B.L. revised and polished the manuscript. All authors contributed to the final draft and approved the final version for submission.

Competing interests

The authors declare no competing interests.

Additional information

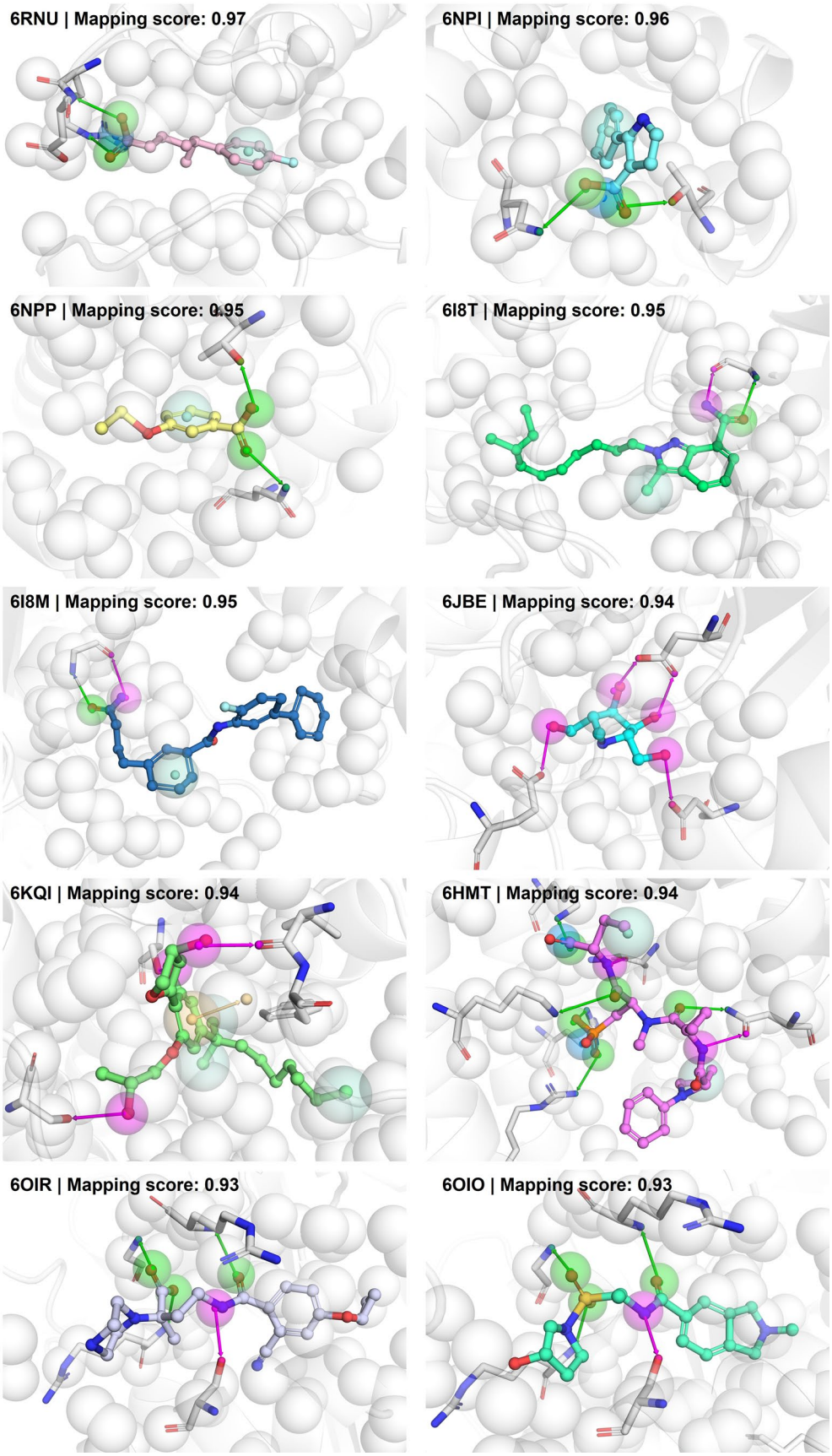
Extended data are available for this paper at https://doi.org/10.1038/s43588-025-00850-5.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43588-025-00850-5.

Correspondence and requests for materials should be addressed to Guo-Bo Li.

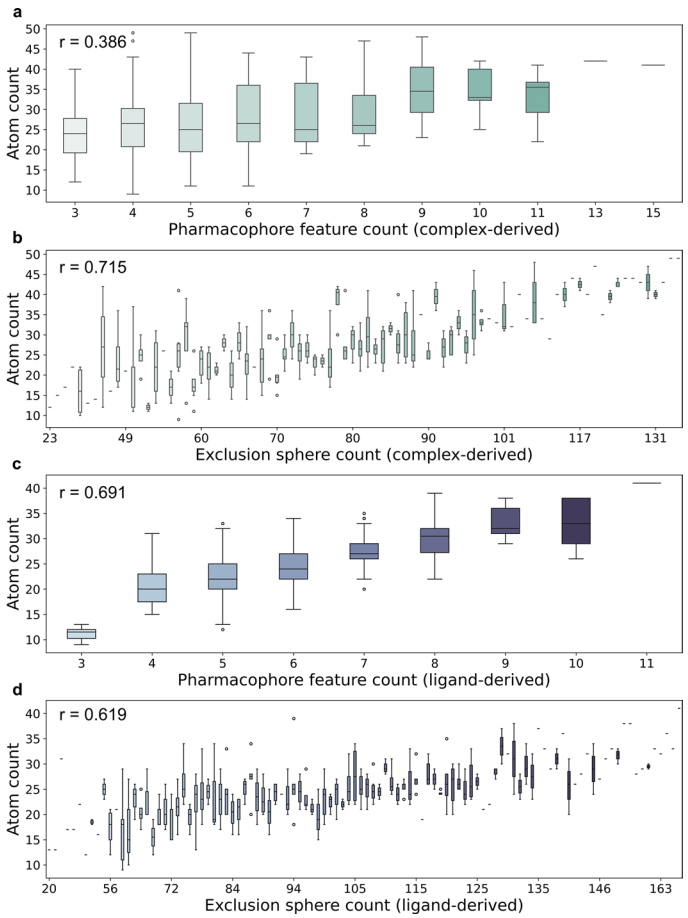
Peer review information *Nature Computational Science* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Kaitlin McCardle, in collaboration with the *Nature Computational Science* team. Peer reviewer reports are available.

https://doi.org/10.1038/s43588-025-00850-5



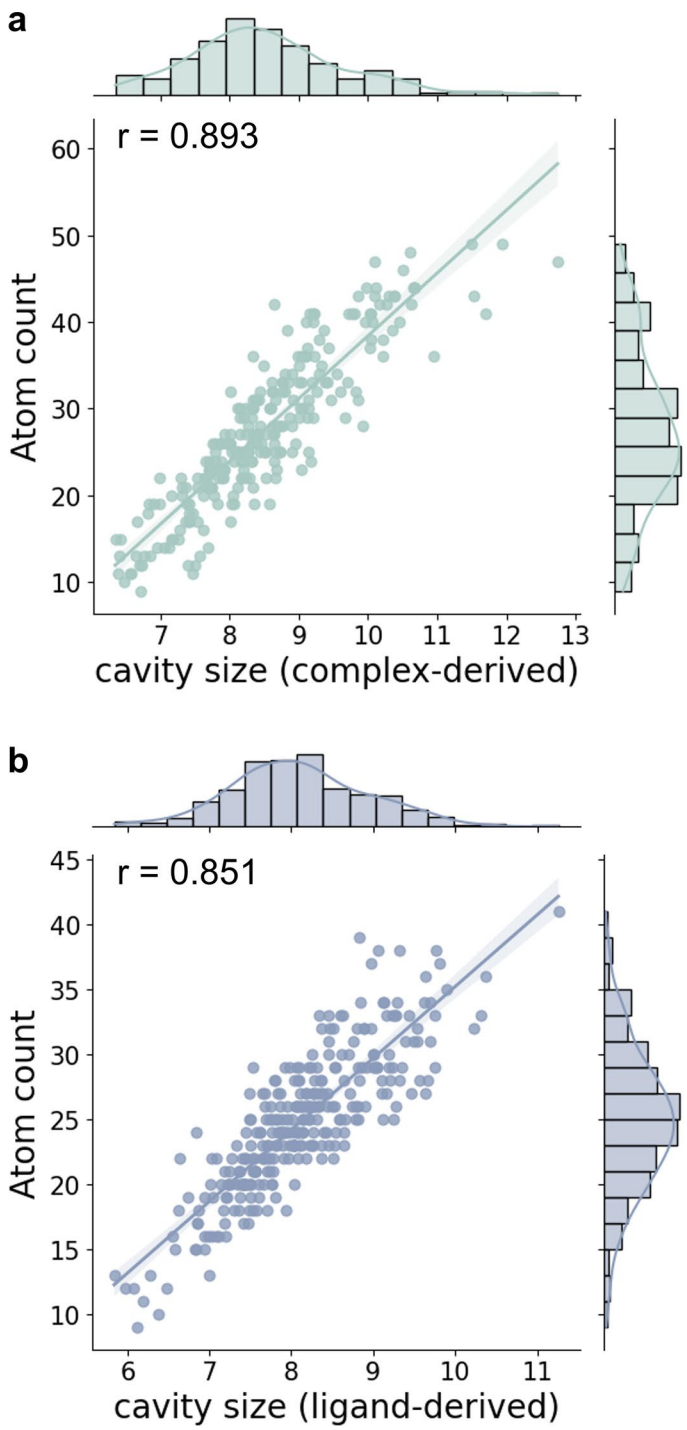
Extended Data Fig. 1 | View of the top-10 ranked generated molecules with their corresponding pharmacophore models. It reveals that for pharmacophore models with common, balanced feature compositions, PhoreGen can generate well-mapped, high-quality 3D molecules. Boron atoms, nitrogen atoms, oxygen atoms, phosphorus atoms, sulfur atoms, and chlorine atoms are pink, blue, red, darkorange, goldenrod, and limegreen respectively, while the other colors are carbon atoms.

https://doi.org/10.1038/s43588-025-00850-5



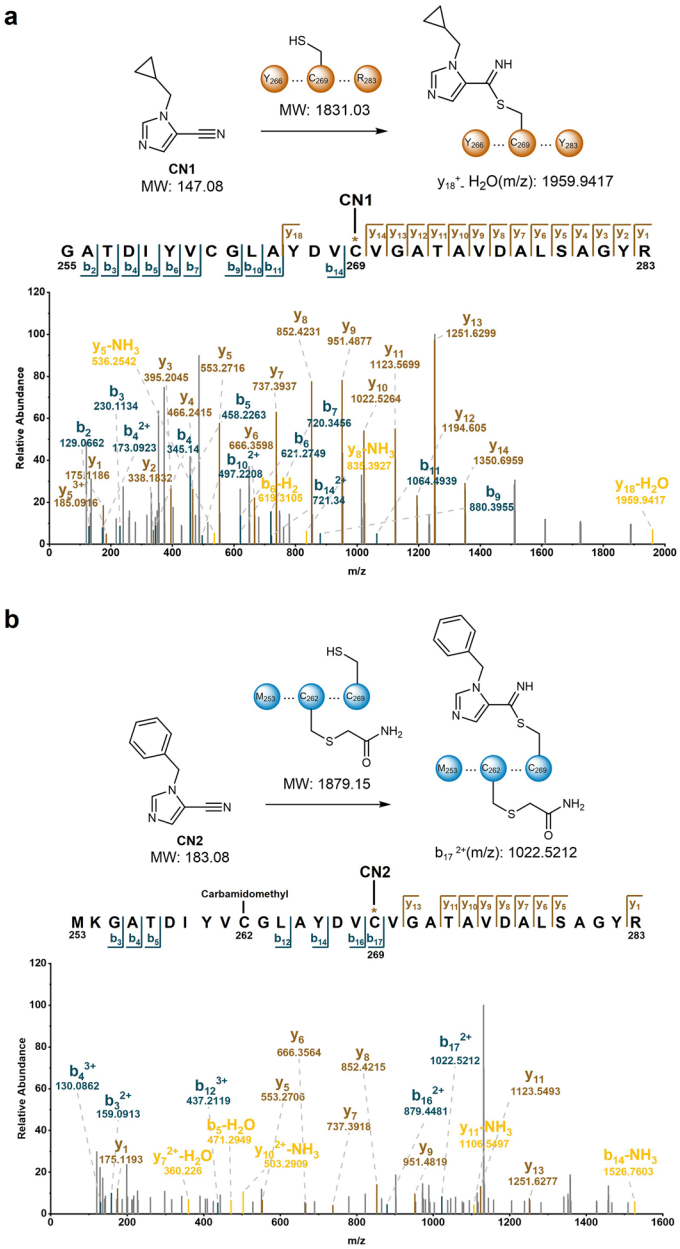
Extended Data Fig. 2 | Distributions of the pharmacophore feature and exclusion sphere counts versus the atom counts of generated molecules. **a**,**b** For the complex-derived pharmacophore models (that is, the CpxPhore test set), the atom counts of the generated molecules show correlation with the pharmacophore feature and exclusion sphere counts. **c**,**d** For the ligand-derived pharmacophore models (that is, the LigPhore test set), the generated molecule's atom counts also correlate with the feature and exclusion sphere counts. The sample size can be found in the Source Data file. Shadings indicate the feature or exclusion sphere counts, with darker shades representing higher counts. The box is the interquartile range (IQR), the line inside the box is the median, the whiskers represent data within ±1.5 × IQR, and the dots outside the whiskers are outliers. The r value refers to the Pearson correlation coefficient.

https://doi.org/10.1038/s43588-025-00850-5



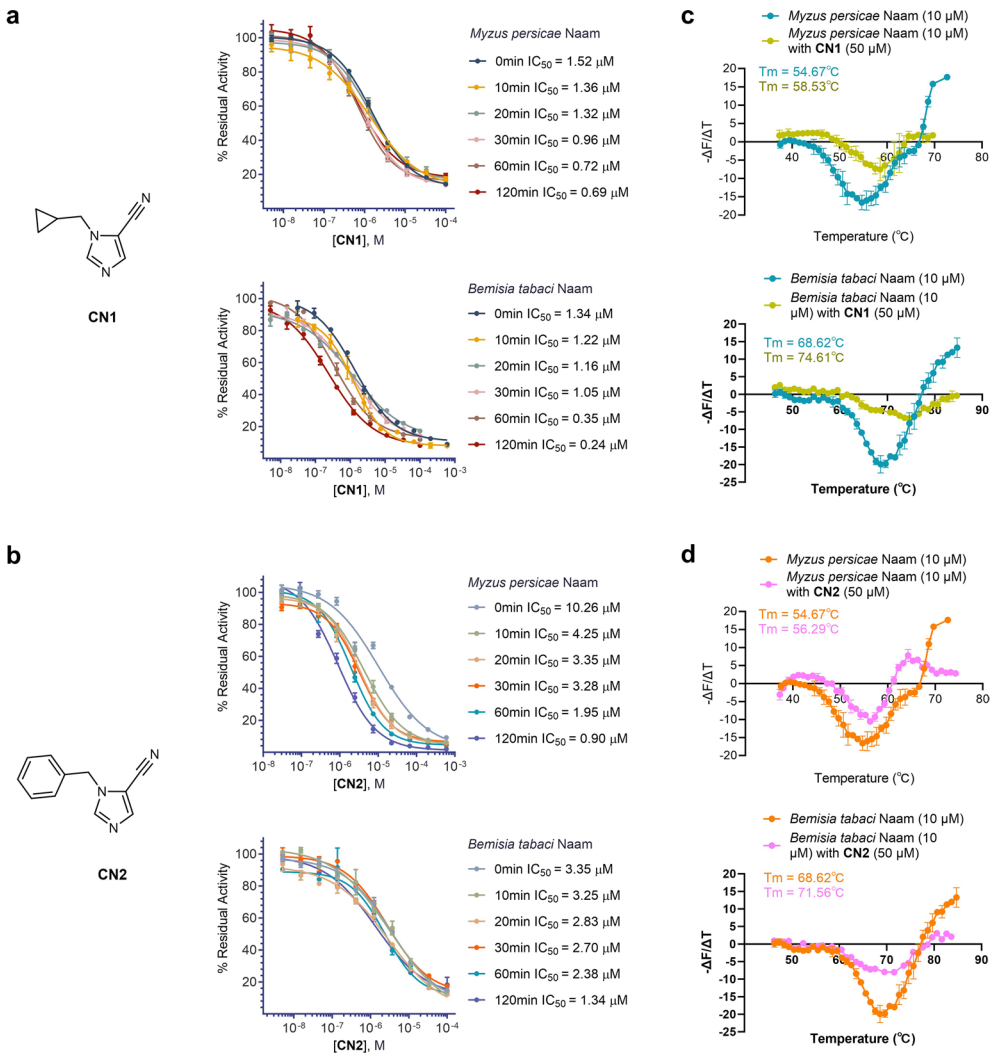
Extended Data Fig. 3 | Correlation between the cavity sizes of pharmacophore models and the atom counts of generated molecules. a,b For both the complex-derived (**a**) and ligand-derived (**b**) pharmacophore models, the atom counts of the generated molecules show a strong correlation with the corresponding cavity sizes, as indicated by the average distance between the pharmacophore features and exclusion spheres. The r value refers to the Pearson correlation coefficient.

https://doi.org/10.1038/s43588-025-00850-5



Extended Data Fig. 4 | CN1 and CN2 form a covalent bond with the catalytic cysteine of Naam. a, The LC–MS/MS spectrum of **CN1**-modified peptide $Y_{266}DVC_{269}VGATAVDALSAGYR_{283}$ from *Drosophila melanogaster* Naam. **b**, The LC–MS/MS spectrum of **CN2**-modified peptide $M_{253}KGATDIYVC_{262}$(Carbamidomethyl)GLAYDVC_{269} from *Drosophila melanogaster* Naam. The results demonstrated that **CN1** and **CN2** form a covalent bond with the catalytic residue Cys269 of *Drosophila melanogaster* Naam. **CN1**/**CN2** (120 μM) was incubated with *Drosophila melanogaster* Naam (25 μM) at 4 °C for 2 h and the mixture was digested by trypsin.

https://doi.org/10.1038/s43588-025-00850-5



Extended Data Fig. 5 | CN1 and CN2 show inhibitory activity against the pests' Naam enzymes. a, b IC_{50} curves are obtained by incubating compounds **CN1** and **CN2** with the pests' *Myzus persicae* Naam (15 nM) and *Bemisia tabaci* Naam (30 nM) enzymes, respectively, for durations ranging from 0 to 120 minutes, revealing that both compounds are time-dependent inhibitors of these Naam enzymes; data are presented as mean values ±SEM from $n = 3$ biological replicates, with error bars representing the SEM. **c**, **d** The melting curves (first-derivative of dissociation) of Myzus persicae Naam (10 μM) and B*emisia tabaci* Naam (10 μM) in presence or absence of **CN1** (50 μM) or **CN2** (50 μM), reveal that these two compounds bind to and stabilize these enzymes; data are presented as mean values ±SEM from $n = 3, 2, 3, 3$ for **c** and $n = 3, 2, 3, 1$ for **d** biological replicates, with error bars representing the SEM.

nature portfolio

Corresponding author(s): Guo-Bo Li

Last updated by author(s): May 27, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All code was written in Python (V3.9.16). For data preparation and model training, we used PyTorch (V1.12.1), PyTorch Geometric (V2.1.0), RDKit (V2022.9.5), OpenBabel (V3.1.1), and NumPy (V1.25.1). Our source code is accessible at https://github.com/ppjian19/PhoreGen.
Data analysis	For data analysis, we used Pandas (V1.5.3), SciPy (V1.11.1), Biopython (V1.81), AutoDock Vina (V1.2.2), TensorBoard (V2.14.0), Seaborn (V0.12.2), and Matplotlib (V3.7.1).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The LigPhore, CpxPhore, and DockPhore datasets for training and evaluation are avaliable at https://zenodo.org/records/15518867. The PDBBind set is available at

http://pdbbind.org.cn. The CrossDocked dataset is available at https://bits.csb.pitt.edu/files/crossdock2020/. The pharmacophore models for generation can be obtained via https://ancphore.ddtmlab.org/Modeling. The complex crystal structures (PDB codes 9KSA and 9U8M) are available at Protein Data Bank (https://www.rcsb.org).

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	NA
Population characteristics	NA
Recruitment	NA
Ethics oversight	NA

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical method was used to predetermine sample size. Following the principle of fair comparison, we referenced prior work (Guan Jiaqi, et al. 2023) and sampled 100 molecules per pharmacophore model or protein target for all models, which ensures sufficient statistical power for reliable comparisons. Reference: Guan Jiaqi, et al. 3D Equivariant Diffusion for Target-Aware Molecule Generation and Affinity Prediction. ICLR, 2023
Data exclusions	Molecules that failed sampling or did not pass RDKit's sanitization step were excluded due to the requirements of the evaluation metrics.
Replication	All models sampled and evaluated molecules under identical conditions, ensuring reproducibility. Wet-lab experiments were replicated three times and all attempts at replication were successful.
Randomization	The LigPhore, CpxPhore, and DockPhore datasets were randomly split into training, validation, and test set, ensuring no data leakage . All models sampled and evaluated molecules using a fixed random seed.
Blinding	No blinding was included in this study because of the automated and unbiased computational data analysis.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging