nature communications

Article

https://doi.org/10.1038/s41467-025-57485-3

Knowledge-guided diffusion model for 3D ligand-pharmacophore mapping

Received: 23 August 2024

Accepted: 21 February 2025

Published online: 06 March 2025

Check for updates

Jun-Lin Yu[1], Cong Zhou[1], Xiang-Li Ning[1], Jun Mou[1], Fan-Bo Meng[1], Jing-Wei Wu[1], Yi-Ting Chen[1], Biao-Dan Tang[1], Xiang-Gen Liu[2] ✉ & Guo-Bo Li[1] ✉

Pharmacophores are abstractions of essential chemical interaction patterns, holding an irreplaceable position in drug discovery. Despite the availability of many pharmacophore tools, the adoption of deep learning for pharmacophore-guided drug discovery remains relatively rare. We herein propose a knowledge-guided diffusion framework for 'on-the-fly' 3D ligand-pharmacophore mapping, named DiffPhore. It leverages ligand-pharmacophore matching knowledge to guide ligand conformation generation, meanwhile utilizing calibrated sampling to mitigate the exposure bias of the iterative conformation search process. By training on two self-established datasets of 3D ligand-pharmacophore pairs, DiffPhore achieves state-of-the-art performance in predicting ligand binding conformations, surpassing traditional pharmacophore tools and several advanced docking methods. It also manifests superior virtual screening power for lead discovery and target fishing. Using DiffPhore, we successfully identify structurally distinct inhibitors for human glutaminyl cyclases, and their binding modes are further validated through co-crystallographic analysis. We believe this work will advance the AI-enabled pharmacophore-guided drug discovery techniques.

Artificial intelligence (AI) is permeating various critical stages of drug discovery, holding immense potential to revolutionize the drug discovery process and profoundly change the current landscape of drug discovery[1–3]. In recent years, AI has made considerable advancements in lead discovery and optimization, target identification, and preclinical/clinical investigations[4,5]. Especially in structure-guided drug discovery, deep learning (DL) algorithms can efficiently accomplish several core tasks such as binding pose generation, binding affinity prediction, and molecular generation[6–12]. Currently, there are three main DL methods for binding pose generation with protein structure constraints: predicting ligand translation, rotation, and torsion to recover binding modes (e.g., DiffDock[13] and EquiBind[14]), employing gradient decent to predict protein-ligand distance matrices (e.g., TANKBind[15]), and adopting E(n)-equivariant graph neural network to iteratively update the movement and position of ligand atoms (e.g., KarmaDock[16] and E3Bind[17]). For structure-guided molecular generation, several DL algorithms have also been well-established recently, including TargetDiff[18], ResGen[19], SurfGen[20], and PockeFlow[21], some of which have undergone successful validations in discovering new hit/lead compounds. The learning capabilities of these DL methods are typically enhanced by incorporating sufficient samples or integrating knowledge such as the complementary principles of protein-ligand recognition.

Pharmacophores, as abstractions of critical chemical interactions, provide alternative means to depict the principles of protein-ligand complementary. Compared with structure-guided methods, pharmacophores possess unique, concise, and position-inclusive features, along with directional matching patterns. Consequently, they are widely employed in practical drug discovery, especially in the early stages of the process[22]. Despite the availability of several widely-used

[1]Key Laboratory of Drug Targeting and Drug Delivery System of Ministry of Education, Department of Medicinal Chemistry, West China School of Pharmacy, Sichuan University, Chengdu, Sichuan, China. [2]College of Computer Science, Sichuan University, Chengdu, Sichuan, China.
✉e-mail: liuxianggen@scu.edu.cn; liguobo@scu.edu.cn

pharmacophore-based drug discovery tools (e.g., AncPhore[23], PhDD[24], PHASE[25], Catalyst[26], and Pharao[27], Pharmit[28]), DL-enabled pharmacophore-guided drug discovery technologies remain relatively rare, with only a few instances reported to date. For example, PGMG implemented pharmacophore-guided molecular generation by establishing latent variables to address many-to-many (non-3D) mapping between pharmacophores and molecules[29]. PharmacoNet utilized DL techniques to model pharmacophores from protein structures, coupled with a graph-matching algorithm, ultimately achieving effective pharmacophore-based virtual screening[30]. The sluggish progress in DL technologies for pharmacophore-guided drug discovery involves multiple factors, including the absence of high-quality datasets and sophisticated algorithms capable of efficiently capturing sparse pharmacophore features.

By enhancing our previously developed anchor pharmacophore tool, AncPhore[23], we created two datasets (CpxPhoreSet and LigPhoreSet) of 3D ligand-pharmacophore pairs, incorporating 10 pharmacophore feature types and exclusion spheres. CpxPhoreSet, derived from experimental protein-ligand complex structures, contains real but biased ligand-pharmacophore mapping (LPM) scenarios. In contrast, LigPhoreSet, generated from energetically favorable ligand conformations by considering both pharmacophore and ligand diversity, covers a broader range of perfectly-matched ligand–pharmacophore pairs. The complementary characteristics of these two datasets enable the development of efficient DL models for LPM and other pharmacophore-based tasks (e.g., de novo design and structural optimization).

Adhering to the ligand-pharmacophore matching principles, we propose DiffPhore, a pioneering knowledge-guided diffusion framework for "on-the-fly" 3D LPM. The main concept behind DiffPhore is to utilize the LPM knowledge to guide the conformation generative process, meanwhile leveraging calibrated sampling to reduce the exposure bias inherent in diffusion models. Specifically, DiffPhore consists of three main modules: knowledge-guided LPM encoder, diffusion-based conformation generator, and calibrated conformation sampler. The LPM encoder extracts the ligand-pharmacophore matching principles based on type and directional alignment, efficiently representing the mapping relationships between 3D ligands and pharmacophores. Based on these LPM representations, the diffusion-based conformation generator processes the ligand-pharmacophore matching information and estimates the directions for conformation denoising. The calibrated sampler adjusts the conformation perturbation strategy to narrow the discrepancy between the training and inference phases, aiming to enhance sample efficiency.

In this work, we evaluate DiffPhore on two independent datasets, PDBBind test set and PoseBusters[31] set, and observe that it outperforms traditional pharmacophore tools and several advanced docking methods in predicting binding conformations. Further assessments on the DUD-E database[32] and the IFPTarget library[33] highlight DiffPhore's effectiveness in virtual screening for both lead discovery and target fishing. We then apply DiffPhore for virtual screening of human glutaminyl cyclases, promising drug targets for neurodegenerative diseases and cancer immunotherapy[34–37], successfully identifying structurally distinct inhibitors. Co-crystallographic analysis reveals consistency between the binding conformations of these inhibitors, as observed in complex crystal structures, and those predicted by DiffPhore.

Results

The complementary datasets of 3D ligand-pharmacophore pairs

To promote the development of pharmacophore-based DL methods, we released two datasets of 3D ligand-pharmacophore pairs, CpxPhoreSet and LigPhoreSet. They were constructed using AncPhore[23] by considering 10 types of pharmacophore features (as shown in Supplementary Fig. 1), including hydrogen-bond donor (HD), hydrogen-bond acceptor (HA), metal coordination (MB), aromatic ring (AR), positively-charged center (PO), negatively-charged center (NE), hydrophobic (HY), covalent bond (CV), cation-π interaction (CR), halogen bond (XB), along with steric constraints represented by exclusion spheres (EX). CpxPhoreSet comprises 15,012 ligand-pharmacophore pairs derived from experimental protein-ligand complex structures, each containing 3–15 pharmacophore features. To better represent a wider space of LPM patterns, we established a sophisticated protocol to derive ligand-pharmacophore pairs from 3D ligand structures (Fig. 1a); it involves Bemis–Murcko scaffold filtering, fingerprint similarity clustering, 3D conformation generation, pharmacophore generation and sampling, and exclusion sphere addition (see details in Methods). Starting with 11.48 million ligands from the In-Stock subset of ZINC20, we ultimately obtained 280,096 representative ligands and 840,288 corresponding ligand-pharmacophore pairs, collectively forming LigPhoreSet.

By performing t-SNE analysis on the dimensionality-reduced ECFP4 descriptors of the ligands, we observed that the ligands in LigPhoreSet exhibit a broader chemical diversity compared to those in CpxPhoreSet (Fig. 1b, Supplementary Fig. 2 and Supplementary Table 1). Meanwhile, LigPhoreSet displays greater diversity in pharmacophore features and a roughly comparable ccurrence frequency of pharmacophore feature type, in contrast to CpxPhoreSet (Fig. 1c and Supplementary Table 2). These attributes, combined with perfect-matching ligand-pharmacophore pairs, make LigPhoreSet suitable for developing DL algorithms to capture generalizable LPM patterns across a broad chemical and pharmacophoric space. By comparison, CpxPhoreSet contained imperfectly-matched ligand-pharmacophore pairs with fitness scores ranging from 0.5 to 1.0, averaging 0.967 (Fig. 1d). It can be used to refine the model for understanding the real-world biased LPMs and recognizing the induced-fit effects of ligand-target interactions. Therefore, we employed LigPhoreSet for the initial warm-up phase of model training and CpxPhoreSet for the subsequent refinement stage.

An overview of the knowledge-guided diffusion framework DiffPhore

DiffPhore is a knowledge-guided diffusion framework designed to generate 3D ligand conformations that maximally map to a given pharmacophore model (Fig. 2a). Essentially, DiffPhore incorporates pharmacophore type and direction matching rules to guide the alignment between ligand conformations and pharmacophore models (Fig. 2b). It comprises three main modules, namely, knowledge-guided LPM encoder, diffusion-based conformation generator, and calibrated conformation sampler.

The knowledge-guided LPM encoder module encodes ligand conformation and pharmacophore model as a geometric heterogenous graph $G_t = (G_{l,t}, G_p, G_{lp})$ composed of a ligand conformation graph $G_{l,t}$, a pharmacophore graph G_p, and a full-connected bipartite graph G_{lp} that represents ligand conformation-pharmacophore relations. The explicit pharmacophore-ligand mapping knowledge, including rules for pharmacophore type and direction matching, are incorporated into G_{lp}. This is achieved by integrating the pharmacophore fingerprints, orientations, and reference angles of all ligand atoms, as well as the types and directions of all pharmacophore features (Fig. 2b; see details in Methods and Supplementary Methods). The pharmacophore type matching vectors V_{lp} are obtained by aligning each ligand atom with all pharmacophore features one by one, which is expedited using pharmacophore fingerprints. Similarly, the pharmacophore direction matching vectors N_{lp} are derived by computing the discrepancy between the intrinsic orientation of each ligand atom and the direction of each directional pharmacophore feature (HA, HD, MB, etc). Leveraging these knowledge-guided encodings, the encoder captures the essence of the alignment between ligand

https://doi.org/10.1038/s41467-025-57485-3

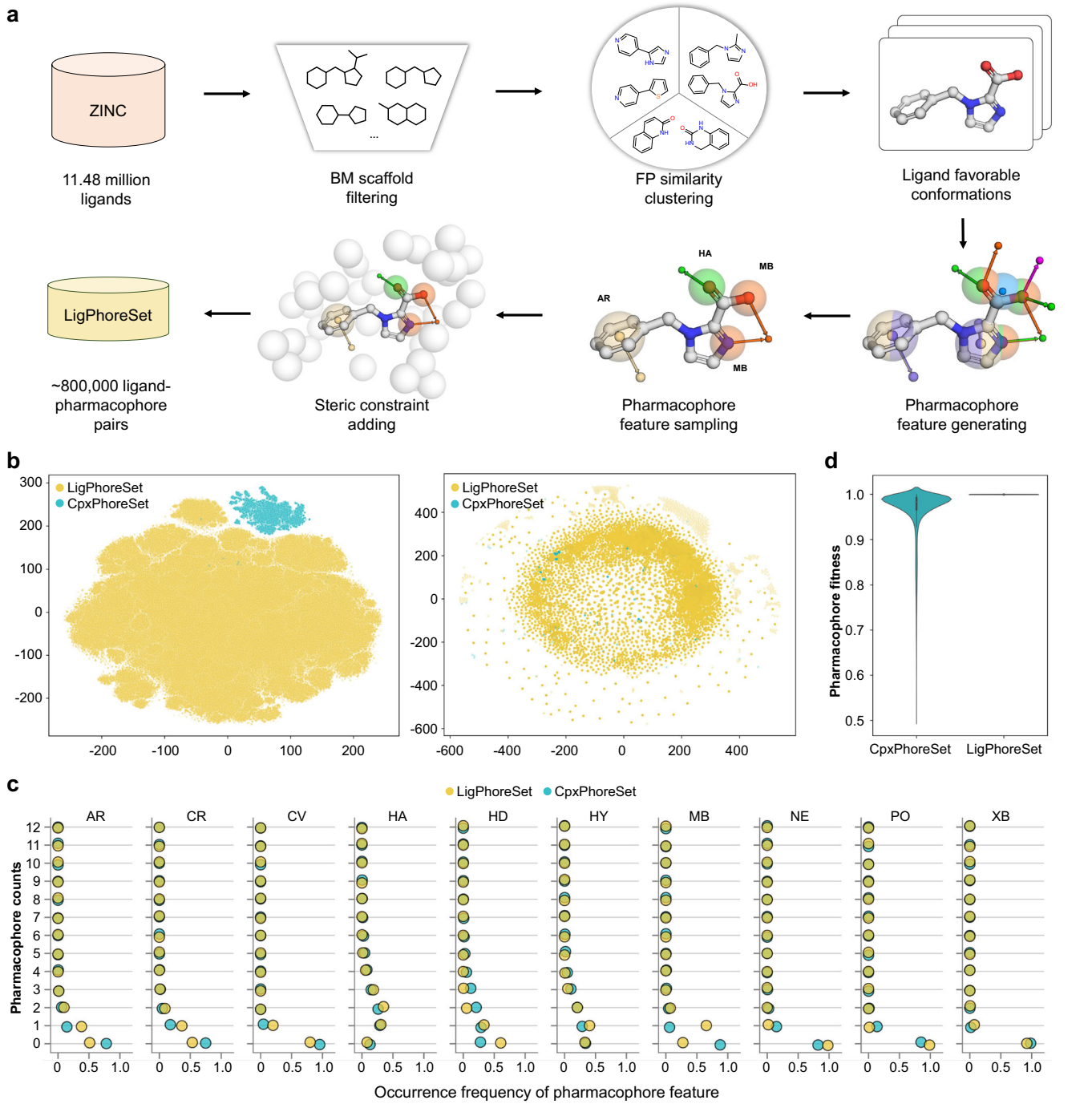

Fig. 1 | The datasets of 3D ligand-pharmacophore pairs. a The construction protocol for LigPhoreSet (see details in Methods section). **b** The t-SNE plots of ligands' ECFP4 (1024-bit) fingerprints and pharmacophore counts reveal that LigPhoreSet covers wider chemical and pharmacophoric spaces compared with CpxPhoreSet. The ECFP4 fingerprints were processed by PCA (with random_state=2024 and n_component = 50) for dimensionality reduction before the t-SNE analysis. The t-SNE analysis was performed with the following hyperparameters: n_component = 2, perplexity = 30, n_iter = 5000, random_state = 2024. **c** LigPhoreSet shares similar occurrence frequency of pharmacophore feature with CpxPhoreSet. **d** Distribution of the fitness scores (i.e., *DfScore*1; see Methods) of ligand-pharmacophore pairs from CpxPhoreSet (n = 15,012) and LigPhoreSet (n = 840,288). The boxes represent data distribution with center lines showing medians, box limits indicating the 25th and 75th percentiles, and whiskers extending to 1.5 times the interquartile range from the lower and upper quartiles. Source data are provided as a Source Data file.

conformations and pharmacophores, resulting in a robust representation of LPM.

The diffusion-based conformation generator module takes the LPM representations as input, and estimates the translation (Δr), rotation (ΔR), and torsion (Δθ) transformations for the ligand conformation at each step (see details in Methods and Supplementary Fig. 3). Crucially, the generator employs a score-based diffusion model, parameterized by an SE(3)-equivariant graph neural network, to uncover the deep geometric features of ligand conformations, pharmacophores, and most importantly, their mapping relationships. This allows for conformation exploration that is informed by both the 3D chemical structure and pharmacophore model. During training, the ground truth ligand conformation is perturbed by applying random transformations sampled from corresponding diffusion kernels at time

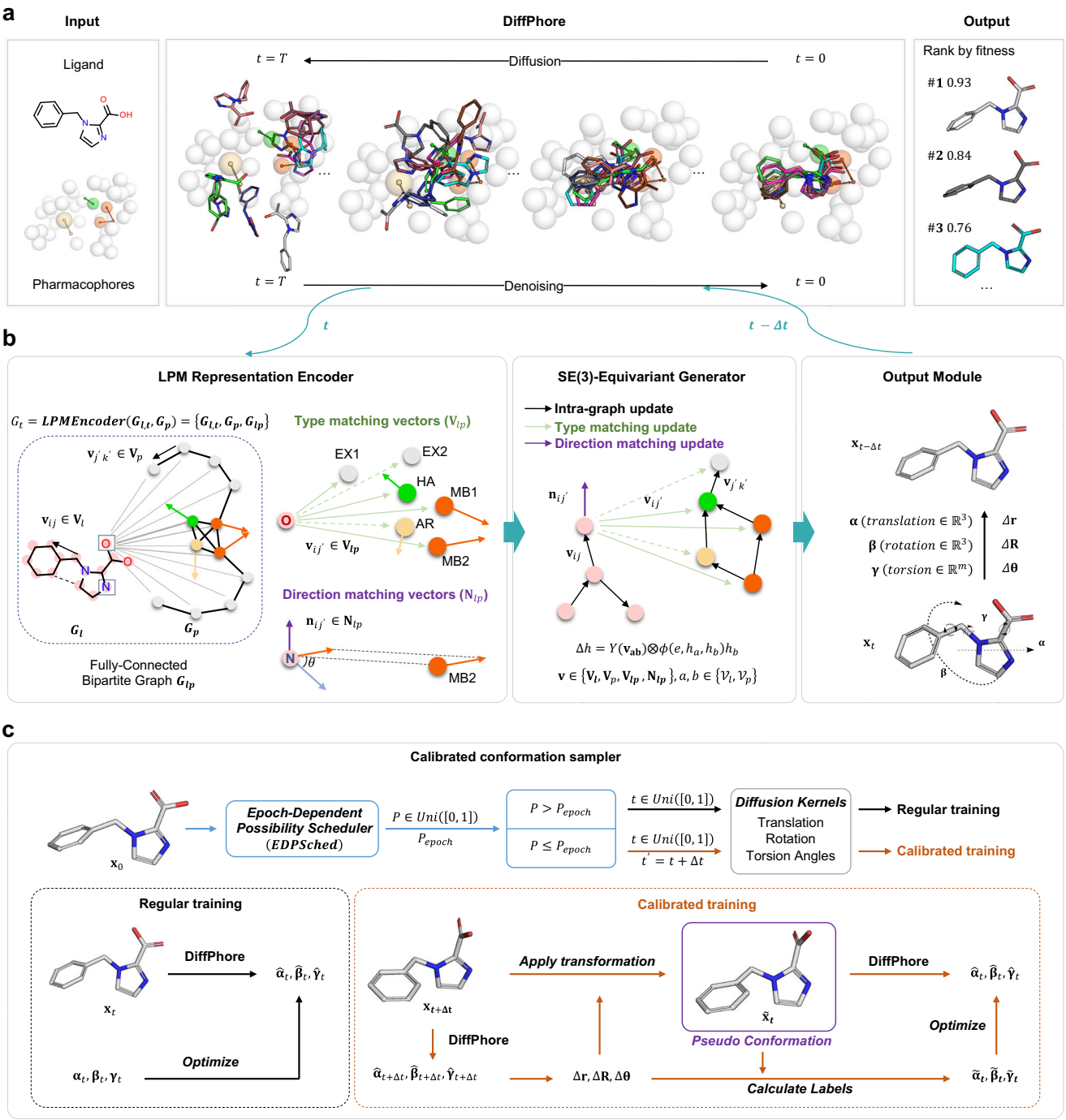

Fig. 2 | The framework of DiffPhore. a DiffPhore adopts the diffusion-denoising process to predict binding conformations mapping with pharmacophore from randomly initialized conformations. **b** DiffPhore incorporates knowledge-guided pharmacophore mapping rules for conformation generation. LPM representation encoder uses a geometric heterogenous graph G_t, including a fully-connected bipartite graph G_{lp} to represent LPM, where V_{lp} and N_{lp} are introduced to deliver type and direction matching information for ligand conformation update. **c** The calibrated conformation sampler randomly takes pseudo conformations (i.e., from intermediate prediction) as inputs for learning the conformation denoising process. The probability to select pseudo conformations is controlled by an annealing temperature P_{epoch}.

$t \in [0,1]$, which are then fed into the network as input. The conformation generator is subsequently tasked with predicting the gradients of the diffusion kernels (i.e., α, β, and γ), which can be used to estimate the actual transformations ($\Delta r, \Delta R, \Delta\theta$) to recover the ligand towards original conformation x_0 (Supplementary Fig. 3). During the conformation generation phase, the generator can gradually refine the ligand conformations until it theoretically aligns maximally with the pharmacophore model.

The auto-regressive conformation generation process faces the issue of exposure bias because the computations performed during the inference phase differ from those during the training phase of the conformation generator. During the training phase, the generator takes the perturbed conformation of the ligand as input, whereas during the inference phase, it receives the predicted conformation as input instead. Since the predicted conformations cannot ensure generation quality, any prediction error in the last step accumulates,

leading to a significant bias in the final generation, especially within large 3D conformation spaces. To address this issue, we proposed a calibrated conformation sampler to narrow the generation discrepancy between the training and inference processes (Fig. 2c; see details in Methods and Supplementary Table 3). In the training process, the calibrated conformation sampler mimics the inference computations and adopts a pseudo conformation to feed into the generator instead of the former perturbed ground truth. In this manner, the conformation generator undergoes a consistent training and inference process, reducing the exposure bias and enhancing the generation quality. In practical model training, we employed the calibrated conformation sampler for refinement training on CpxPhoreSet after an initial warm-up training on LigPhoreSet.

Our ablation experiment results indicated that removing either feature type matching or direction matching decreases the accuracy of binding conformation prediction (Supplementary Table 4), underscoring the importance of incorporating pharmacophore-specific knowledge for enhancing model performance. In terms of training schemes, skipping the warm-up training with LigPhoreSet led to reduced conformation prediction accuracy (Supplementary Table 5). Notably, excluding refinement training with CpxPhoreSet resulted in a substantial drop in prediction accuracy, demonstrating the critical role of learning from real-world ligand-pharmacophore matches. Additionally, omitting the calibrated conformation sampler impaired the model's predictive capability (Supplementary Table 5). These results highlight the importance of integrating pharmacophore mapping knowledge, conformation sampling, and the complementary use of LigPhoreSet and CpxPhoreSet datasets.

DiffPhore enables accurate prediction of ligand binding conformations

In this section, we assessed the capability of DiffPhore in generating ligand binding conformations. We selected two traditional pharmacophore programs, AncPhore and MOE, combined with two conformation generation tools, Openbabel (OB) and Conformator (CF), as baseline comparisons. MOE is recognized as one of the state-of-the-art pharmacophore tools, while AncPhore is selected for parallel comparison because it employs identical pharmacophore definitions as those used in DiffPhore. We here employed *Df Score*1 (see Methods) to rank the generated poses of DiffPhore, as it effectively reflects the quality of the generated poses (Supplementary Fig. 4). To ensure a fair comparison between DiffPhore and the baselines, we evaluated them on two independent test sets (PDBBind test set and PoseBusters set), and employed the same number of input initial conformations in each evaluation.

We observed that with 40 initial conformations, DiffPhore achieved a high top-1 success rate (73.82%) in generating the conformations with RMSD less than 2 Å on PDBBind test set, and a 67.13% success rate with only 10 initial conformations, substantially outperforming AncPhore and MOE regardless of the conformation generation methods (Fig. 3a and Table 1). The superior predictive capability and minimal impact of the number (and diversity) of initial conformations, at least partly, indicates that DiffPhore effectively reduces local optimum issues during the conformational space search. Similarly, DiffPhore exhibited a substantially higher success rate in binding pose prediction than AncPhore and MOE on PoseBusters (Fig. 3b and Table 1). Notably, DiffPhore showed comparable performance when evaluated on the subsets of new proteins from the PDBBind test set and PoseBusters set, which were excluded from the training datasets (Fig. 3c). This partly reflects that DiffPhore captures the underlying principles of ligand-pharmacophore mapping, rather than merely memorizing the training samples.

Subsequently, PoseBusters validity tests were carried out to assess the chemical and physical plausibility of the predicted binding poses, measured by the "%RMSD < 2 Å & PB-Valid" & "%RMSD < 2 Å and PB-Valid (without protein)" metrics. Compared with the baseline methods, DiffPhore achieved high success rates for plausible poses on PDBBind test set, with "%RMSD < 2 Å & PB-Valid" at 31.80% and "%RMSD < 2 Å & PB-Valid (without protein)" at 72.48%; similarly, on the PoseBusters set, it recorded 53.68% and 86.70%, respectively (Supplementary Table 6). Analysis of the energy ratios of the top-1 predicted conformations versus the ground truth ones revealed that the DiffPhore-generated conformations have relatively lower conformational energies compared to those predicted by AncPhore and MOE (Fig. 3d, e). Additionally, the bond lengths, angles, and dihedral angles of DiffPhore-generated conformations displayed distributions closely resembling those of the ground truth poses (Supplementary Figs. 5–7). These results clearly indicate that DiffPhore can produce chemically and energetically reasonable conformations while maintaining optimal mapping to the given pharmacophore models.

Next, four open-access molecule docking tools—AutoDock Vina[38], Uni-dock[39], SMINA[40], GNINA[41]—were selected for comparison; all these docking tools were evaluated under a pocket-given situation. DiffPhore outperformed the tested traditional docking tools and had comparable performance to the recently reported advanced DL-based docking tools KarmaDock[16] and SurfDock[42] on the PDBBind test set, albeit without a speed advantage (Table 1 and Supplementary Fig. 8). Notably, DiffPhore generated more intramolecularly plausible conformations than the docking tools, as evidenced by the "%RMSD < 2 Å & PB-Valid (without protein)" metric in Supplementary Table 6, while maintaining comparable effectiveness in handling protein clashes.

Given that ligand flexibility and pharmacophore complexity are known to influence the predictive accuracy and speed of pharmacophore tools, we further explored how these factors affect the performance of DiffPhore. Here, AncPhore was chosen for comparison because it uses the same pharmacophore definitions and input files as DiffPhore. It can be observed that as molecular complexity increases, the success rate of conformation prediction decreases for both DiffPhore and AncPhore (Fig. 4). However, DiffPhore shows a notable advantage over AncPhore in handling more flexible compounds. For compounds with fewer than 47 heavy atoms or 19 rotatable bonds, DiffPhore achieved a success rate of ~80% (Fig. 4a, b). Regarding pharmacophore complexity, it manifested superior predictive capability for pharmacophores with 3 to 12 features (Fig. 4c). Unlike AncPhore, whose prediction speed is significantly affected by ligand flexibility and pharmacophore complexity, DiffPhore demonstrates resilience to these factors, only experiencing a modest reduction in speed when handling more flexible ligands (Supplementary Fig. 9). DiffPhore's robust performance reflects the advancements and sophistication of our established datasets and knowledge-guided diffusion framework.

DiffPhore manifests superior screening power for lead discovery and target fishing

The robust performance of DiffPhore in generating binding conformations enables it to serve as a core engine for virtual screening in lead discovery and target fishing tasks. To examine its screening ability for lead discovery, we selected 28 structurally and mechanistically different targets, including metalloenzymes, from the DUD-E dataset[32], by considering the specificity of pharmacophore models and the impact of anchor pharmacophore features in practical drug discovery. Among the selected targets, half are exclusive to the training set, while the other half overlap with it. For all these targets, the corresponding pharmacophore models (Supplementary Table 7) were established by comparing multiple complex crystal structures; notably, they all exhibited typical anchor pharmacophore features that frequently occur across multiple structures and/or are important for natural substrate binding or catalysis. To fully leverage the advantage of DiffPhore in virtual screening, we proposed four different fitness scorings (*Dfscore*1 to *Dfscore*4; see "Methods") to evaluate the ligand-pharmacophore matching. AncPhore, MOE, and four docking tools

https://doi.org/10.1038/s41467-025-57485-3

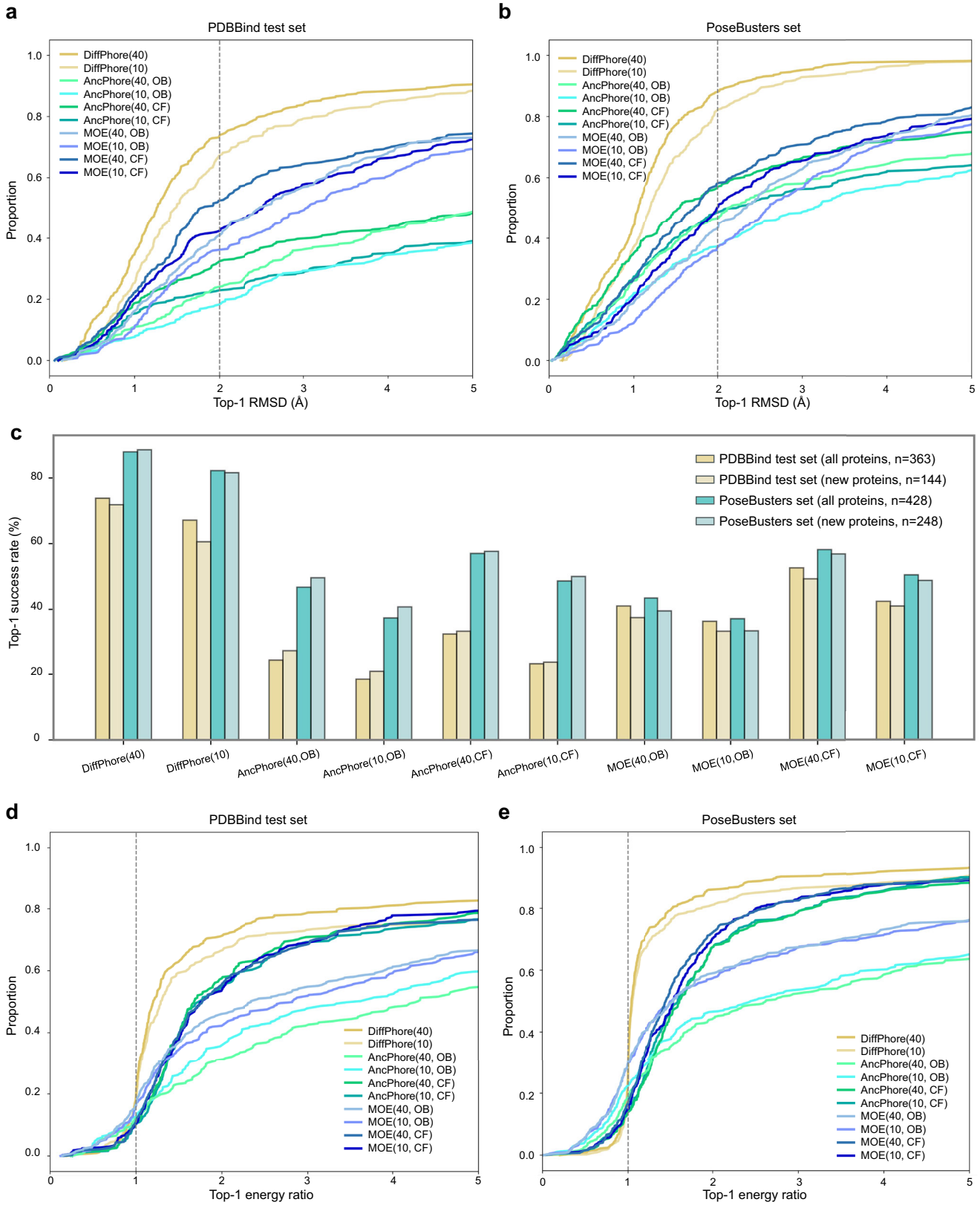

Fig. 3 | The performance of DiffPhore on ligand binding conformation prediction. Plots of cumulative distribution describing the proportion of observations falling below each RMSD value by different methods on (**a**) the PDBBind test set and (**b**) PoseBusters set. **c** The top-1 success rates for different methods evaluated on the full set (all proteins) or new protein subset (new proteins, not included in the training set) of PDBBind test set and PoseBusters set. Cumulative distribution plots describing the proportion of observations below each energy ratio value for different methods on (**d**) the PDBBind test set and (**e**) PoseBusters set. The energy ratio is calculated as $ratio = E_{pred}/E_{true}$, where E_{pred} and E_{true} represent the UFF force field energies (from PoseBusters validity test) of the predicted and ground truth poses, respectively. Source data are provided as a Source Data file.

Table 1 | Comparison of DiffPhore and other methods on the time-split PDBBind test set and the PoseBusters set

Dataset	Method[a]	Top-1 RMSD (Å)			Top-5 RMSD (Å)			Runtime (s)[b]
		%<1	%<2	Med.	%<1	%<2	Med.	
Time-split PDBBind test set	AncPhore(10, OB)	7.71	18.46	8.47	10.47	21.21	6.40	11.61
	AncPhore(40, OB)	11.02	24.24	5.39	14.04	31.13	4.11	48.23
	AncPhore(10, CF)	15.43	23.14	8.03	16.80	28.10	6.45	31.40
	AncPhore(40, CF)	18.73	32.51	5.56	22.58	38.84	4.04	53.51
	MOE(10, OB)	11.29	36.36	3.03	15.70	44.90	2.21	3.88
	MOE(40, OB)	16.25	41.05	2.47	23.42	50.96	1.93	8.06
	MOE(10, CF)	20.39	42.42	2.51	27.00	53.72	1.76	25.20
	MOE(40, CF)	22.31	52.62	1.77	32.78	63.09	1.44	28.14
	Uni-dock	17.63	34.16	4.00	26.17	47.93	2.14	3.97
	Glide SP*	17.36	44.63	2.27	31.13	60.06	1.54	-
	GNINA	20.11	42.42	2.55	25.62	55.92	1.68	84.95
	SMINA	17.63	29.48	4.19	23.14	45.18	2.29	122.63
	AutoDock Vina	17.20	30.32	4.40	21.57	45.48	2.22	67.93
	KarmaDock*	-	56.20	-	-	-	-	-
	SurfDock*	**40.96**	68.41	**1.18**	**54.18**	75.11	**0.94**	-
	DiffPhore(10)	25.35	67.13	1.48	36.77	79.11	1.21	6.97
	DiffPhore(40)	34.82	**73.82**	1.26	49.3	**80.78**	1.01	27.51
PoseBusters	AncPhore(10, OB)	22.20	37.38	3.14	24.53	46.03	2.27	21.50
	AncPhore(40, OB)	26.17	46.73	2.12	29.44	56.07	1.59	83.69
	AncPhore(10, CF)	27.57	48.60	2.16	33.41	53.97	1.67	24.55
	AncPhore(40, CF)	35.28	57.01	1.54	44.16	65.19	1.25	77.82
	MOE(10, OB)	12.38	37.15	2.60	17.76	52.57	1.92	3.21
	MOE(40, OB)	18.93	43.46	2.32	25.00	57.94	1.75	10.55
	MOE(10, CF)	20.79	50.47	1.99	29.21	63.08	1.55	7.47
	MOE(40, CF)	26.87	58.18	1.73	37.85	67.29	1.32	13.64
	Uni-dock	16.35	34.96	3.72	23.83	45.56	2.23	3.41
	GNINA	34.35	61.21	1.44	42.99	81.07	1.13	12.99
	SMINA	27.10	49.30	2.03	36.45	66.36	1.35	14.09
	AutoDock Vina	26.87	46.03	2.47	33.64	62.38	1.49	14.61
	DiffPhore(10)	37.29	82.19	1.20	51.78	92.87	0.98	4.40
	DiffPhore(40)	**44.42**	**87.89**	**1.05**	**67.22**	**96.67**	**0.76**	19.51

[a] "*" means data from references[16,42]. "-" indicates unavailable data. The numbers in parentheses for DiffPhore, AncPhore, and MOE represent the number of initial conformations. The abbreviations following the numbers in the parentheses denote the conformation tools for evaluation, where 'OB' refers to Openbabel and 'CF' refers to Conformator. The best and the second-best results are highlighted in **bold** and underlined, respectively.

[b] Although all methods were tested on the same operating system and hardware, the runtimes presented here are not directly comparable due to differences in processing devices: DiffPhore, GNINA, and Uni-dock utilized GPUs, while the other methods relied on CPUs. The time taken by conformation generation was included for AncPhore and MOE.

(AutoDock Vina, Uni-dock, SMINA, and GNINA) were chosen for comparison. To provide a more comprehensive comparison, we referenced the data for additional docking and re-scoring methods, whose screening power on these targets has been evaluated[43].

The four fitness scores of DiffPhore exhibited similar performance across different metrics (Fig. 5a–f). DiffPhore shows a superior ability to distinguish between active and decoy ligands, as evidenced by the AUROC metric, which surpasses traditional pharmacophore tools including AncPhore and MOE, and is comparable to the leading docking tools including RFscore-VS, Glide SP, and EquiScore (Fig. 5a, d). Regarding the BEDROC metric, a similar trend was observed, with the exception that MOE shows performance comparable to DiffPhore (Fig. 5b, e). Notably, all four DiffPhore fitness scores exhibited strengths in terms of enrichment factors at 0.5% and 1%, closely approaching the performance of MOE while surpassing AncPhore and nearly all docking tools (Fig. 5c, f and Supplementary Fig. 10). This highlights DiffPhore's excellent ability to prioritize active molecules according to the given pharmacophore model, which is intrinsically related to the essence of pharmacophore methods. Importantly, when comparing performance on overlapping and non-overlapping targets from the training set, DiffPhore showed no obvious differences in the metrics such as AUROC, BEDROC, and enrichment factors at 0.5% and 1% (Fig. 5a–f and Supplementary Fig. 10). This partly reflects the advantage of the pharmacophore methods in addressing the challenges of target preferences and the generalizability to unseen proteins that are often faced by DL-based docking tools. In comparison, the success rate of pharmacophore-based virtual screening is closely tied to the quality of defined pharmacophore models (i.e., whether the pharmacophore models accurately represent key target information). Theoretically, incorporating multiple pharmacophore models in virtual screening could potentially increase enrichment factors and enhance overall hit rates and chemical diversity.

As reported, the DUD-E dataset poses challenges for evaluating virtual screening methods due to the inherent biases[44,45]. Analog bias occurs when actives for a given target share similar scaffolds, creating recognizable patterns that DL models can easily exploit. Decoy bias, often introduced by selection criteria that prioritize dissimilarity to actives, can also be leveraged by models, leading to false positives. Although the DUD-E bias cannot be avoided in our evaluations, DiffPhore primarily performs the task of matching ligands to the pharmacophore

https://doi.org/10.1038/s41467-025-57485-3

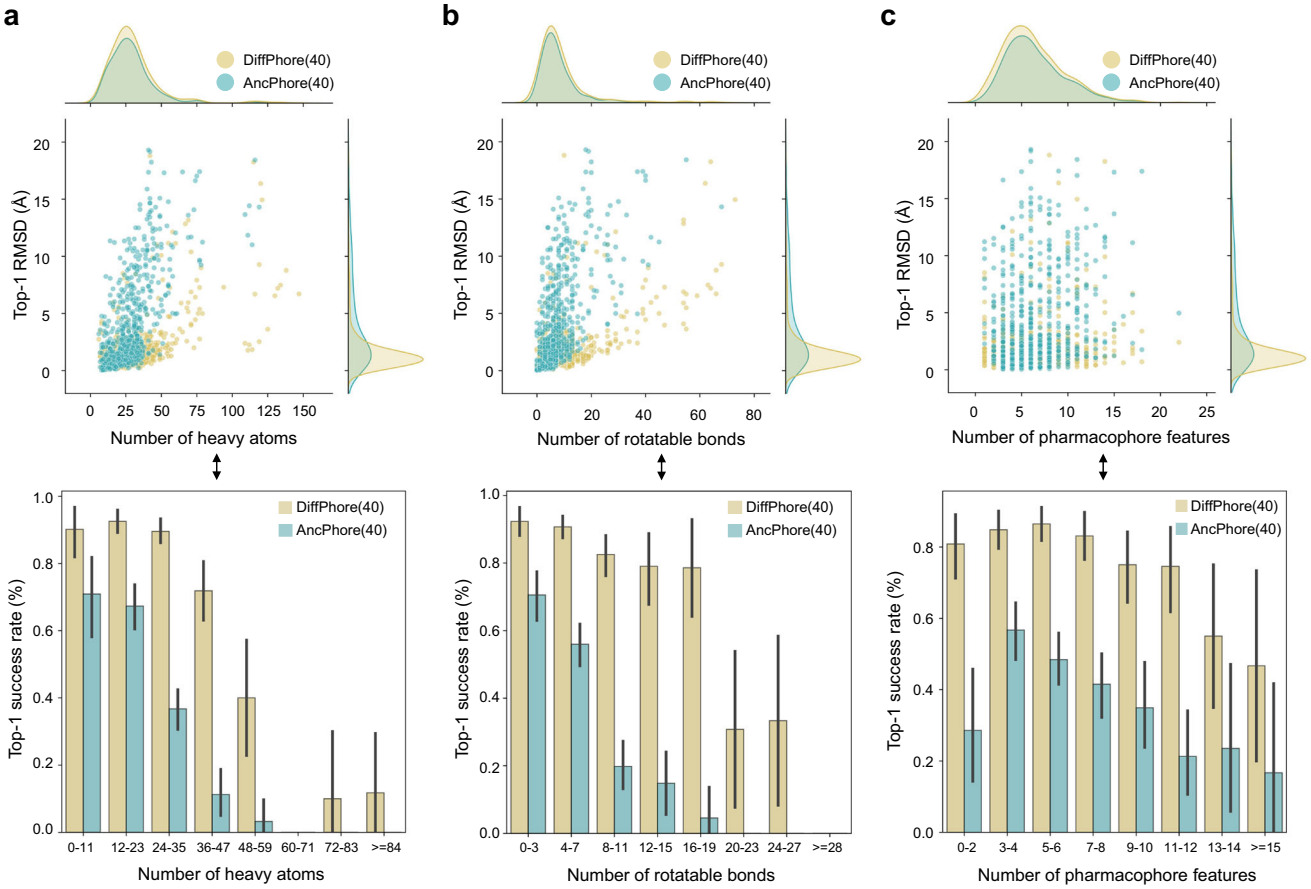

Fig. 4 | The impact of ligand flexibility and pharmacophore complexity on the predictive accuracy of DiffPhore. Plots of the Top-1 RMSD values (upper) or success rates (lower) versus the number of heavy atoms (**a**), rotatable bonds (**b**), and pharmacophore features (**c**) reveal the impacts of ligand flexibility and pharmacophore complexity on the conformation prediction performance of DiffPhore and AncPhore. The numbers in parentheses represent the number of initial conformations; top-1 success rate means generating conformations with RMSD < 2 Å. Data are presented as mean values ±95% confidence interval. Source data are provided as a Source Data file.

models, allowing for appropriate feature-matching deviations, which fundamentally distinguishes it apart from docking methods. We used four different fitness scores to assess the virtual screening performance of DiffPhore, which likely helps mitigate the impact of database bias. Overall, DiffPhore achieves comprehensive and balanced performance across all metrics, indicating its potential as a promising tool for pharmacophore-based virtual screening in lead discovery.

We next evaluated the screening ability of DiffPhore in target fishing using the IFPTarget library, which contains 2842 unique targets and 11,890 complex structures[33]. To better accommodate the target fishing tasks, we employed *DfScore*5, which is specifically designed to reduce the impact of the number of pharmacophore features on target ranking (see Methods). We selected 4OH-Tamoxifen for testing due to its known binding to more than 12 different targets. DiffPhore outperformed AncPhore in target ranking, achieving an average percent rank of 12.03%, and demonstrated comparable or slightly superior performance relative to the tested docking methods (Fig. 5g). We observed a notable limitation of DiffPhore for certain targets such as human fibroblast collagenase, mainly due to that the derived pharmacophore models cannot represent these targets. The findings indicate that appropriate pharmacophore representations are crucial for DiffPhore to achieve effective virtual screening, whether for lead discovery or target fishing.

DiffPhore identifies lead compounds for human glutaminyl cyclases

Human secretory glutaminyl cyclase (sQC) and Golgi-resident glutaminyl cyclase (gQC), responsible for N-terminal pyroglutamation for multiple protein substrates, are attractive therapeutic targets for various human diseases, including neurodegenerative diseases and cancers[34–37]. Based on the reaction pathway of sQC-catalyzed pyroglutamation of the tripeptide substrate NH_2-Gln-Phe-Ala-$CONH_2$ (QFA), we constructed a pharmacophore model derived from its initial binding mode for virtual screening using DiffPhore against about 1.4 million compounds from the Vitas-M library (see Methods). We picked 15 structurally distinct top-ranked compounds (using *DfScore*1) for experimental verification. Of them, 7 displayed inhibitory activity against sQC and gQC with IC_{50} values less than 100 μM (Supplementary Table 8 and Supplementary Fig. 11). Compounds **5** and **13** manifested IC_{50} of 6.94 μM (K_i = 6.71 μM) and 3.44 μM (K_i = 3.33 μM) to sQC, and 15.73 μM (K_i = 15.31 μM) and 3.93 μM (K_i = 3.82 μM) to gQC, respectively. Notably, both compounds exhibited the ability to thermodynamically stabilize both sQC (ΔT_m of 4.98 °C for **5** and 7.05 °C for **13**) and gQC (ΔT_m of 2.06 °C for **5** and 5.50 °C for **13**) proteins (Fig. 6a).

Through co-crystallization studies, we obtained the crystal structures of sQC in complex with **5** and **13** (Supplementary Tables 9, 10). Clear mF_o - DF_c electron density was observed in the active site for both structures, enabling confident modeling of **5** and **13** (Fig. 6b, c, Supplementary Fig. 12). Compound **5** is positioned to coordinate with the active site Zn^{2+} (N3-Zn^{2+} distance of 2.2 Å), make a hydrogen bond with the catalytic triad Asp_{248} (N1-OD2 distance of 2.8 Å), two hydrogen bonds with Gln_{304} (N1′-O2 distance of 2.9 Å, and O3′-NH distance of 3.6 Å), and face-to-face π-π stacking interactions with Trp_{207} (Fig. 6b). Compound **13** makes a coordination bond with Zn^{2+} (N3-Zn^{2+} distance of 2.0 Å), a hydrogen bond with Asp_{248} (N1-OD2 distance of 2.7 Å), and

https://doi.org/10.1038/s41467-025-57485-3

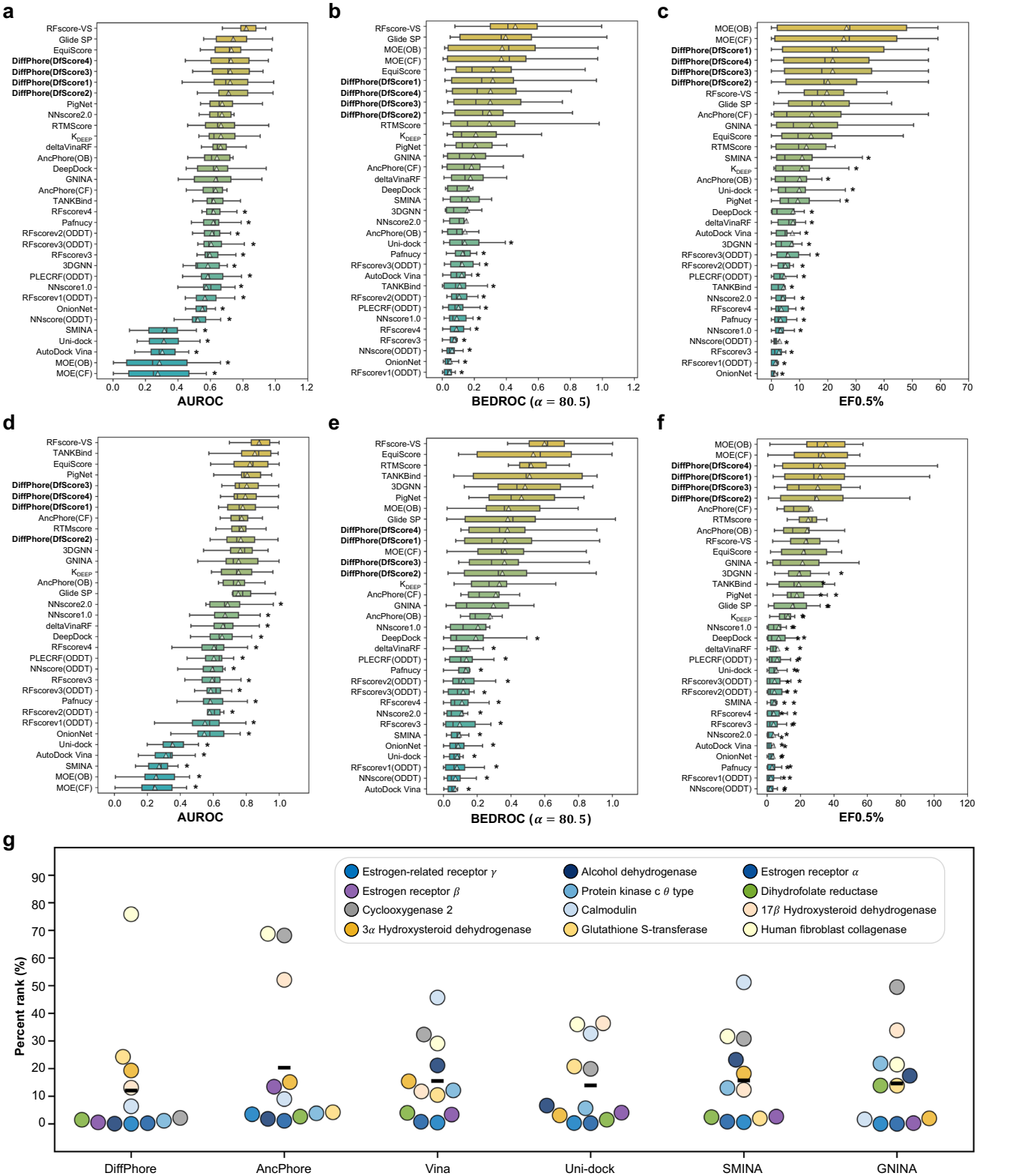

Fig. 5 | The DiffPhore screening power for lead discovery and target fishing. Comparison of different DiffPhore scorings with other methods in virtual screening for lead discovery, evaluated respectively on (**a**–**c**) non-overlapping and (**d**–**f**) overlapping targets using the metrics AUROC, BEDROC, and EF0.5%. Boxes are ranked based on their mean values, indicated by triangle markers. The "*" symbol denotes a statistically significant difference (unpaired two-sided student's t-tests, *p*-value < 0.05, $n = 14$) between the baseline and DiffPhore (*Df Score*1). Exact *p* values are provided in the Source Data file. The boxes represent data distribution with center lines showing medians, box limits indicating the 25th and 75th percentiles, and whiskers extending to 1.5 times the interquartile range from the lower and upper quartiles. AUROC area under the receiver operating characteristic curve, BEDROC Boltzmann-enhanced discrimination of receiver operating characteristic, EF0.5% enrichment factor at 0.5%. **g** Comparison of DiffPhore with other baselines in predicting the 12 targets of 4OH-tamoxifen. Percent rank = (rank order/number of total complex structures in IFPTarget) × 100. Source data are provided as a Source Data file.

https://doi.org/10.1038/s41467-025-57485-3

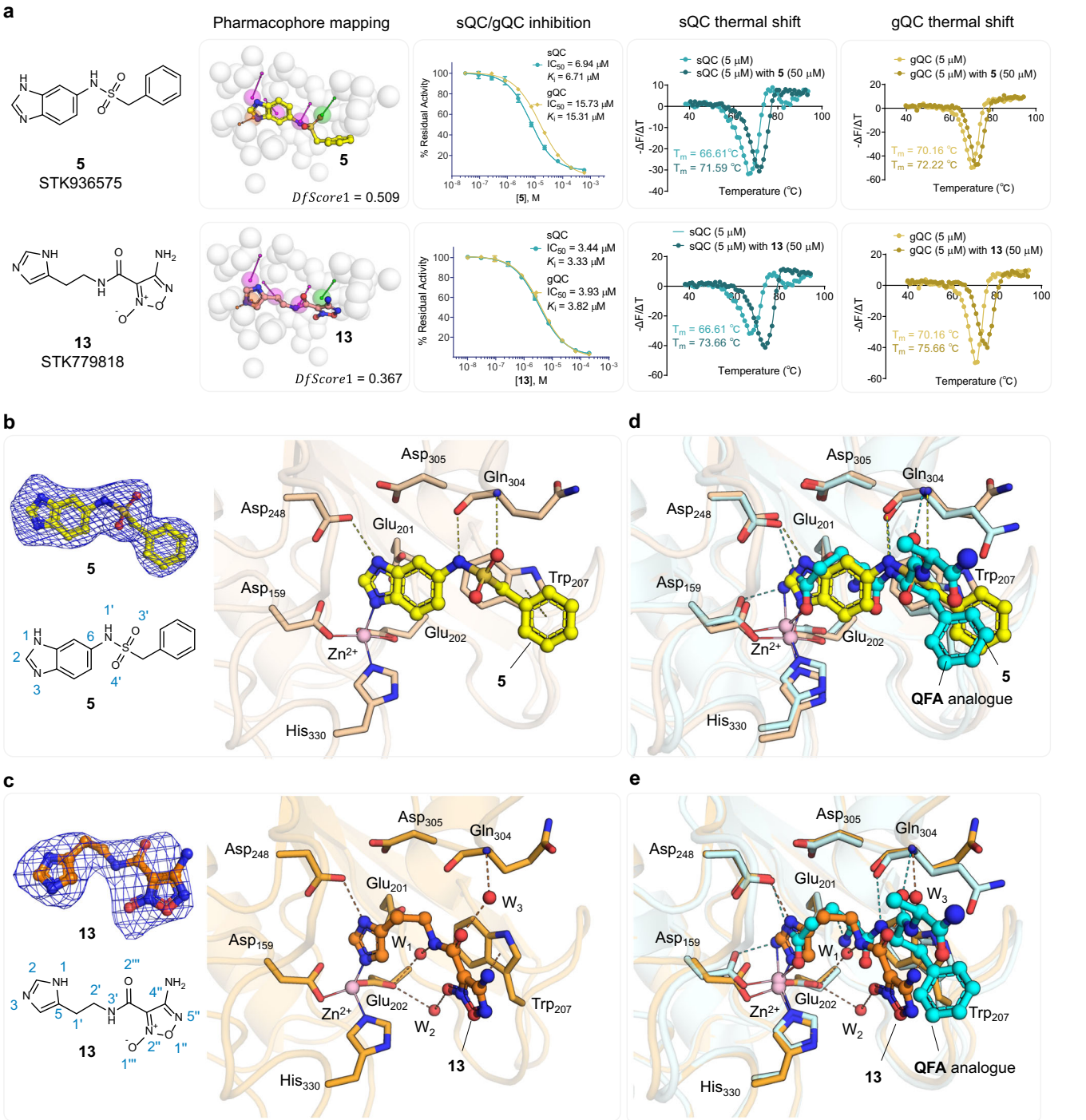

Fig. 6 | The sQC/gQC inhibitors identified by DiffPhore. a The chemical structures of compounds **5** and **13**, along with their predicted conformations mapping to the pharmacophore model derived from the binding mode of QFA with sQC; the IC_{50} curves of the two inhibitors of sQC/gQC (all determinations are tested in triplicate; data are presented as mean values ±SEM); the melting curves (first-derivative of dissociation) of sQC (yellow) and gQC (cyan) in the presence or absence of **5** (50 μM) or **13** (50 μM). Views from the (**b**) sQC:**5** (PDB code 9ISD) and (**c**) sQC:**13** (PDB code 9IVV) complex structures, revealing the modes of **5** and **13** inhibiting sQC; the mF_o-DF_c electron density (OMIT maps, blue mesh, contoured at 3.0σ) around **5** and **13** are calculated from the last refinement models. Superimpositions of (**d**) sQC:**5** and (**e**) sQC:**13**, respectively, with sQC:QFA analog (PDB code 6YI1)[46], reveal that **5** and **13** have a similar mode as that of QFA analog with sQC.

water-bridging interaction with Glu_{202} and Gln_{304} (Fig. 6c). Superimposing these structures with the sQC:QFA analog structure (PDB code 6YI1)[46] reveals that both compounds, especially **5**, closely resemble QFA analog binding with sQC (and gQC) through similar pharmacophore features, notably in zinc coordination and hydrogen-bonding interactions with catalytically important residues (Fig. 6d, e, Supplementary Fig. 13, and Fig. 14). These results highlight the effectiveness of DiffPhore in pharmacophore-guided lead discovery.

Discussion

Pharmacophore and molecular docking methods are both fundamentally based on the principles of receptor–ligand recognition, but they are implemented in entirely different ways. Molecular docking captures a wide array of receptor-ligand interactions, considering the full spectrum of possible contacts. In contrast, pharmacophores distill these interactions down to their most essential features, focusing on the abstracted representation and precise matching of key

interactions, including directional alignment. The pharmacophore approach sidesteps the complexity of less significant interactions, providing a streamlined and efficient mode for drug discovery. More importantly, pharmacophore can avoid the target preference issue often encountered in traditional or DL-based molecular docking methods. Therefore, developing DL-enabled pharmacophore technologies is a promising direction. DiffPhore stands as the pioneering DL model for the LPM task, potentially acting as a catalyst for advancing this kind of technologies.

Since there are no standardized datasets for constructing DL-based pharmacophore models, we established LigPhoreSet and CpxPhoreset by considering 10 types of pharmacophore features and exclusion spheres. The combined use of these two datasets has been demonstrated to develop high-quality models for the LPM task. Certainly, these datasets will be useful for developing additional DL models for other tasks, such as molecular generation. Moreover, our proposed protocol for constructing ligand-derived datasets may inspire the creation of more robust and comprehensive datasets, driving progress in pharmacophore-guided drug discovery.

In DiffPhore, the pharmacophore principles have been elaborately fused into neural networks to search for ligand binding conformations according to a given pharmacophore model. This approach greatly enhances the efficiency of conformation search and reduces the likelihood of combinatorial explosion, which is especially beneficial for highly flexible ligands, as observed in the test results. The knowledge-guided diffusion-based framework, which focuses on 3D transformations rather than the generation of absolute atomic coordinates, can produce more chemically plausible and energetically favorable conformations. The proposed calibrated conformation sampler proved effective in resolving the discrepancies between training and inference stages and mitigating the model exposure issues in diffusion models. Notably, DiffPhore shows robust performance in generating binding conformations even for unseen proteins, reflecting, at least in part, that it has learned the essential principles of ligand-pharmacophore matching rather than merely memorizing the training data.

In practical application, we utilized DiffPhore to implement a substrate-mimicking strategy (i.e., deriving pharmacophore models from catalytic reaction pathways), and successfully discovered structurally distinct inhibitors for the clinically important metalloenzymes sQC/gQC. Co-crystallographic analysis revealed that the binding modes of the inhibitors closely resemble that of the sQC/gQC substrate, especially concerning anchor pharmacophore features such as zinc coordination and hydrogen bonding with the catalytically important residues. This case study clearly reveals that DiffPhore, equipped with precisely defined pharmacophore models, can efficiently discover high-quality lead compounds. It also highlights the unique advantage of pharmacophore models in identifying metalloenzyme inhibitors involving metal coordination.

The encouraging performance of DiffPhore in predicting binding conformations and conducting virtual screening highlights the advancements of our proposed datasets and knowledge-guided diffusion-based framework. Further development of algorithms is warranted to improve computational efficiency and accuracy, to address differences in bond lengths and dihedral angles arising from conformational changes, to consider intramolecular interactions (e.g., intramolecular hydrogen bonds), as well as to tackle conformational prediction challenges posed by more complex ligands (e.g., macrocyclic structures).

Methods

Dataset construction

We constructed two 3D ligand-pharmacophore pair datasets, CpxPhoreSet and LigPhoreSet, for LPM learning, by using the enhanced version of AncPhore[23]. CpxPhoreSet was established by analyzing a total of 19,443 protein-ligand complex structures collected in PDBBind (version 2020)[47,48]. We followed a time-split scheme[14] and divided PDBBind into train (16,379 entries), validation (968 entries), and test (363 entries) sets. The train and validation set were used to establish the CpxPhoreSet and the remaining test set was used for performance evaluation. For each complex structure, AncPhore was used to generate one pharmacophore model considering 10 pharmacophore feature types (HD, HA, MB, AR, PO, NE, HY, CV, CR, and XB) and exclusion spheres (EX) according to protein-ligand interactions. These models with less than 3 features and more than 15 features were disregarded. The retained pharmacophore models, along with ligand conformations, constituted the CpxPhoreSet, encompassing a total of 15,012 ligand–pharmacophore pairs.

We started LigPhoreSet construction with ~11.48 million ligands (with molecular weights less than 800 and LogP less than 5) obtained from the In-Stock subset of the ZINC20 database[49] (downloaded in May, 2023). After removing duplicates, multiple components, and unidentifiable SMILESs using the RDKit software, the remaining ligands were clustered based on Bemis–Murcko scaffold[50] rules. From each ligand cluster, a representative ligand was randomly chosen. These selected ligands underwent further filtration based on Morgan fingerprint[51] similarity (with a radius of 2) to ensure a broad but non-redundant diversity of ligand chemotypes. Then, each of these filtered ligands generated a corresponding, energetically favorable 3D conformation using the RDKit MMFF force field. Next, the pharmacophore models for all ligand conformations are generated as follows: (1) generating initial pharmacophore models by considering 10 pharmacophore feature types (identical to those used for CpxPhoreSet) for all 3D ligand conformations; (2) retaining the pharmacophore models bearing at least 2 features among MB, HA, HD, AR, NE, or PO; (3) sampling three pharmacophore models with different pharmacophore feature combinations for each ligand conformation; (4) generating exclusion spheres as steric constraints via a pseudo-receptor manner, with exclusion spheres introduced around the ligand conformation at distances ranging from 3 Å to 5 Å. We finally obtained a version of LigPhoreSet comprising 840,288 ligand-pharmacophore pairs derived from 280,096 ligands. To facilitate initial model training and hyperparameter search, we randomly extracted a subset from LigPhoreSet, containing 84,030 samples and 28,010 ligands.

Problem formulation

Generally, the 3D ligand-pharmacophore mapping (LPM) task is to identify a reasonable ligand conformation that maximally matches with a given pharmacophore model. Given the 3D structure of the pharmacophore G_p and an initial ligand conformation G_l, the LPM model predicts a ligand conformation that satisfies the pharmacophore constraints.

$$\widehat{G}_l = Model\left(G_p, G_l\right) \tag{1}$$

where G_p, G_l and $\widehat{G}_l$ represent the 3D representations of the input pharmacophore model, the input ligand conformation and the generated ligand conformation, respectively. The generated ligand conformation $\widehat{G}_l$ shares the same chemical structure with G_l and only differs in conformation.

The LPM problem treated by diffusion-based generative modeling

To solve the LPM problem, we need to approximate the conditional probability density of ligand binding conformation $P(\widehat{G}_l|G_p, G_l)$. In general, the gradient of the probability density $\nabla P(\widehat{G}_l|G_p, G_l)$ is called the score function. Song and Ermon introduced score-based generative modeling to learn this score function from data and to generate samples with Langevin dynamics[52]. Given an initial sample $\widehat{G}_{l,0}$ from any prior distribution $\pi(\widehat{G}_{l,0})$ (e.g., Gaussian distribution), Langevin dynamics is incorporated to perform denoising using the following

iterative update:

$$\widehat{G}_{l,t} = \widehat{G}_{l,t-1} + \epsilon \nabla \log P\left(\widehat{G}_{l,t-1} | G_p, G_l\right) + \sqrt{2\epsilon} z_t, 1 \le t \le T \quad (2)$$

where ϵ is the step size and T is the number of iterations. z_t is the sample noise from $\mathcal{N}(0,1)$. $\widehat{G}_{l,0} = G_l$ is the input ligand conformation and $\widehat{G}_{l,T} = \widehat{G}_l$ is the predicted ligand conformation. Under certain modest conditions[53,54], and with sufficiently small step size ϵ and large T, the distribution of $\widehat{G}_{l,T}$ will approximate the true distribution of ligand binding conformation $P_{data}(G_l^*|G_p, G_l)$, where G_l^* represents an optimal ligand conformation.

In this paper, we proposed a score-based diffusion model DiffPhore, to iteratively generate the ligand conformation $\widehat{G}_l$. Specifically, the denoising probability can be formulated as $P_\sigma(G_{l,t}|G_l^*, G_p) = \mathcal{N}(G_{l,t}|G_l^*, \sigma^2 I)$ given the ligand structure G_l^*. $G_{l,t}$ stands for the perturbed data point at step t. Thus, the perturbed data distribution $P_\sigma(G_{l,t}|G_p, G_l) = \int P_{data}(G_l^*|G_p, G_l) P_\sigma(G_{l,t}|G_l^*) dG_l^*$. We consider a sequence of decreasing noise scales $\{\sigma_t\}_{t=1}^{T}$ ($\sigma_1 > \sigma_2 > \ldots > \sigma_T$). DiffPhore introduces a score network[52,55] $CFGenerator(G_{l,t}, G_p, t)$ to estimate the score function at each noise level σ_t. Thus, the training loss of this score network is

$$L = \frac{1}{T}\sum_{t=1}^{T} \sigma_t^2 \mathbb{E}_{P_{data}} \mathbb{E}_{P_{\sigma_t}(G_{l,t}|G_l^*)} \left[\left|\left| CFGenerator\left(G_{l,t}, G_p, \sigma_t\right) - \nabla \log P_{\sigma_t}\left(G_{l,t}|G_l^*\right) \right|\right|^2 \right] \quad (3)$$

As for the generation stage, DiffPhore sequentially performs T steps of Langevin MCMC to obtain a reasonable ligand conformation mapped with the given pharmacophore.

$$\widehat{G}_{l,t} = \widehat{G}_{l,t-1} + \epsilon_{t-1} CFGenerator(\widehat{G}_{l,t-1}, G_p, \sigma_{t-1}) + \sqrt{2\epsilon_{t-1}} z_t, 1 \le t \le T \quad (4)$$

Where ϵ_{t-1} is the step size at $t-1$. In the next section, we will explain the detailed implementations of the above generative modeling framework.

DiffPhore architecture

To iteratively generate reasonable ligand conformations given the pharmacophore, DiffPhore adopts the following three main components, including a knowledge-guided LPM representation encoder (*LPMEncoder*), a conformation generator (*CFGenerator*), and a calibrated conformation sampler (*CCSampler*). As the input of the score network, the LPM representation leverages pharmacophore principles to characterize 3D mapping relationships of the ligand conformation-pharmacophore pairs. The conformation generator takes LPM representations as inputs to iteratively search ligand conformations to fit with a pharmacophore model, continuing until the maximum alignment is achieved. The calibrated conformation sampler is designed to eliminate the exposure bias of the iterative conformation search process.

Knowledge-guided LPM representations

Accurately representing LPMs is the prerequisite for successfully performing the task of fitting ligand conformations to pharmacophores. We proposed a heterogenous geometric graph G_t to characterize LPMs in the 3D space,

$$G_t = LPMEncoder\left(G_{l,t}, G_p\right) = \left\{G_{l,t}, G_p, G_{lp}\right\} \quad (5)$$

where $G_{l,t} = \{\mathcal{V}_l, \mathrm{x_t}, \mathcal{E}_l, \mathrm{V}_l\}$ is a ligand graph at the t-th step, $G_p = \left\{\mathcal{V}_p, \mathrm{x_p}, \mathcal{E}_p, \mathrm{V}_p\right\}$ is a pharmacophore graph, and G_{lp} is a bipartite graph (Fig. 2b; see details in Supplementary Methods). $\mathcal{V}_l$ and $\mathrm{x_t}$ stand for the ligand atoms and their 3D coordinates. $\mathcal{E}_l$ represents the covalent bonds as well as unbonded edges within 5 Å in the ligand. $\mathcal{V}_p$ and $\mathrm{x_p}$ represent the pharmacophore points and their 3D coordinates, respectively. $\mathcal{E}_p$ denotes the connections between each pair of pharmacophore features, and connections of each exclusion sphere to the nearest pharmacophore point in $\mathcal{V}_p$. $\mathrm{V}_l = \{\mathrm{v}_{ij} | i, j \in \mathcal{V}_l\}$ and $\mathrm{V}_p = \{\mathrm{v}_{j'k'} | j', k' \in \mathcal{V}_p\}$ are vectors connecting the neighboring nodes in G_l and G_p, respectively.

The bipartite graph $G_{lp} = \{\mathcal{V}_l, \mathcal{V}_p, \mathcal{E}_{lp}, \mathrm{V}_{lp}, \mathrm{N}_{lp}\}$ is exploited to describe the ligand-pharmacophore matching relations, where $\mathcal{E}_{lp}$ connects each ligand atom to all the pharmacophore feature points, $\mathrm{V}_{lp} = \{\mathrm{v}_{ij'} | i \in \mathcal{V}_l, j' \in \mathcal{V}_p\}$ and $\mathrm{N}_{lp} = \left\{\mathrm{n}_{ij'} \middle| i \in \mathcal{V}_l, j' \in \mathcal{V}_p\right\}$ are the pharmacophore type and direction matching vectors, respectively. We detail the featurization and implementation of $G_{l,t}$, G_p and G_{lp} in Supplementary Methods.

Diffusion-based conformation generator

Given the LPM representations G_t at random time t, the generator aims to predict the conformations $\hat{\mathrm{x}}_{t-\Delta t}$ at the former step, given by

$$\hat{\mathrm{x}}_{t-\Delta t} = CFGenerator\left(G_t, t\right) \quad (6)$$

Since the degrees of freedom in 3D coordinates are significantly higher than needed, the ligand conformation here is represented by a combination of translations, rotations, and changes to torsion angles. The ligand conformation space can be formulated as an $(m+6)$-dimensional submanifold, m is the number of rotatable bonds, and 6 refers to the roto-translations[13]. More specifically, the ligand conformation space can be formally defined as:

$$g = (\mathrm{r}, \mathrm{R}, \theta) \in \mathbb{P} \quad (7)$$

$$\mathbb{P} = \mathbb{T}^3 \times SO(3) \times SO(2)^m \quad (8)$$

$$\mathrm{x}_{t-\Delta t} = A\left((\mathrm{r}, \mathrm{R}, \theta), \mathrm{x_t}\right) = A_{tr}\left(\mathrm{r}, A_{rot}\left(\mathrm{R}, A_{tor}\left(\theta, \mathrm{x_t}\right)\right)\right) \quad (9)$$

where $\mathbb{P}$ is the product space of 3D translation group $\mathbb{T}^3$, 3D rotation group $SO(3)$ and changes in torsion angles $SO(2)^m$, A stands for the total transformation with respect to an element $g = (\mathrm{r}, \mathrm{R}, \theta)$ in the product space $\mathbb{P}$, and A_{tr}, A_{rot}, A_{tor} refer to the actual translation, rotation, and torsion angle transformations. To reduce the prediction complexity, we formulate the prediction task as the learning of the change directions (or scores) in the ligand translation (α), rotation (β), and torsion angles (γ) instead:

$$\alpha, \beta, \gamma = CFGenerator\left(G_t, t\right) \quad (10)$$

$$\alpha_t, \beta_t, \gamma_t \leftarrow \nabla p_t^{tr}\left(\Delta \mathrm{r}_{0\to t} | 0\right), \nabla p_t^{rot}\left(\Delta \mathrm{R}_{0\to t} | 0\right), \nabla p_t^{tor}\left(\Delta \theta_{0\to t} | 0\right) \quad (11)$$

The scores (α, β, γ) correspond to the gradient estimates of the translation (p_t^{tr}), rotation (p_t^{rot}) and torsion angle (p_t^{tor}) diffusion kernels, which follow Gaussian distribution, IGSO(3) distribution[56] and the wrapped normal distribution[57], respectively. The corresponding gradients (e.g., $\nabla p_t^{tr}\left(\Delta \mathrm{r}_{0\to t} | 0\right)$ can be easily calculated in advance[13].

The SE(3)-equivariant conformation generator comprises the embedding, update, and output modules (Supplementary Fig. 3a). The update module consists of L message-passing layers, each with intra- and inter-graph update layers. The intra-graph layer extracts the topological features of the ligand and the pharmacophore separately; the inter-graph layer performs the feature fusion between two graphs, establishing deep representations of the ligand-pharmacophore interactions. Finally, the output module predicts

the change directions in the ligand translation (α), rotation (β), and torsion angles (γ).

Initially, the embedding module processes input ligand and pharmacophore graphs as well as cross-edges between them, and integrates the random diffusion time (t) into the graph features (Supplementary Fig. 3b; see details in Supplementary Methods). This yields the initial embeddings for ligand (h_l^0, e_l) pharmacophore (h_p^0, e_p) and cross edges (e_{lp}). Since all the computations of the conformation generator correspond to a specific step, we omit the notation t for the subsequent features for simplicity.

$$h_l^0, e_l, h_p^0, e_p, e_{lp} = Embedding(G_t, t) \tag{12}$$

Next, the update module iteratively refines the initial embeddings via message passing layers (Supplementary Fig. 3c), with each layer comprising intra-graph and inter-graph updates. Intra-graph updates compute messages ($m_{l,i-intra}$ and $m_{p,j'-intra}$, Eqs. 13–15) to incorporate the information of internal topological structure within each graph. The updates are computed as tensor products of the node features and the spherical harmonic representations of neighboring edge vectors, weighted by the edge embedding e_{ij}, the outgoing h_j and the incoming node features h_i.

$$m_{l,i-intra} = TP_{l\to l}\left(h_{l,i}^{l-1}, \mathrm{V}_l, e_l\right) \tag{13}$$

$$m_{p,j'-intra} = TP_{p\to p}\left(h_{p,j'}^{l-1}, \mathrm{V}_p, e_p\right) \tag{14}$$

$$TP(h_i, \mathrm{V}, e) = BN\left(\frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i, \mathrm{v}_{ij}\in\mathrm{V}} Y(\mathrm{v}_{ij}) \otimes \phi(e_{ij}, h_i, h_j)h_j\right) \tag{15}$$

In the formulas, TP denotes the tensor products layer, $\mathcal{N}_i = \{j|(i,j) \in \mathcal{E}_l \text{ or } \mathcal{E}_p\}$ stands for neighbor nodes and Y is the spherical harmonic. BN refers to the batch normalization, ϕ stands for a MLP layer, and $\otimes$ refers to tensor product operation.

The inter-graph layer simulates the ligand-pharmacophore recognition and alignment with the constructed bipartite graph G_{lp}. The pharmacophore type and direction matching vectors are incorporated for the calculation of the inter-graph updates ($m_{l,i-inter}$, $m_{p,j'-inter}$) via similar tensor product layers (Eqs. 16–17). The intra- and inter-graph messages are aggregated to update ligand (h_l^l) and pharmacophore (h_p^l) node embeddings (Eqs. 18–19). All tensor product layers here are implemented using the 'FullyConnectedTensorProduct' layer from the E3NN package (see details in Supplementary Methods).

$$m_{l,i-inter} = TP_{p\to l,type}\left(h_{l,i}^{l-1}, \mathrm{V}_{lp}, e_{lp}\right) + TP_{p\to l,direction}\left(h_{l,i}^{l-1}, \mathrm{N}_{lp}, e_{lp}\right) \tag{16}$$

$$m_{p,j'-inter} = TP_{l\to p,type}\left(h_{p,j'}^{l-1}, \mathrm{V}_{lp}, e_{lp}\right) + TP_{l\to p,direction}\left(h_{p,j'}^{l-1}, \mathrm{N}_{lp}, e_{lp}\right) \tag{17}$$

$$h_{l,i}^l = h_{l,i}^{l-1} + m_{l,i-intra} + m_{l,i-inter} \tag{18}$$

$$h_{p,j'}^l = h_{p,j'}^{l-1} + m_{p,j'-intra} + m_{p,j'-inter} \tag{19}$$

Finally, the output module receives the updated ligand features h_l^L to predict the translation, rotation, and torsion scores *w.r.t* their diffusion kernels. The translation and rotation are rigid transformations operating on the center of mass of the ligand so that edges between ligand atoms and the center of mass $\mathcal{E}_{ca}$ are constructed to compute the corresponding scores α, β with tensor product layer. Here, ψ, ϕ are MLP layers, and $\mu(\cdot)$ denotes radical bias embedding of edge length. The vector v_{ca} stands for the 3D vectors from the center of mass to the ligand atom a.

$$\alpha, \beta = \frac{1}{|\mathcal{V}_l|}\sum_{a\in\mathcal{V}_l} Y(\mathrm{v}_{ca}) \otimes \phi\left(e_{ca}, h_{l,a}^L\right)h_{l,a}^L \tag{20}$$

$$e_{ca} = \psi(\mu(||\mathrm{v}_{ca}||)) \tag{21}$$

To estimate the torsion score γ, the output layer focuses on the rotatable bonds and the adjacent atoms to predict the corresponding changes in torsion angles. For the torsion score γ_r of the rotatable r, the output layer performs the tensor product operation between the bond features and the adjacent atom features, given by

$$\gamma_r = \frac{1}{|\mathcal{N}_r|}\sum_{b\in\mathcal{N}_r} T_r(\mathrm{v}_{rb}) \otimes \phi\left(e_{rb}, h_r, h_b^L\right)h_b^L \tag{22}$$

$$e_{rb} = \omega(\mu(||\mathrm{v}_{rb}||)) \tag{23}$$

$$h_r = CONCAT\left(\left[h_{r_{in}}, h_{r_{out}}\right]\right) \tag{24}$$

$$T_r := Y^2(\mathrm{v}_r) \otimes Y(\mathrm{v}_{rb}) \tag{25}$$

where $\mathcal{N}_r$ is the set of ligand atoms connected with the rotatable bonds. T_r is a convolutional filter constructed for each rotatable bond r calculating the tensor product of the spherical harmonic representation of the bond axis v_r (Y^2 here means max level is 2) and the vector v_{rb} from bond center of r to the atom b. ω refers to a MLP layer, h_r stands for the feature of the rotatable bond r formed by the involved node features of the incoming node $h_{r_{in}}$ and the outgoing node $h_{r_{out}}$.

The loss function ($L_{diffphore}$) of DiffPhore consists of three components corresponding to the translation, rotation and torsion diffusion kernels:

$$L_{diffphore} = ||\alpha - \alpha'||^2 + ||\beta - \beta'||^2 + ||\gamma - \gamma'||^2 \tag{26}$$

where α', β', γ' are the precomputed labels in the training dataset. In this way, the loss function ($L_{diffphore}$) enforces the conformation generator to estimate the denoising directions at each step.

Calibrated conformation sampler

Due to the auto-regressive generation fashion, diffusion models usually suffer from the exposure bias, which is caused by the input mismatch between the training and the inference phases. In particular, the conformation generator in the training process takes a perturbed conformation x_t as input and the corresponding scores ($\alpha_t, \beta_t, \gamma_t$) as the labels. By contrast, the conformation generator is fed with the predicted conformation $\hat{\mathrm{x}}_t$ during the inference process. To narrow the discrepancy between the training and inference phases, we proposed a calibrated conformation sampler, which mimics the inference process to construct pseudo ligand conformations ($\widetilde{\mathrm{x}}_t$) and corresponding scores ($\widetilde{\alpha}_t, \widetilde{\beta}_t, \widetilde{\gamma}_t$) for model training (see details in Supplementary Table 3 and Supplementary Methods). The pseudo ligand conformations are estimated based on the denoised data points by

DiffPhore and thus alleviating the exposure bias problem.

$$CCSampler(G_l, t) = \begin{cases} x_t, (\alpha_t, \beta_t, \gamma_t)\ if & p > P_{epoch} \\ \tilde{x}_t, (\tilde{\alpha}_t, \tilde{\beta}_t, \tilde{\gamma}_t)\ if & p \le P_{epoch} \end{cases} \quad (27)$$

$$P_{epoch} = p_{\max}\left(1 - \frac{\mu}{\mu + \exp\left(\frac{c \cdot epoch}{\mu}\right)}\right) \quad (28)$$

$$p \sim Uni([0,1]) \quad (29)$$

However, solely relying on these calibrated data points for the model training is infeasible, as the quality of them is inferior to the real data points. Therefore, we utilized an epoch-dependent possibility scheduler to sample the calibrated data as inputs with the probability P_{epoch} and real data points with the probability $(1 - P_{epoch})$. During the training process, the sampling probability P_{epoch} starts from a small value and gradually increases along with the increment of the training epoch as shown in Eq. 28. where $p_{\max}, \mu, c$ are hyperparameters (Supplementary Table 11) that balance the utilization of the two types of training data.

The detailed algorithms of *LPMEncoder*, *CFGenerator*, and *CCSampler*, and the model training details can be found in Supplementary Methods and the source codes.

Pharmacophore fitness scorings

To leverage the advantages of DiffPhore in different virtual screening scenarios, we introduced four fitness scorings to evaluate the degree of the alignment between the generated ligand conformations and the reference pharmacophore model. *Df Score*1 is a basic scoring function considering pharmacophore feature alignment and exclusion sphere collision, which is calculated using an in-situ max-matching approach as provided in Eqs. 30–32:

$$Df\,Score1 = \frac{V_{overlap}}{V_{ref}} - MIN\left(\frac{V_{overlapEX}}{\epsilon}, 1\right) \quad (30)$$

$$V_{overlap} = \sum_{i=1}^{n} C_i W_i \lambda_i f(\theta) \exp\left(\frac{-d_{i,L-i,P}^2}{\sigma_{i,L} + \sigma_{i,P}}\right) \quad (31)$$

$$f(\theta) = \begin{cases} \cos(\theta - \theta_0) & For\ HA, HD\ and\ MB \\ |\cos(\theta)| & For\ AR \\ 1 & For\ other\ features \end{cases} \quad (32)$$

where $V_{overlap}$ represents the total overlap volume between the ligand conformation pharmacophore features (L) and the reference pharmacophore features (P). It is calculated as the sum of individual feature overlap volume, considering scaling factors (C_i), basic weights (W_i), chemical group weights (λ_i), directional differences ($f(\theta)$), tolerance ranges ($\sigma_{i,L}, \sigma_{i,P}$), the distance of the match pharmacophore pair ($d_{i,L-i,P}$). θ_0 is set as 0 (the number of root atoms equals 1) or $\frac{\pi}{3}$ (the number of root atoms larger than 1). V_{ref} represents the total volume of the reference pharmacophore features. $V_{overlapEX}$ denotes the sum of volumes where the ligand atoms overlap with reference exclusion volumes. ϵ is a maximum tolerance for ligand clashing with pharmacophore, set to 500. *Df Score*1 is adopted as the default fitness score for DiffPhore.

Building upon *Df Score*1, *Df Score*2 includes a bias factor that accounts for the percentage of matched pharmacophore feature pairs, with the aim to consider the tolerance of pharmacophore features. It is calculated by Eq. 33, where n is the number of matched pharmacophore pairs and N_{ref} is the total number of reference pharmacophore features. Here, two pharmacophore features are regarded as a matched pair if the distance between them is less longer than their tolerance range.

$$Df\,Score2 = 0.5 * Df\,Score1 + 0.5 * \frac{n}{N_{ref}} \quad (33)$$

Recognizing the significance of anchor pharmacophore features in protein-ligand recognition and practical drug discovery[23], we introduced *Df Score*3 to specifically measure the alignment of anchor pharmacophore features:

$$Df\,Score3 = 0.5 * Df\,Score1 + 0.5 * \frac{V_{overlapAnchor}}{V_{Anchor}} \quad (34)$$

where V_{Anchor} is the total volume of the anchor features in reference pharmacophore model. $V_{overlapAnchor}$ is the sum of volumes accounting for the ligand pharmacophore features overlapping with the reference anchor features.

Taking all the factors into consideration, we also proposed a comprehensive fitness score *Df Score*4:

$$Df\,Score4 = \frac{Df\,Score1 + \frac{n}{N_{ref}} + \frac{V_{overlapAnchor}}{V_{Anchor}}}{3} \quad (35)$$

In addition, *Df Score*5 is specifically designed for target fishing. It further considers the extent to which the number of ligand's pharmacophore features matches the number of pharmacophore features representing the target.

$$Df\,Score5 = Df\,Score1 * \frac{n}{N_{ref} + N_{mol} - n} \quad (36)$$

where N_{mol} is the count of molecule pharmacophore features.

Baseline setup and implementation

The conformation generation tools, including OpenBabel (version 2.4.1) and Conformator (version 1.2.1), were obtained from their official websites. We utilized the official "AutoPH4" plugin to perform pharmacophore modeling in MOE (version 2020.09) and employed the "Compute | Pharmacophore | Search" functionality for ligand-pharmacophore alignment. The aligned poses in MOE are ranked using the default "rmsdx" metric.

The open-access docking programs, including AutoDock Vina, Uni-dock, SMINA, and GNINA, were implemented following their official source code repositories and instructions. We utilized the "prepare_receptor" and "prepare_ligand" scripts in ADFR toolkit for AutoDock Vina, GNINA, and SMINA to prepare the PDBQT files of protein and ligand structures. As for Uni-dock, its official "unidocktools" was employed. To define the conformation search area, we used a box of 20Å × 20Å × 20Å centered on the ligand in complex crystal structure for AutoDock Vina and Uni-dock, and utilized the "--autobox_ligand" option with default buffer range (4 Å) for SMINA and GNINA. The number of binding conformations was set to 10 for all docking baselines, with other parameters kept at their default settings.

All these calculations were performed on a Linux Rocky 9.2 operating system, utilizing Intel(R) Xeon(R) Platinum 8378C CPU @ 2.80 GHz and NVIDIA RTX 4090 GPU.

Evaluation metrics

The metrics including Root Mean Square Deviation (RMSD), PoseBusters test validity (PB-Valid), Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic (BEDROC), Enrichment Factor (EF), and Area Under the Receiver Operating Characteristic Curve

(AUROC), were used for performance evaluations (see details in Supplementary Methods).

Case study: human glutaminyl cyclases

We used DiffPhore to conduct virtual screening for identifying potential inhibitors for sQC and gQC by employing a substrate-mimicking strategy. First, the binding mode of the tripeptide substrate NH_2-Gln-Phe-Ala-$CONH_2$ (QFA) with sQC was obtained *via* our QM/MM calculations. Then, we constructed the pharmacophore modeling from the calculated sQC:**QFA** complex structure to represent the key binding features of QFA with sQC (Supplementary Fig. 13). Given the pharmacophore model, we then utilized DiffPhore to screen potential hit compounds for sQC against the commercial Vitas-M compound library (https://vitasmlab.biz/) with about 1.4 million compounds available for quick purchase. To efficiently screen the large compound library, we implemented a pharmacophore fingerprint filtering approach, which evaluates the matching context based on the number and types of pharmacophore features, without accounting for their 3D alignment. A pharmacophore fingerprint similarity cutoff of 0.6 was set. To this end, we identified 91,229 ligands from the Vitas-M compound library for subsequent screening by DiffPhore (20 poses generated for each ligand). Through manual inspection, we picked 15 structurally distinct compounds from the top-ranked hits with *Df Score*1 for experimental verification.

sQC/gQC/PGP-1 protein expression and purification

We followed the protocols from our previous study[58] for the expression and purification of human sQC (amino acids 33-361), gQC (amino acids 53-382), and the auxiliary enzyme PGP-1 (amino acids 1-215) (see details in Supplementary Methods).

sQC/gQC/PGP-1 inhibition activity assays

All compounds were tested for their inhibitory activity on sQC, gQC, and PGP-1 in the assay buffer (25 mM Tris-HCl, 150 mM NaCl, 10% glycerol, pH 8.0) as described previously[58] (see details in Supplementary Methods). All determinations were tested in triplicate.

Thermal shift assays

The sQC or gQC enzymes (5 μM) were first incubated with test compounds (50 μM) or a vehicle at room temperature for 20 minutes in Tris-HCl buffer (25 mM Tris-HCl, 150 mM NaCl, 10% glycerol, pH 8.0). Then, the SYPRO ORANGE dye (10× concentration) was added, and the fluorescence was promptly quantified using a fluorescence quantitative PCR instrument. The temperature was incrementally increased from 30 to 95 °C, rising by 1 °C per cycle. The resulting fluorescence intensity versus temperature was analyzed using GraphPad Prism to determine the melting temperature (T_m) values.

Co-crystallization, data collection, and analysis

The hanging-drop vapor diffusion method was employed for co-crystallization experiments. The purified sQC proteins (8 mg/mL) were incubated with **5**/**13** (3.9 mM) on ice for 2 h and then centrifuged at 15,777 × *g* for 10 min to remove insoluble materials. Crystals were grown under the condition: 12–16% (v/v) polyethylene glycol 4000, 0.2 M $MgCl_2$ and 0.1 M Tris-HCl at pH 8.5. The protein solution was mixed with the reservoir solution at a 1:1 ratio. The crystals were cryoprotected with the mother liquor supplemented with 30% (v/v) glycerol prior to harvesting. Data collection was performed at the BL18U1 beamline at the Shanghai Synchrotron Radiation Facility. The diffraction data were processed using XDS[59] or AutoXP[60], followed by structural determination with PHENIX[61] and WinCoot[62]. We utilized the existing crystal structure of sQC (PDB code 3PBB) as the template in the molecular replacement step. The crystal structures of sQC:**5** (PDB code 9ISD) and sQC:**13** (PDB code 9IVV) are available in the Protein Data Bank.

Reporting summary

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

Data availability

The LigPhoreSet and CpxPhoreSet datasets for training and evaluation are available in Zenodo[63] (https://doi.org/10.5281/zenodo.14819917). The PDBBind set is available at http://pdbbind.org.cn. The PoseBusters set is available at https://zenodo.org/record/8278563. The DUD-E set is available at http://dude.docking.org. The ZINC database is available at https://zinc20.docking.org. The crystal structure of sQC used as the template in structural determination is available in the Protein Data Bank under the accession code 3PBB. Crystallographic data for sQC:**5** and sQC:**13** reported in this study are available in Protein Data Bank under the accession codes 9ISD and 9IVV. The crystal structure of sQC:QFA analog used for structure comparison is available in Protein Data Bank under the accession codes 6YI1. Source data is provided with this paper as a Source Data file.

Code availability

The source code is available in Zenodo[64] (https://doi.org/10.5281/zenodo.14818730), GitHub repository[65] (https://github.com/VicFisher/DiffPhore) and our project website (https://diffphore.ddtmlab.org).

References

1. Wang, H. et al. Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60 (2023).
2. Schneider, P. et al. Rethinking drug design in the artificial intelligence era. *Nat. Rev. Drug Discov.* **19**, 353–364 (2020).
3. Mullowney, M. W. et al. Artificial intelligence for natural product drug discovery. *Nat. Rev. Drug Discov.* **22**, 895–916 (2023).
4. Theodoris, C. V. et al. Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).
5. Ren, F. et al. A small-molecule TNIK inhibitor targets fibrosis in preclinical and clinical models. *Nat. Biotechnol.* **43**, 63–75 (2024).
6. Catacutan, D. B. et al. Machine learning in preclinical drug discovery. *Nat. Chem. Biol.* **20**, 960–973 (2024).
7. Pandey, M. et al. The transformational role of GPU computing and deep learning in drug discovery. *Nat. Mach. Intell.* **4**, 211–221 (2022).
8. Perez-Lopez, R. et al. A guide to artificial intelligence for cancer researchers. *Nat. Rev. Cancer* **24**, 427–441 (2024).
9. Allenspach, S. et al. Neural multi-task learning in drug design. *Nat. Mach. Intell.* **6**, 124–137 (2024).
10. Tropsha, A. et al. Integrating QSAR modelling and deep learning in drug discovery: the emergence of deep QSAR. *Nat. Rev. Drug Discov.* **23**, 141–155 (2024).
11. Du, Y. et al. Machine learning-aided generative molecular design. *Nat. Mach. Intell.* **6**, 589–604 (2024).
12. Munson, B. P. et al. De novo generation of multi-target compounds using deep generative chemistry. *Nat. Commun.* **15**, 3636 (2024).
13. Corso, G. et al. DiffDock: diffusion steps, twists, and turns for molecular docking. In *International Conference on Learning Representations* (ICLR, 2023).
14. Stärk, H. et al. EquiBind: geometric deep learning for drug binding structure prediction. In *Proc. International Conference on Machine Learning*. **162**, 20503–20521 (PMLR, 2022).
15. Lu, W. et al. TANKBind: Trigonometry-aware neural networks for drug-protein binding structure prediction. *Adv. Neural Inf. Process. Syst.* **35**, 7236–7249 (2022).
16. Zhang, X. et al. Efficient and accurate large library ligand docking with KarmaDock. *Nat. Comput. Sci.* **3**, 789–804 (2023).
17. Zhang, Y. et al. E3Bind: an end-to-end equivariant network for protein-ligand docking. In *International Conference on Learning Representations* (ICLR, 2023).

18. Guan, J. et al. 3D Equivariant diffusion for target-aware molecule generation and affinity prediction. In *International Conference on Learning Representations* (ICLR, 2023).
19. 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).
20. Zhang, O. et al. Learning on topological surface and geometric structure for 3D molecular generation. *Nat. Comput. Sci.* **3**, 849–859 (2023).
21. Jiang, Y. et al. PocketFlow is a data-and-knowledge-driven structure-based molecular generative model. *Nat. Mach. Intell.* **6**, 326–337 (2024).
22. Schaller, D. et al. Next generation 3D pharmacophore modeling. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **10**, e1468 (2020).
23. 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).
24. Huang, Q. et al. PhDD: A new pharmacophore-based de novo design method of drug-like molecules combined with assessment of synthetic accessibility. *J. Mol. Graph. Model.* **28**, 775–787 (2010).
25. Dixon, S. L. et al. PHASE: a novel approach to pharmacophore modeling and 3d database searching. *Chem. Biol. Drug Des.* **67**, 370–372 (2006).
26. Barnum, D. et al. Identification of common functional configurations among molecules. *J. Chem. Inf. Comput. Sci.* **36**, 563–571 (1996).
27. Taminau, J. et al. Pharao: pharmacophore alignment and optimization. *J. Mol. Graph. Model.* **27**, 161–169 (2008).
28. Sunseri, J. et al. Pharmit: interactive exploration of chemical space. *Nucleic Acids Res.* **44**, W442–W448 (2016).
29. Zhu, H. et al. A pharmacophore-guided deep learning approach for bioactive molecular generation. *Nat. Commun.* **14**, 6234 (2023).
30. Seo, S. & Kim, W. Y. PharmacoNet: accelerating large-scale virtual screening by deep pharmacophore modeling. *Chem. Sci.* **15**, 19473–19487 (2024).
31. Buttenschoen, M. et al. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chem. Sci.* **15**, 3130–3139 (2024).
32. Mysinger, M. M. et al. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* **55**, 6582–6594 (2012).
33. Li, G.-B. et al. IFPTarget: a customized virtual target identification method based on protein–ligand interaction fingerprinting analyses. *J. Chem. Inf. Model.* **57**, 1640–1651 (2017).
34. Xu, C. et al. Glutaminyl cyclase, diseases, and development of glutaminyl cyclase inhibitors. *J. Med. Chem.* **64**, 6549–6565 (2021).
35. Coimbra, J. R. M. et al. Therapeutic potential of glutaminyl cyclases: Current status and emerging trends. *Drug Discov. Today* **28**, 103644 (2023).
36. Logtenberg, M. E. W. et al. Glutaminyl cyclase is an enzymatic modifier of the CD47- SIRPα axis and a target for cancer immunotherapy. *Nat. Med.* **25**, 612–619 (2019).
37. Barreira da Silva, R. et al. Loss of the intracellular enzyme QPCTL limits chemokine function and reshapes myeloid infiltration to augment tumor immunity. *Nat. Immunol.* **23**, 568–580 (2022).
38. Eberhardt, J. et al. AutoDock Vina 1.2.0: new docking methods, expanded force field, and Python bindings. *J. Chem. Inf. Model.* **61**, 3891–3898 (2021).
39. Yu, Y. et al. Uni-Dock: GPU-accelerated docking enables ultralarge virtual screening. *J. Chem. Theory Comput.* **19**, 3336–3345 (2023).
40. Koes, D. R. et al. Lessons learned in empirical Scoring with smina from the CSAR 2011 benchmarking exercise. *J. Chem. Inf. Model.* **53**, 1893–1904 (2013).
41. McNutt, A. T. et al. GNINA 1.0: molecular docking with deep learning. *J. Cheminform.* **13**, 43 (2021).
42. Cao, D. et al. SurfDock is a surface-informed diffusion generative model for reliable and accurate protein-ligand complex prediction. *Nat Methods* **22**, 310–322 (2025).
43. Cao, D. et al. Generic protein–ligand interaction scoring by integrating physical prior knowledge and data augmentation modelling. *Nat. Mach. Intell.* **6**, 688–700 (2024).
44. Chaput, L. et al. Benchmark of four popular virtual screening programs: construction of the active/decoy dataset remains a major determinant of measured performance. *J. Cheminform.* **8**, 56 (2016).
45. Chen, L. et al. Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. *PLoS One* **14**, e0220113 (2019).
46. Kupski, O. et al. Hydrazides are potent transition-state analogues for glutaminyl cyclase implicated in the pathogenesis of Alzheimer's Disease. *Biochemistry* **59**, 2585–2591 (2020).
47. Liu, Z. et al. Forging the basis for developing protein–ligand interaction scoring functions. *Acc. Chem. Res.* **50**, 302–309 (2017).
48. Su, M. et al. Comparative assessment of scoring functions: the CASF-2016 update. *J. Chem. Inf. Model.* **59**, 895–913 (2019).
49. Irwin, J. J. et al. ZINC20—a free ultralarge-scale chemical database for ligand discovery. *J. Chem. Inf. Model.* **60**, 6065–6073 (2020).
50. Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **39**, 2887–2893 (1996).
51. Morgan, H. L. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *J. Chem. Doc.* **5**, 107–113 (1965).
52. Song, Y. & Ermon, S. Generative modeling by estimating gradients of the data distribution. *Adv. Neural Inf. Process. Syst.* **32**, 11918–11930 (2019).
53. Roberts, G. O. & Tweedie, R. L. J. B. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* **2**, 341–363 (1996).
54. Welling, M. & Teh, Y. W. Bayesian learning via stochastic gradient Langevin dynamics. In: *Proc. 28th International Conference on International Conference on Machine Learning* (2011).
55. Song, Y. et al. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations* (ICLR, 2021).
56. Nikolayev, D. et al. Normal Distribution on the Rotation Group So(3). *Texture Stress Microstruct.* **29**, 201–233 (1997).
57. Jing, B. et al. Torsional diffusion for molecular conformer generation. *Adv. Neural Inf. Process. Syst.* **35**, 24240–24253 (2022).
58. Mou, J. et al. X-ray structure-guided discovery of a potent benzimidazole glutaminyl cyclase inhibitor that shows activity in a Parkinson's Disease mouse model. *J. Med. Chem.* **67**, 8730–8756 (2024).
59. Brehm, W. et al. XDSGUI: a graphical user interface for XDS, SHELX and ARCIMBOLDO. *J. Appl. Crystallogr.* **56**, 1585–1594 (2023).
60. Wang, L. et al. AutoPX: a new software package to process X-ray diffraction data from biomacromolecular crystals. *Acta Crystallogr. D* **78**, 890–902 (2022).
61. Liebschner, D. et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr. D* **75**, 861–877 (2019).
62. Emsley, P. et al. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
63. Yu, J. et al. LigPhoreSet and CpxPhoreSet for training DiffPhore. *Zenodo*, https://doi.org/10.5281/zenodo.14819917 (2025).
64. Yu, J. et al. DiffPhore:v1.0. *Zenodo*, https://doi.org/10.5281/zenodo.14818730 (2025).
65. Yu, J. et al. DiffPhore:v1.0. *GitHub*, https://github.com/VicFisher/DiffPhore (2024).

Acknowledgements
This work is financially supported by the National Key R&D Program of China (2023YFF1204901 to G.B.L.), the National Natural Science Foundation of China (82122065, 82473845, and 82073698 to G.B.L.), the Basic Research Foundation of Sichuan University (2023SCUH0073 to G.B.L.), and the Sichuan Science and Technology Program (2025YFHZ0085 to G.B.L.). We thank the staff from beamlines BL18U1 and BL19U1 at Shanghai Synchrotron Radiation Facility of the National Facility for Protein Science (Shanghai, China) for their great support. We also thank Professor Jin-Liang Yang and Dr. Zhi-Xiong Zhang (Sichuan University) for providing MOE for comparison purposes.

Author contributions
G.B.L. conceived, planned, and supervised this study; J.L.Y. and C.Z. collected and processed the datasets; J.L.Y. designed and trained the model supervised by G.B.L. and X.G.L.; J.L.Y. and J.W.W. performed model validation and case study; X.L.N., J.M., F.B.M., and Y.T.C. performed protein purification, enzymatic activity testing, and co-crystallographic studies; J.L.Y., C.Z., X.L.N., J.M., F.B.M., J.W.W., B.D.T., X.G.L., and G.B.L. analyzed the data; J.L.Y., X.L.N., X.G.L. and G.B.L. wrote 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
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-57485-3.

Correspondence and requests for materials should be addressed to Xiang-Gen Liu or Guo-Bo Li.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.