Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal | View Issue

Cite this: *Chem. Sci.*, 2025, **16**, 17374

All publication charges for this article have been paid for by the Royal Society of Chemistry

Decoding the limits of deep learning in molecular docking for drug discovery

Yue Li,†[a] Jiacai Yi,†[b] Hui Li,[a] Kun Li,[a] Fenghua Kang,[a] Youchao Deng,[a] Chengkun Wu,[b] Xiangzheng Fu,[c] Dejun Jiang*[a] and Dongsheng Cao *[ac]

Received 19th July 2025
Accepted 18th August 2025

DOI: 10.1039/d5sc05395a

rsc.li/chemical-science

Structure-based molecular docking, a cornerstone of computational drug design, is undergoing a paradigm shift fueled by deep learning (DL) innovations. However, the rapid proliferation of DL-driven docking methods has created uncharted challenges in translating *in silico* predictions to biomedical reality. Here, we delve into the performance and prospects of traditional methods and state-of-the-art DL docking paradigms—encompassing generative diffusion models, regression-based architectures, and hybrid frameworks—across five critical dimensions: pose prediction accuracy, physical plausibility, interaction recovery, virtual screening (VS) efficacy, and generalization across diverse protein–ligand landscapes. We reveal that generative diffusion models achieve superior pose accuracy, while hybrid methods offer the best balance. Regression models, however, often fail to product physically valid poses, and most DL methods exhibit high steric tolerance. Furthermore, our analysis reveals significant challenges in generalization, particularly when encountering novel protein binding pockets, limiting the current applicability of DL methods. Finally, we explore failure mechanisms from a model perspective and propose optimization strategies, offering actionable insights to guide docking tool selection and advance robust, generalizable DL frameworks for molecular docking.

Introduction

Prolonged timelines, substantial costs, and inherent uncertainties impede drug development, a process critically dependent on effective lead discovery and optimization.[1–3] Structure-based molecular docking methods have become indispensable computational tools for lead discovery and optimization over recent decades, offering unique advancements in predicting protein–ligand interactions.[4,5] The efficacy of a drug hinges on the specific interactions between the drug molecule and its target, typically a protein. Effective drug–target interaction necessitates close proximity and appropriate orientation, allowing key molecular surface regions to fit precisely. Subsequently, driven by this interaction, the molecular conformations adjust appropriately, ultimately forming a relatively stable complex conformation and exerting the expected biological activity.

Molecular docking technology, as a powerful computational method, aims to computationally simulate and find the stable complex conformation between a protein and a ligand. It also quantitatively evaluates the binding affinity through scoring functions (SFs),[6,7] providing the corresponding binding free energy. Traditional physics-based docking tools, such as Glide SP[8] and AutoDock Vina,[9] typically consist of two components: SF and conformational search algorithm. The SF estimates the binding energy of a ligand in a hypothesized binding pose, while the search algorithm explores the conformational space to find the pose with the most favorable score assigned by the SF.[4] However, these traditional methods face significant limitations. Their reliance on empirical rules and heuristic search algorithms results in computationally intensive processes and inherent inaccuracies, constraining the precision of docking outcomes.

Recent advances in computational power and the accumulation of massive data have promoted the rapid development of artificial intelligence (AI) particularly DL, in the pharmaceutical field. AlphaFold's[10] groundbreaking success in protein structure prediction has inspired researchers to re-envision traditional molecular docking with DL methodologies, potentially transforming this critical process.[11–16] DL-based docking methods offer distinct advantages by overcoming the limitations of traditional approaches. These methods directly utilize the 2D chemical information of ligands and the 1D sequence or 3D structural data of proteins as inputs, leveraging the robust learning and processing capabilities of DL models to predict protein–ligand binding conformations and their associated binding free energies. This approach bypasses computationally

[a]*Xiangya School of Pharmaceutical Sciences, Central South University, Changsha 410013, Hunan, P.R. China. E-mail: jiang_dj@zju.edu.cn; oriental-cds@163.com*
[b]*College of Computer, National University of Defense Technology, Changsha 410073, Hunan, China*
[c]*School of Chinese Medicine, Hong Kong Baptist University, Hong Kong, SAR 999077, China*
† The first two authors should be regarded as joint first authors.

intensive conformational searches by leveraging the parallel computing power of DL models, enabling efficient analysis of large datasets and accelerated docking. Moreover, DL models can extract complex patterns from vast datasets, significantly enhancing the accuracy of docking predictions and providing a more reliable foundation for drug discovery.[17]

However, most DL-based docking studies have primarily focused on binding pose prediction, often relying on a single evaluation metric, such as the root-mean-square deviation (RMSD) of the ligand. Buttenschoen *et al.*[18] developed the PoseBusters toolkit to systematically evaluate docking predictions against chemical and geometric consistency criteria, including bond length/angle validity, stereochemistry preservation, and protein–ligand clash detection, revealing that many DL methods produce physically implausible structures despite favorable RMSD scores. More importantly, these methods often overlook the biological relevance of predicted poses—specifically, their ability to recapitulate key protein–ligand interactions. Recent work has demonstrated that even when RMSD is acceptable, AI-based docking models frequently fail to recover critical molecular interactions essential for biological activity.[19] Moreover, a critical concern for drug researchers is the ability of molecular docking methods to accurately identify hit compounds in VS,[20–23] which demands not only precise binding pose prediction but also robust generalization and screening capabilities. Recognizing these challenges, Gu *et al.*[23] conducted a comprehensive benchmark of both AI-powered and traditional physics-based docking methods across rigorously curated datasets designed to mimic real-world VS scenarios. However, the generalization performance of docking methods beyond

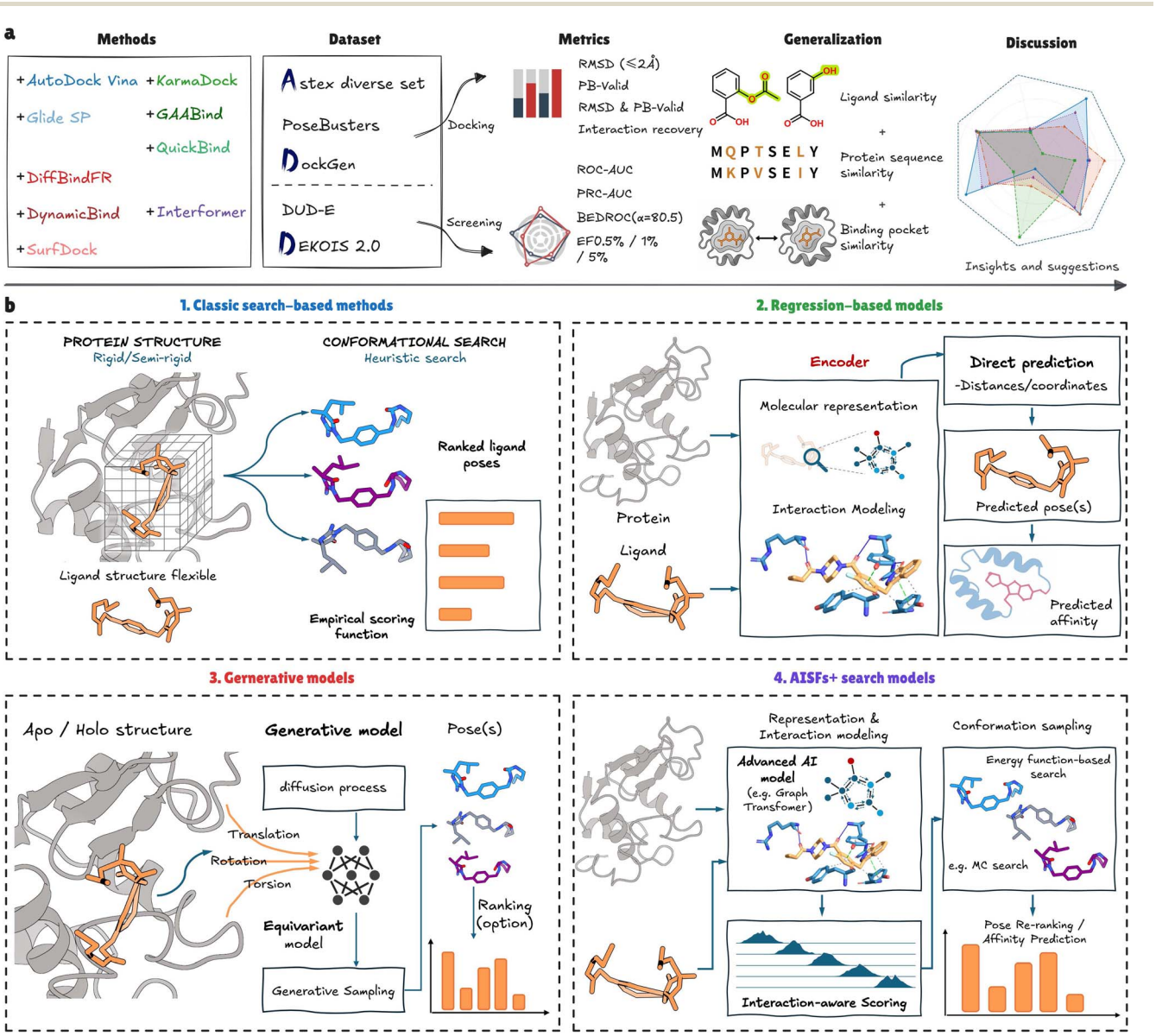

Fig. 1 Overview of systematic benchmarking workflow and molecular docking paradigms. (a) Schematic representation of the overall workflow employed in this study. (b) Conceptual illustration of the four molecular docking paradigms, delineating their distinct approaches and methodologies.

training datasets and their practical utility in lead discovery remain underexplored, significantly limiting their widespread adoption in drug development.

To address these challenges, this study conducts a systematic, multidimensional evaluation of existing small molecule-protein docking methods, encompassing traditional physics-based approaches (Glide SP[8] and AutoDock Vina[9]), generative diffusion models (SurfDock,[11] DiffBindFR[14] and DynamicBind[12]), regression-based models (KarmaDock,[12] GAABind[9] and QuickBind[24]), and hybrid methods (Interformer[15]) that integrate traditional conformational searches with AI-driven SFs (Fig. 1b). We evaluated these methods across diverse benchmark datasets, assessing their performance in binding pose prediction, physical validity, interaction recovery, VS efficacy, and generalization across three dimensions: protein sequence similarity, ligand topology, and protein binding pocket structural similarity (Fig. 1a). Our study offers several critical contributions to the field:

- We provide a comprehensive multidimensional evaluation of traditional and contemporary DL-based molecular docking methods. This involved rigorous comparison across multiple datasets and performance indicators, with a particular emphasis on generalization to unseen protein sequences, binding pockets, and structurally distinct ligands.
- We deliver a holistic assessment of binding and affinity, critically evaluating the practical screening performance by integrally considering both binding conformation and affinity prediction—aspects crucial for real-world drug development.
- We formulate targeted optimization strategies based on a detailed analysis of each method's strengths, weaknesses, and optimal application contexts. These strategies offer actionable pathways for enhancing diffusion model sampling, refining regression model loss functions, and improving hybrid method search efficiency.

Results and discussion

Comparative docking accuracy and physical validity across benchmarks

We evaluated docking performance using three benchmark datasets designed to rigorously test method capabilities: the Astex diverse set[25] (known complexes), the PoseBusters benchmark set[18] (unseen complexes), and the DockGen[26] dataset (novel protein binding pockets). Details of the evaluation protocols are provided in the Materials and methods section (Docking methods).

As depicted in Fig. 2a and S1, a striking pattern emerges, enabling the classification of the nine evaluated docking methods into four distinct performance tiers based on PB-valid and combined success rates (RMSD $\leq$ 2 Å & PB-valid): traditional methods > hybrid AI scoring with traditional conformational search > generative diffusion methods > regression-based methods. Notably, DynamicBind, designed specifically for blind docking, exhibits performance slightly lagging behind other generative diffusion methods and aligns with regression-based methods in a separate tier. This stratification underscores the diverse strengths and limitations of each approach across to known complexes, unseen complexes, and novel binding pockets.

The generative diffusion method (the red series in Fig. 2a) SurfDock exhibited exceptional pose accuracy (Fig. 2c), achieving RMSD $\leq$ 2 Å success rates exceeding 70% across all datasets: 91.76% (Astex), 77.34% (PoseBusters), and 75.66% (DockGen). This highlights its proficiency in generating accurate docking poses, likely due to its advanced generative modeling capabilities. However, its suboptimal PB-valid scores (63.53%, 45.79%, 40.21%)—reveal deficiencies in modeling critical physicochemical interactions, such as steric clashes or hydrogen bonding, leading to moderate combined success rates (RMSD $\leq$ 2 Å & PB-valid) of 61.18%, 39.25%, and 33.33%, respectively. The DiffBindFR variants (MDN and SMINA) displayed moderate pose accuracy, with RMSD $\leq$ 2 Å rates of 75.29% and 75.30% (Astex), 50.93% and 47.66% (PoseBusters), and 30.69% and 35.98% (DockGen). Yet, their physical validity faltered on more challenging datasets, with PB-valid rates of 47.20% and 46.73% (PoseBusters) and 47.09% and 45.50% (DockGen), resulting in combined success rates of 33.88%, 34.58% (PoseBusters), and 18.52%, 23.28% (DockGen). These results suggest that while diffusion models excel in pose generation, their reliance on learned distributions may overlook physical constraints, particularly on unseen or novel systems.

In contrast, the traditional method (the blue series in Fig. 2a) Glide SP consistently excelled in physical validity, maintaining PB-valid rates above 94% across all datasets: 97.65% (Astex), 97.90% (PoseBusters), and 94.18% (DockGen). This robustness translated into high combined success rates of 70.59%, 57.94% and 40.21%. AutoDock Vina also demonstrated strong physical validity, with PB-valid rates of 82.35%, 78.97% and 88.36%, and competitive combined success rates, notably 40.74% on DockGen. These findings reaffirm the enduring efficacy of traditional approaches, particularly in maintaining structural integrity across diverse datasets, underscoring the enduring reliability of physics-driven approaches.

The hybrid method (the purple series in Fig. 2a) Interformer, which couples traditional conformational search with DL-enhanced scoring, represents a promising synthesis of data-driven and physics-driven approaches. Interformer-Energy achieved competitive accuracy (81.18% RMSD $\leq$ 2 Å on Astex, 59.58% on PoseBusters, 46.56% on DockGen) while retaining robust physical validity (72.94%, 71.96%, and 69.84% PB-valid, respectively), yielding combined success rates of 68.24%, 46.26%, and 34.39%. Interformer-PoseScore, relying on DL rescoring alone, underperformed relative to Interformer-Energy (71.76%, 57.24%, and 42.86% RMSD $\leq$ 2 Å; 71.76%, 70.56%, and 70.37% PB-valid; 55.29%, 44.16%, and 29.63% combined), suggesting that integrating conformational sampling with DL scoring enhances overall performance. This synergy highlights the potential of hybrid strategies to balance accuracy and physical plausibility, offering a pathway to overcome limitations inherent in purely DL-based methods.

Regression-based DL methods (QuickBind, GAABind and KarmaDock) (the green series in Fig. 2a), which predicts a single optimal pose without sampling a distribution of possible conformations, generally underperformed, characterized by

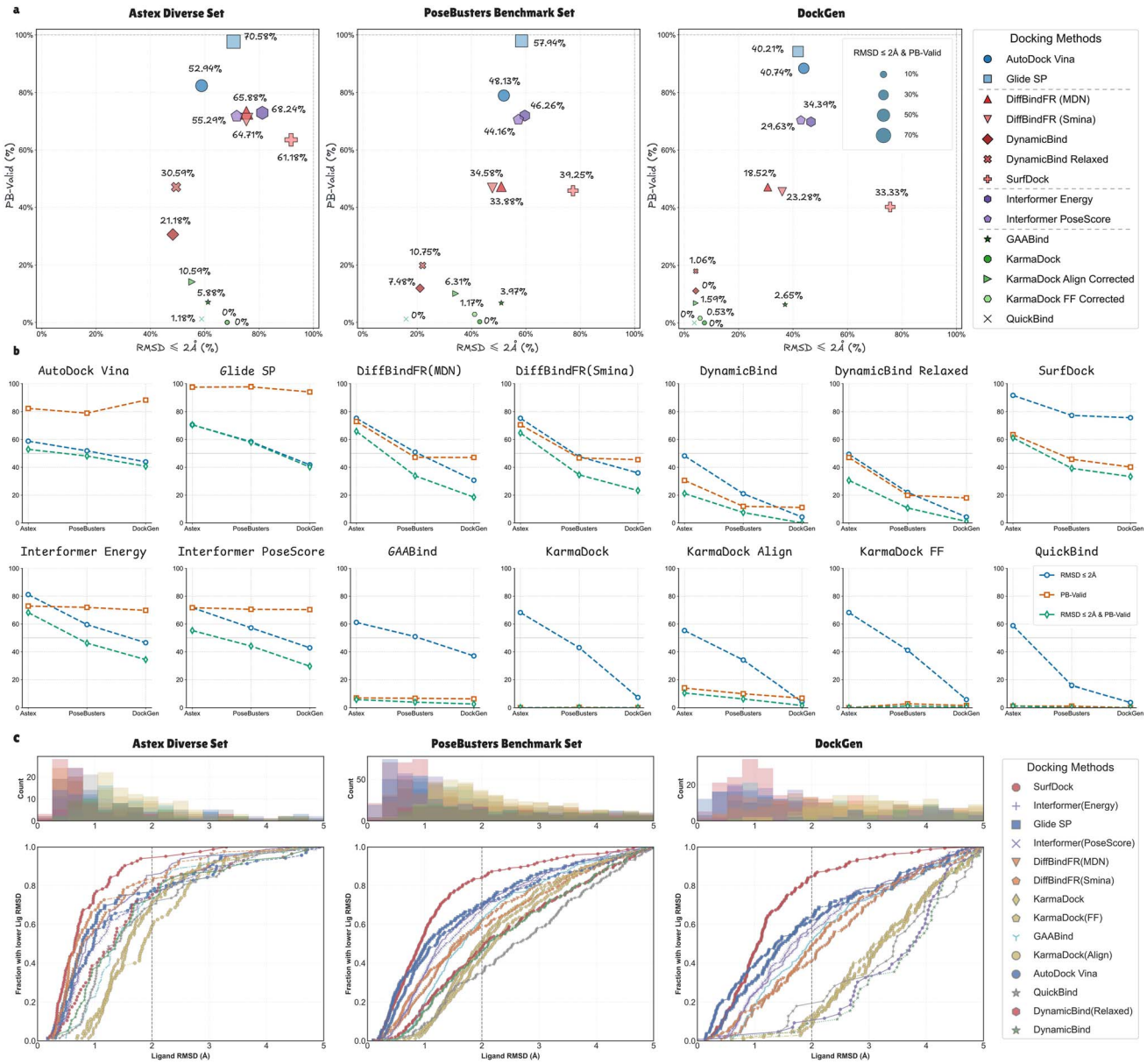

Fig. 2 Docking accuracy and physical validity across benchmark datasets. Comparative performance of docking methods across Astex diverse set, PoseBusters benchmark set and DockGen datasets. (a) Scatter plots show the percentage of successful docking cases for each method, evaluated by RMSD ≤ 2 Å (*x*-axis) and PB-valid rate (*y*-axis) on Astex, PoseBusters, and DockGen sets, with marker sizes and annotated values reflecting the combined success rate (RMSD ≤ 2 Å & PB-valid). (b) Trends in RMSD ≤ 2 Å, PB-valid, and combined success rates across the three datasets for each method. (c) Cumulative distribution of ligand RMSD values. Comparison of ligand RMSD (Å) distributions for the top-ranked pose predicted by each docking method on the three datasets. Top row: histograms showing the distribution counts of RMSD values for representative methods (colors match bottom row). Bottom row: cumulative distribution function plots showing the fraction of predictions with an RMSD lower than the value on the *x*-axis. The vertical dashed gray line indicates the 2 Å threshold.

notably low physical validity and poor combined success rates. KarmaDock exhibited PB-valid rates of 0.00% on Astex and DockGen, with a marginal 0.23% on PoseBusters, resulting in combined success rates of 0.00% across all datasets. QuickBind followed a similar trend, with PB-valid rates of 1.18%, 1.17% and 0.00%, and combined success rates of 1.18%, 0.00%, and 0.00%, respectively. Even corrected variants of KarmaDock (Align-corrected: 6.31% combined success rates on PoseBusters; FF-corrected: 1.17%) showed only marginal improvements, underscoring inherent limitations in regression-based approaches for ensuring physical plausibility. GAABind, with PB-valid rates of 7.06% (Astex), 6.78% (PoseBusters), and 6.35% (DockGen), achieved combined success rates of 5.88%, 3.97%, and 2.65%, reflecting a consistent inability to model complex intermolecular interactions effectively.

A marked decline in DL method performance from the Astex diverse set to PoseBusters and further to DockGen (Fig. 2b) reveals significant generalization limitations, particularly for

novel binding pockets. SurfDock's RMSD ≤ 2 Å rate decreased from 91.76% to 77.34% to 75.66%, while its combined success rate dropped from 61.18% to 39.25% to 33.33%. Similarly, interformer-Energy's combined success declined from 68.24% to 46.26% to 34.39%, and DiffBindFR's performance tapered from 65.88% (Astex) to 33.88–34.58% (PoseBusters) to 18.52–23.28% (DockGen). Surprisingly, PB-valid rates for DL methods consistently decreased across datasets—e.g., SurfDock from 63.53% to 45.79% to 40.21%, and KarmaDock-Align from 14.12% to 10.05% to 6.88%—a trend less pronounced in traditional methods (*e.g.*, Glide SP: 97.65% to 97.90% to 94.18%). This observation raises a profound question: Can DL models be trained to prioritize physical plausibility without sacrificing the flexibility of generated conformations?

Dissecting factors influencing physicochemical validity in deep learning docking methods

Given the observed deficiencies in physicochemical validity among DL-based docking methods, as highlighted in the previous section, we sought to investigate the specific factors contributing to these shortcomings. To this end, we analyzed interaction recovery across the Astex diverse set, PoseBusters benchmark set, and DockGen datasets to determine whether DL methods effectively learn protein–ligand (PL) binding interactions or merely fit to dataset-specific biases. Additionally, we dissected the PB-valid metric into its three components—chemical validity and consistency, intramolecular validity, and intermolecular validity—to pinpoint the drivers of performance limitations.

Interaction recovery: learning true binding or fitting biases? To assess whether DL methods capture true PL binding interactions, we evaluated interaction recovery, defined as the percentage of correctly predicted key interactions (*e.g.*, hydrogen bonds, ionic interactions and π–π stacking) between the ligand and protein binding pocket, across a range of thresholds (0.1 to 1.0) (Fig. 3a). For consistency, we focus on the threshold of 0.5 (50% recovery), which balances sensitivity and specificity in identifying meaningful interactions (Fig. 3c).

As shown in Fig. 3c, SurfDock's performance in interaction recovery was highly competitive with the traditional method Glide SP across all three datasets, achieving 92.68%, 77.99%, and 71.75% across Astex, PoseBusters, and DockGen, respectively, compared to Glide SP's 82.93%, 78.95%, and 64.41%. These results indicate that SurfDock effectively learns critical PL interactions, rather than overfitting to dataset biases, challenging the notion that DL methods lack the capacity to model binding physics. Other diffusion-based methods, such as DiffBindFR (MDN: 73.17%, 57.89%, 48.59%; SMINA: 76.83%, 60.29%, 52.54% across all three datasets) and DynamicBind (52.44%, 28.95%, 11.30%), showed more variability, with DynamicBind particularly struggling on DockGen, suggesting challenges in generalizing to novel binding pockets.

Among traditional methods, AutoDock Vina maintained solid performance (73.17%, 63.88%, 63.84%), though it lagged behind Glide SP and SurfDock. The hybrid method Interformer-Energy also performed well (80.49%, 68.90%, 63.84%), reinforcing its balanced approach. Regression-based methods, however, underperformed significantly: KarmaDock (54.88% on Astex, 46.17% on PoseBusters, 13.56% on DockGen) and QuickBind (52.44%, 21.77%, 4.52%) exhibited low recovery rates, particularly on novel binding pocket, likely due to their single-point prediction paradigm limiting the exploration of diverse interaction modes. However, GAABind maintained a robust interaction recovery, offering a promising avenue for improvement.

Despite SurfDock's strong interaction recovery, its PB-valid rates remained suboptimal (63.53% on Astex, 45.79% on PoseBusters, 40.21% on DockGen), prompting a deeper investigation into the factors hindering its physicochemical validity. This discrepancy raises a critical question: If deep learning methods like SurfDock can accurately predict binding interactions, what barriers prevent them from achieving high physical validity?

Dissecting the PB-valid metric: chemical, intramolecular, and intermolecular validity. To identify the specific factors driving the observed limitations in physicochemical validity, we decomposed the PB-valid metric into its three constituent components: (1) chemical validity and consistency, ensuring ligand molecular accuracy (*e.g.*, valency, stereochemistry, protonation); (2) intramolecular validity, verifying ligand geometry and conformational energy without steric clashes; and (3) intermolecular validity, verifying no spatial conflicts between the ligand and protein or cofactors. These metrics were analyzed across all methods on the three datasets, as visualized in Fig. 3b.

Our analysis revealed that diffusion-based methods, SurfDock, DiffBindFR and DynamicBind, achieved levels of chemical validity and consistency and intramolecular validity comparable to traditional conformational sampling algorithms like Glide SP and AutoDock Vina. These results indicate that diffusion-based methods effectively model ligand-specific properties at a level competitive with traditional methods.

However, a stark contrast emerged in intermolecular validity, which assesses spatial conflicts between the ligand and protein. SurfDock's intermolecular validity scores were significantly lower—70.59% (Astex), 48.36% (PoseBusters), and 43.39% (DockGen)—compared to Glide SP's 100.0%, 99.07%, and 98.41%, and AutoDock Vina's 87.06%, 82.48%, and 94.18% and Interformer-Energy's 90.59%, 82.94%, and 81.48%. DiffBindFR (MDN) showed similar trends (75.29%, 48.13%, 48.68%) and DiffBindFR (SMINA) (72.94%, 47.90%, 47.09%), while DynamicBind performed worse (32.94%, 12.62%, 6.88%), with its relaxed variant improving slightly (49.41%, 20.33%, 8.99%). This suggests that spatial conflicts with the protein are the primary factor dragging down the PB-valid metric for diffusion-based methods. These methods, while adept at generating accurate poses (*e.g.*, SurfDock's RMSD ≤ 2 Å of 91.76% on Astex) and recovering interactions (92.68% at 0.5), often position ligands in ways that lead to steric clashes.

Regression-based methods (QuickBind, GAABind, KarmaDock) exhibited no distinct advantage across any of the PB-valid components. On Astex, QuickBind showed 68.24% chemical validity, 3.53% intramolecular validity, and 56.47%

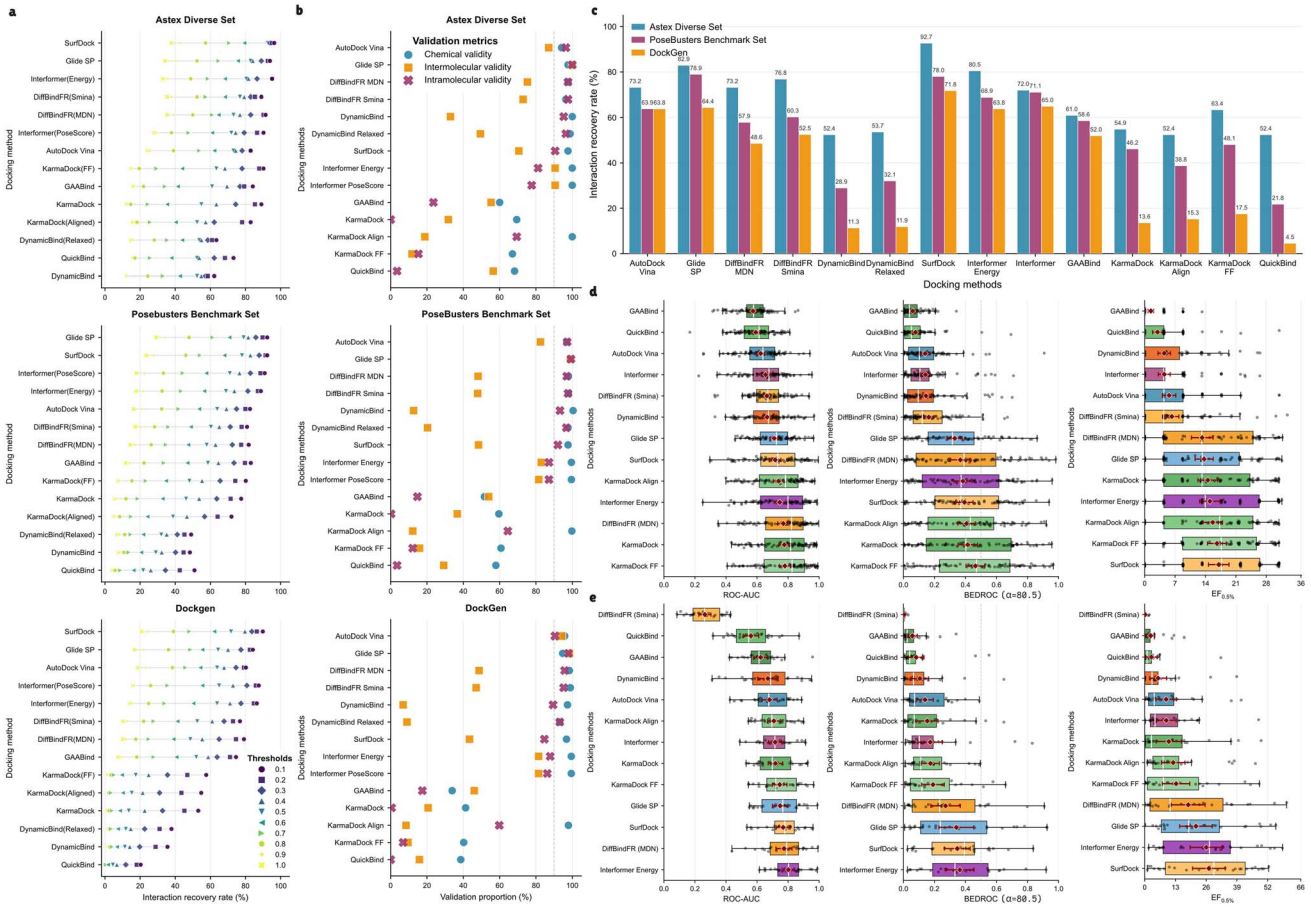

Fig. 3 Evaluating prediction performance across diverse datasets. (a) Evaluation of each docking method's ability to recover native protein–ligand interactions (identified by ProLIF analysis of crystal structures) across the Astex diverse set, PoseBusters benchmark set, and DockGen datasets. The scatter plots show the performance of docking methods at various thresholds (0.1 to 1.0), where a threshold indicates the minimum percentage of interactions recovered considered as successful (*e.g.*, 0.1 represents 10% recovery). Each point represents the recall rate at a specific threshold, with colors and shapes corresponding to the methods as indicated in the legends. (c) Bar chart comparing interaction recovery across the three datasets at a 0.5 threshold (50% recovery). (b) Breakdown of the PoseBusters physical validity (PB-valid) metric into its three main components for each docking method across the Astex diverse set, PoseBusters benchmark set, and DockGen datasets. The scatter plots display the performance of docking methods in terms of three PB-valid components: chemical validity (blue circles), intermolecular validity (orange squares), and intramolecular validity (purple crosses). Virtual screening performance on DEKOIS2.0 (d) and DUD-E (26) (e) datasets, including $EF_{0.5\%}$, ROC-AUC and BEDROC ($\alpha = 80.5$). Boxplots indicate the median (center line), interquartile range (IQR, box limits), and 1.5 × IQR whiskers. Mean values are marked with a red diamond.

intermolecular validity, while KarmaDock had 69.41%, 0.00%, and 31.76%. On PoseBusters, QuickBind recorded 57.94%, 3.50%, and 29.21%, and KarmaDock 59.58%, 0.23%, and 36.68%. On DockGen, QuickBind had 38.62%, 0.00%, and 15.87%, and KarmaDock 41.27%, 0.53%, and 20.63%. These low scores, particularly in intramolecular validity, reflect their direct prediction of atomic coordinates in 3D space, as small errors in predicted coordinates can lead to significant distortions, such as bond length violations or steric clashes within the ligand.

Comprehensive analysis of virtual screening performance

A central concern in drug discovery is whether molecular docking tools can effectively identify lead compounds by screening hit or lead candidates from large-scale chemical libraries.[20] This section evaluates the VS performance of docking models using two benchmark datasets, DEKOIS2.0[27] and DUD-E[28] (26 representative targets).

Results, summarized in Table 1, Fig. 3d and e, with target-specific statistics presented as heatmaps in Fig. S3 and S4. The traditional method Glide SP achieved above-average performance, with averages of ROC-AUC 0.714 and 0.750 and medians of 0.726 and 0.766 across DEKOIS2.0 (Fig. 3d) and DUD-E (Fig. 3e), respectively, and average $EF_{0.5\%}$ from 14.758 to 21.669 with medians of 13.243 and 18.637. Its physics-based scoring ensures consistent ranking, as the tight gap between average and median (*e.g.*, ROC-AUC median close to average) suggests stability across targets. In contrast, AutoDock Vina underperformed, with ROC-AUC of 0.623, 0.681 (Avg.) and 0.638, 0.676 (Med.), and $EF_{0.5\%}$ of 5.630, 9.068 (Avg.) with medians of 4.429 and 4.028. The lower medians and wider gaps between average and median values (*e.g.*, $EF_{0.5\%}$ on DUD-E: 9.068 *vs.* 4.028) indicate inconsistent performance on DUD-E,

Table 1 Virtual screening performance of docking models on the DEKOIS2.0 and DUD-E (26 representative targets)[a]

		ROC-AUC		PRC-AUC		BEDROC ($\alpha = 80.5$)		$EF_{0.5\%}$		$EF_{1\%}$		$EF_{5\%}$	
Dataset	Model	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.
DEKOIS 2.0	AutoDock Vina	0.623	0.638	0.091	0.066	0.140	0.103	5.630	4.429	4.768	2.385	2.919	2.500
	Glide SP	0.714	0.726	0.224	0.185	0.352	0.315	14.758	13.243	11.979	9.958	5.891	5.500
	Interformer	0.656	0.677	0.096	0.066	0.142	0.107	4.509	4.386	4.605	2.383	3.358	2.625
	Interformer_Energy	0.746	0.800	0.271	0.225	0.384	0.372	15.008	13.951	13.527	11.885	7.066	6.634
	DiffBindFR_MDN	0.770	0.822	0.278	0.241	0.367	0.391	13.251	13.157	11.963	11.673	7.073	7.487
	DiffBindFR_Smina	0.664	0.670	0.105	0.066	0.164	0.122	6.255	4.386	5.500	4.712	3.400	2.539
	DynamicBind	0.665	0.674	0.106	0.070	0.145	0.095	4.506	0	4.426	2.360	3.276	2.959
	SurfDock	0.717	0.733	0.267	0.204	0.393	0.369	**17.064**	**17.714**	14.588	14.273	6.712	6.484
	GAABind	0.574	0.578	0.057	0.043	0.060	0.039	1.421	0	1.708	0	1.665	1.490
	KaramDock	0.776	0.817	0.306	0.241	0.414	0.402	14.535	13.211	13.410	11.875	7.739	7.500
	KaramDock_aligned	0.739	0.784	0.273	0.247	0.407	0.431	15.687	17.271	14.092	14.250	6.786	6.886
	KaramDock_FF	**0.780**	**0.826**	**0.341**	**0.319**	**0.464**	**0.472**	16.633	17.643	**15.278**	**14.950**	**8.375**	**8.862**
	QuickBind	0.592	0.613	0.062	0.048	0.078	0.051	2.952	0	2.385	0	2.062	1.500
DUD-E (26)	AutoDock Vina	0.681	0.676	0.077	0.041	0.139	0.072	9.068	4.028	7.490	3.309	4.102	3.144
	Glide SP	0.750	0.766	0.243	0.118	0.341	0.238	21.669	18.637	17.619	13.162	7.177	5.773
	Interformer	0.717	0.716	0.103	0.048	0.173	0.102	9.113	4.380	8.044	5.429	4.760	4.472
	Interformer_Energy	**0.802**	**0.803**	**0.251**	0.185	**0.363**	0.332	26.034	27.380	20.452	18.999	**8.501**	8.361
	DiffBindFR_MDN	0.774	0.791	0.187	0.109	0.271	0.234	18.419	10.749	15.766	11.103	7.290	6.560
	DiffBindFR_Smina	0.264	0.262	0.010	0.010	0.005	0.002	0.189	0	0.204	0	0.227	0.148
	DynamicBind	0.670	0.688	0.067	0.032	0.106	0.065	5.684	3.109	5.270	3.414	3.644	2.688
	SurfDock	0.769	0.780	0.231	**0.242**	0.347	**0.393**	**27.148**	**29.259**	**20.747**	**23.335**	7.938	**8.379**
	GAABind	0.623	0.613	0.039	0.023	0.058	0.024	2.524	0	2.428	0.916	2.335	2.033
	KaramDock	0.719	0.698	0.098	0.036	0.154	0.069	10.187	2.865	8.374	2.979	4.716	3.217
	KaramDock_aligned	0.709	0.696	0.096	0.042	0.174	0.112	12.062	7.796	9.311	7.146	4.858	4.235
	KaramDock_FF	0.746	0.721	0.121	0.059	0.191	0.119	13.291	7.892	10.420	5.679	5.652	4.501
	QuickBind	0.561	0.546	0.059	0.036	0.082	0.040	2.858	1.018	2.887	1.258	2.111	1.444

[a] The best result is emphasized by bold formatting, while the second-ranked result is underline.

likely attributable to the limitations of its linear SF in accommodating the structural diversity of binding sites.

Consistent with docking observations, the hybrid method Interformer-Energy excelled, achieving a superior balance (average ROC-AUC: 0.746, 0.802; $EF_{0.5\%}$: 15.008, 26.043; BEDROC: 0.384, 0.363), this performance surpasses Glide SP and significantly outpaces AutoDock Vina, underscoring the advantage of its hybrid paradigm again, which integrates physics-driven conformational exploration with data-driven scoring precision. This synergy retains the strengths of both approaches, enabling robust ranking and early enrichment.

Leveraging generative modeling to capture complex binding distributions, diffusion-based methods, notably SurfDock's surface-guided approach, led in early enrichment (average $EF_{0.5\%}$: 17.064, 27.148), while DiffBindFR-MDN's mixture density network ensured robust ranking (average $EF_{0.5\%}$: 13.251, 18.419), alongside high ROC-AUC(average: 0.770, 0.774, Med.: 0.822, 0.791). However, DiffBindFR-Smina's failure on DUD-E (average $EF_{0.5\%}$: 0.189) exposed a critical vulnerability: generative models rely heavily on precise SFs to translate latent pose distributions into discriminative rankings. Similarly, DynamicBind's consistent underperformance revealing a gap between generative and discriminative objectives.

Regression-based methods, exemplified by GAABind and QuickBind, exhibited poor performance (QuickBind: average ROC-AUC 0.592, 0.561; GAABind: 0.574, 0.623), with near-zero early enrichment (*e.g.*, GAABind average $EF_{0.5\%}$: 1.421, 2.524). In contrast, KarmaDock showed promise on DEKOIS2.0, with KarmaDock FF ranking first (average ROC-AUC: 0.780, $EF_{0.5\%}$: 16.633, BEDROC: 0.464) across most metrics except $EF_{0.5\%}$, but delivered below-average performance on DUD-E (average ROC-AUC: 0.746, $EF_{0.5\%}$: 13.291, BEDROC: 0.191), likely due to its regression-based scoring struggling with DUD-E's diverse binding sites, highlighting both the strengths and limitations of optimized regression approaches.

Performance disparities between datasets underscore the success of Interformer-Energy, SurfDock, and DiffBindFR-MDN in integrating data-driven modeling with physical or geometric constraints. Further analysis of target-specific performance across protein families (Fig. S5–7) revealed substantial variability within the same family, challenging the notion that protein family classification adequately stratifies docking method efficacy. Notably, screening performance on cytochrome P450 and GPCR targets was markedly lower than for other families, a finding consistent with observations by Shen *et al.* in their analysis of SFs[29] and intricately linked to the unique structural and functional properties of these targets. Cytochrome P450 enzymes play a critical role in drug metabolism, featuring binding sites that are sufficiently large and flexible to accommodate a wide array of substrates and inhibitors.[30] In contrast to the deep binding pockets characteristic of many enzymes, G protein-coupled receptors (GPCRs) often

exhibit shallow, exposed, or membrane-embedded binding sites, where ligand–protein interactions tend to be less stable.[31]

The observed decline in docking performance across increasingly generalized datasets (Astex, PoseBusters, DockGen) and the pronounced variability within protein families in VS datasets underscore a critical insight: the generalization capacity of DL methods requires thorough investigation. This trend suggests that current models struggle to extrapolate beyond training distributions, necessitating a deeper understanding of the factors driving this limitation.

Generalization performance analysis

Generalization remains a key bottleneck for the widespread deployment of molecular docking methods,[32,33] as evidenced by the limitations of DL approaches observed in our initial results. Beyond performance on unseen data, generalization reflects a model's adaptability and robustness across diverse protein–ligand contexts.

In this section, we systematically evaluate docking methods across variations in protein sequence, ligand topology and 10 Å binding pocket similarity. We aim to identify the strengths and limitations of DL-based approaches, uncover generalization bottlenecks, and offer guidance for improving model transferability in real-world drug discovery settings.

Generalization docking power performance across protein sequence similarity levels. To further assess the generalization capabilities of the docking methods, we evaluated their performance across the three docking datasets, focusing on protein sequence similarity relative to the DL training data (Fig. 4a and S8–9a). This analysis reveals the impact of sequence divergence on model robustness. Notably, the reduced sample size in the Astex diverse set after similarity stratification (4–16 data points) may introduce uncertainty in the results (Fig. S8a), necessitating cautious interpretation and further validation with larger datasets.

Traditional methods, Glide SP and AutoDock Vina, maintained stable performance across all datasets regardless of sequence similarity. This robustness underscores their physics-based foundations, enabling reliable docking without reliance on protein sequence similarity—a critical advantage for drug discovery targeting novel protein targets. Consistent with prior observations, the hybrid method Interformer maintained

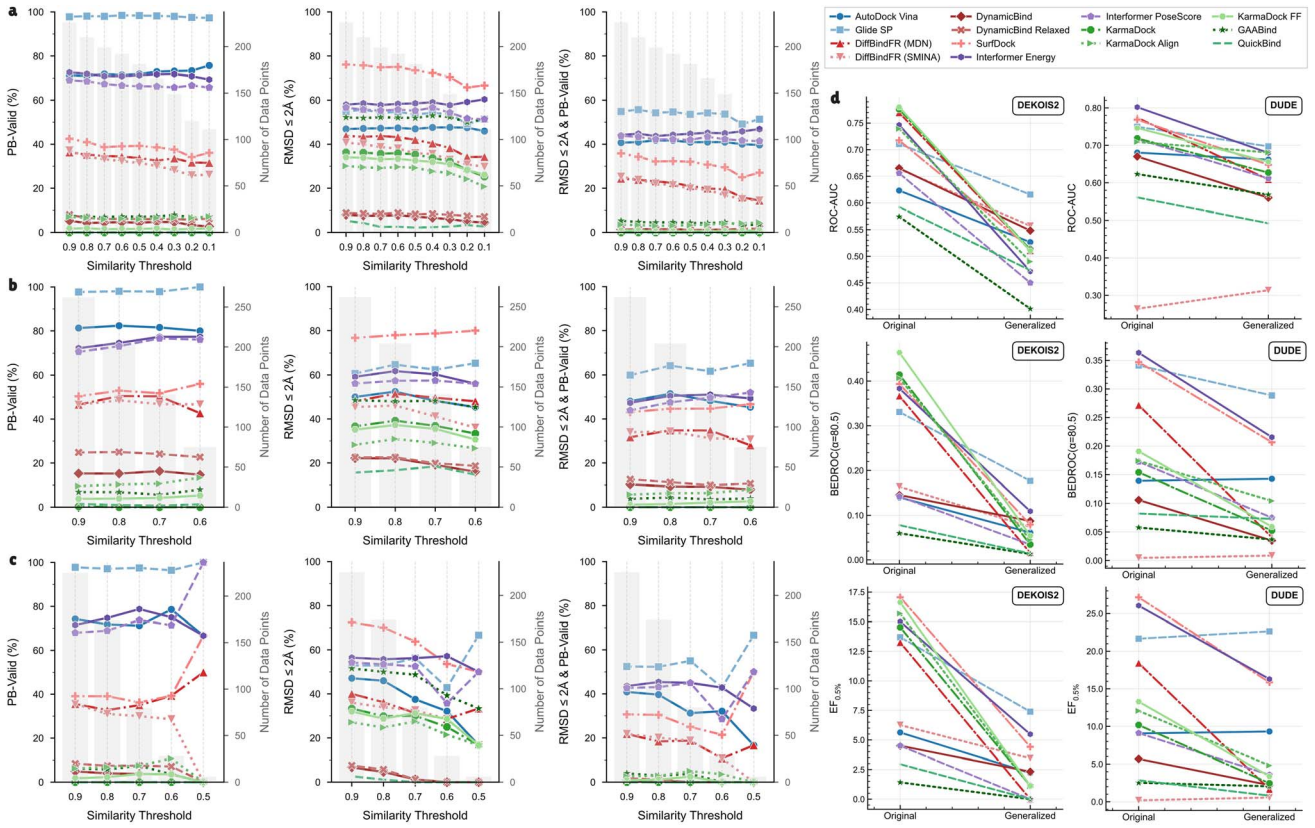

Fig. 4 Generalization performance of docking methods. (a–c) Docking performance metrics as a function of maximum allowed protein sequence similarity (calculated using MMseqs2) (a), ligand similarity (Tanimoto coefficient based on RDKit topological fingerprints) (b) and protein binding pocket similarity (TM-score calculated using USalign on heavy atoms within 10 Å of the ligand) (c) to the PDBBindv2020 training set. Test complexes with higher similarity than the threshold (*x*-axis) were excluded. Performance is evaluated using the PoseBusters benchmark set, with metrics including pose success rate (RMSD ≤ 2 Å), physical validity rate (PB-valid), and combined success rate (RMSD ≤ 2 Å & PB-valid). Light gray bars indicate the number of data points retained after applying the similarity threshold. (d) Impact of binding pocket generalization on virtual screening performance. Comparative analysis of virtual screening performance for docking methods on DEKOIS2.0 and DUD-E before and after generalization, including ROC-AUC, BEDROC($\alpha = 80.5$) and $EF_{0.5\%}$.

balanced performance, demonstrating stability across diverse similarity thresholds, reinforcing its resilience and adaptability to diverse protein sequences, leveraging the synergy of physics-driven conformational sampling and data-driven scoring. In contrast, both diffusion- and regression-based DL methods showed performance declines with decreasing sequence similarity, indicating a dependency on training protein sequence, highlighting a generalization gap compared to traditional methods, this trend, consistent with observations by Buttenschoen *et al.*,[18] suggests a reliance on training set sequence homology, exposing a generalization gap relative to traditional methods. This Interformer's dual approach mitigates the sequence dependency observed in purely DL-based methods, reinforcing its potential as a robust framework for generalizable docking.

A striking pattern emerges in the rate of decline: the drop in performance was more rapid on PoseBusters (unseen complexes) (Fig. 4a) than on DockGen (novel binding pockets) (Fig. S9a). For example, SurfDock combined success decreased from 35.84% to 25.00% across similarity levels on PoseBusters, compared to a milder reduction from 36.36% to 34.29% on DockGen. This differential decline may be attributed to the distinct design and complexity of the datasets, compounded by the effects of similarity-based filtering. DockGen, curated to explore novel protein binding pockets, likely encompasses a higher intrinsic difficulty due to its focus on uncharted structural and chemical spaces, which may demand greater generative and adaptive capacity from DL models. In contrast, PoseBusters, designed to challenge models with unseen complexes, emphasizes time out-of-distribution scenarios. Additionally, the extent of data exclusion *via* similarity filtering may play a role: PoseBusters, with a larger initial sample (111–226 data points), experiences a more substantial reduction in usable data at higher similarity thresholds (*e.g.*, 111 at 0.1 to 226 at 0.9), potentially skewing the remaining subset toward more challenging cases. DockGen, with a more moderate range (140–165), retains a relatively stable sample size, possibly preserving a more representative distribution of pocket complexities. This sampling effect, combined with DockGen's novel pocket focus, may mitigate the severity of the generalization gap observed on PoseBusters.

These results highlight that DL-based docking methods require enhancement strategies—such as broader training datasets, physics-informed loss functions, or hybrid frameworks.

Generalization docking power performance across ligand similarity levels. Previous studies have suggested that overfitting to ligand structures may impair the generalization of DL-based SFs.[34] To investigate this, we analyzed docking performance in relation to ligand similarity (Fig. 4b and S8–9b). The analysis of protein sequence similarity previously indicated that the PoseBusters benchmark set (Fig. 4b) offers a more authentic representation of out-of-distribution (OOD) scenarios than DockGen (Fig. S9b) for our generalization performance analysis method, as its dataset construction deliberately avoids selective removal of data based on specific similarity metrics.

Consistent with findings on protein sequence similarity, traditional methods demonstrated robust performance across the Astex diverse set (Fig. S8b) and PoseBusters benchmark set. However, performance fluctuations on the DockGen underscores the compounded challenges posed by novel ligands and binding pockets. The hybrid model Interformer again showed stable performance across similarity levels, further validates the efficacy of combining DL with physicochemical principles, offering a balanced approach to handle OOD scenarios.

Diffusion-based methods displayed mixed behavior. SurfDock showed declining performance with decreasing ligand similarity on Astex, but surprisingly improved on PoseBusters and DockGen, suggesting resilience to ligand novelty in more complex scenarios. Other diffusion-based and all regression-based DL methods exhibited decreasing performance on Astex and PoseBusters, but remained stable—or even improved slightly—on DockGen, likely implying that unfamiliar pockets, rather than ligands, pose the greater generalization barrier.

This section unveils several noteworthy insights into the generalization capabilities of docking methods across varying levels of ligand similarity. Notably, the robust performance of traditional methods, such as Glide SP, and the hybrid model Interformer-Energy underscores their reliability in navigating diverse chemical spaces, leveraging physics-based principles to maintain accuracy. The exceptional performance of SurfDock further highlights the potential of diffusion-based approaches in addressing complex ligand topology OOD scenarios, demonstrating adaptability to novel ligand environments. Intriguingly, the anomalous stability or enhanced performance of DL methods on the DockGen dataset suggests that unfamiliar binding pockets, rather than ligand dissimilarity, may constitute the primary generalization bottleneck—a finding that merits in-depth investigation. Collectively, these results emphasize the imperative for tailored training strategies and physics-guided methodologies to surmount current limitations, thereby laying a foundation for more adaptable docking solutions in the advancement of drug discovery.

Generalization docking power performance across 10 Å protein pocket similarity levels. Given the significant impact of protein binding pockets on DL docking performance, this section assesses model docking performance relative to 10 Å binding pocket similarity (Fig. 4c, and S8–9c).

Traditional methods showed variable robustness. Glide SP maintains stable performance on the Astex (Fig. S8c) and PoseBusters datasets (Fig. 4c), but its RMSD $\leq$ 2 Å success rate declined on DockGen (Fig. S9c) as pocket similarity decreased. AutoDock Vina displayed a consistent performance decline across all three datasets with decreasing similarity, revealing limitations of physics-based methods in addressing diverse binding environments.

The hybrid model Interformer-Energy exhibited mixed trends in RMSD $\leq$ 2 Å success rate (declining on Astex, stable on PoseBusters, and increasing on DockGen). Overall, its comprehensive metrics remained robust, outperforming traditional methods and underscoring the potential of integrating AI-driven scoring with traditional conformation searches. In contrast, Interformer-PoseScore's performance across all

metrics declined with decreasing similarity on all datasets, suggesting that rescoring with AI-based SFs is less effective for enhancing the generalization of binding pocket compared to coupled scoring approaches.

Diffusion-based methods showed a gradual decline in RMSD $\leq$ 2 Å success rate as pocket similarity decreased, though PB-valid scores remained relatively stable, indicating divergence between structural accuracy and physical plausibility. Regression-based methods, particularly on PoseBusters, showed pronounced sensitivity to pocket similarity. Interestingly, the KarmaDock series exhibited improved RMSD success on DockGen as similarity decreased,may be attributed to a combination of small sample effects and the series' inherently low overall performance.

These findings underscore the significant challenges deep learning methods faced with novel binding pockets, revealing a critical issue of overfitting to training pocket features. This overfitting severely hampers generalization to unseen binding environments, highlighting the need for innovative training paradigms and hybrid approaches, exemplified by Interformer-Energy, to enhance docking performance in diverse binding environments.

Generalization virtual screening power performance across 10 Å protein pocket similarity levels. Our prior observations have demonstrated that the generalization docking performance of DL methods is profoundly influenced by the structural characteristics of protein binding pockets, a finding reinforced by the substantial performance variability observed among different DL methods, even within the same protein family across diverse targets. Building upon these insights, this section extends the evaluation to the VS capabilities of docking models, with a specific focus on their transferability to novel targets under realistic VS scenarios. As is well established, structure dictates function. To this end, we assessed model VS performance on filtered subsets of DEKOIS2.0 and DUD-E, where all targets share <0.8 binding pocket similarity to the training set (Table 2, Fig. 3d, e and S10–12).

Across both datasets, Glide SP emerged as the preeminent performer, exhibiting the highest VS efficacy as measured by all key metrics, a testament to its physics-based robustness in navigating novel binding pockets. On the filtered DEKOIS2.0 traditional methods exhibit relative robustness with performance declines of approximately 15% in ROC-AUC, though performance declined compared to the full dataset, in stark contrast to the substantial drop observed for DL-based methods with performance declines more than 25% in ROC-AUC (Fig. 3d and S10–12a). The limited target set ($n = 4$) may exacerbate this variability, potentially amplifying noise in the observed trends. On the larger filtered DUD-E dataset, traditional methods maintained robustness, with Glide SP leading across all metrics, while DL-based methods continued to show significant performance declines, with BEDROC and $EF_{0.5\%}$ drops exceeding 35% in and ROC-AUC decreasing by over 12% (Fig. 3e and S10–12b). These results underscore the considerable challenges DL docking methods face in generalizing novel protein binding pockets. Among DL approaches, the hybrid method Interformer-Energy and diffusion-based SurfDock showed the greatest potential, despite notably reduced performance relative to the full datasets. Previously high-performing regression models, such as KarmaDock, struggled significantly, while QuickBind, GAABind, and DiffBindFR variants (particularly Smina) consistently underperformed.

These findings reaffirm the need for DL models to enhance robustness to unseen binding sites, a pivotal factor for effective lead discovery. The resilience of Glide SP, coupled with the partial adaptability of Interformer-Energy and SurfDock, suggests that integrating physicochemical constraints or enhanced sampling strategies could help alleviate this gap, providing valuable guidance for tool selection in real-world applications.

Method-specific insights and optimization strategies

In solution, ligands do not adopt a single, fixed structure but exist as an ensemble of conformations, the distribution of which is governed by thermodynamic principles. Notably, the active conformation—the one that binds effectively to the target—is often not the ligand's lowest-energy state in isolation but rather the conformation that minimizes the free energy of the entire ligand–target complex.[35,36] Traditional docking methods rely on SFs that assume a linear relationship between binding free energy and conformation to guide conformational search algorithms toward the active ligand pose (Fig. 5a). In contrast, Interformer leverages the robust nonlinear fitting capabilities of AI-based SFs to direct these searches more efficiently, pinpointing the active conformation with greater precision (Fig. 5b). Regression-based methods, which minimize expected squared error, tend to predict a single conformation (Fig. 5c)—often approximating a (weighted) average of feasible binding poses—whereas diffusion-based generative models aim to capture the full distribution of viable binding conformations, encompassing the most critical binding modes (Fig. 5d).

Detailed analysis of docking performance reveals that DL-based methods are significantly influenced by the PB-valid metric, which comprises three components: chemical validity and consistency, intramolecular validity, and intermolecular validity. Our analysis elucidates the factors driving performance across these dimensions. Diffusion-based methods achieve chemical and intramolecular validity levels comparable to traditional algorithms, reflecting their alignment with physical constraints of ligand during sampling, while regression-based methods show no distinct advantage in this regard. Moreover, in terms of interaction recovery, diffusion-based methods SurfDock perform on par with traditional approaches in recovering interactions observed in crystal structures, representing a notable advancement in addressing the physical and chemical implausibility issues that afflict regression-based methods.

From a modeling perspective, diffusion-based methods align with traditional search algorithms by sampling ligand translations, rotations, and internal torsions, preserving inherent bond length and angle constraints, thereby ensuring consistent chemical and intramolecular validity. However, current diffusion models in molecular docking exhibit certain limitations.

Table 2 Generalization virtual screening performance of docking models on the DEKOIS2.0 (4 targets) and DUD-E (8 targets)[a]

		ROC-AUC		PRC-AUC		BEDROC ($\alpha = 80.5$)		$EF_{0.5\%}$		$EF_{1\%}$		$EF_{5\%}$	
Dataset	Model	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.
DEKOIS 2.0 (4)	AutoDock Vina	0.527	0.524	0.052	0.036	0.061	0.036	2.213	0	1.788	1.192	1.374	1.250
	Glide SP	**0.616**	**0.617**	**0.106**	**0.110**	**0.177**	**0.173**	**7.391**	**4.701**	**4.779**	**3.958**	**3.507**	**2.923**
	Interformer	0.450	0.442	0.033	0.032	0.034	0.033	0.000	0	1.771	1.179	1.238	1.241
	Interformer_Energy	0.472	0.490	0.067	0.032	0.109	0.040	5.479	2.202	3.540	1.186	1.733	0.745
	DiffBindFR_MDN	0.511	0.499	0.039	0.034	0.016	0.014	0.000	0	0.000	0	0.774	0.525
	DiffBindFR_Smina	0.557	0.564	0.045	0.045	0.074	0.078	3.449	4.382	2.399	2.436	1.408	1.298
	DynamicBind	0.548	0.506	0.057	0.051	0.087	0.087	2.323	2.191	2.994	2.307	1.486	1.489
	SurfDock	0.515	0.552	0.049	0.042	0.079	0.060	4.415	2.209	2.972	1.189	1.889	1.536
	GAABind	0.401	0.412	0.025	0.025	0.013	0.007	0.000	0	0.640	0	0.259	0.250
	KaramDock	0.513	0.482	0.044	0.041	0.035	0.019	1.128	0	0.564	0	0.988	0.994
	KaramDock_aligned	0.490	0.488	0.039	0.040	0.045	0.033	1.092	0	1.176	0	1.482	0.994
	KaramDock_FF	0.511	0.484	0.040	0.042	0.054	0.029	1.128	0	1.717	1.179	0.863	0.744
	QuickBind	0.474	0.527	0.035	0.035	0.014	0.013	0.000	0	0.000	0	0.625	0.500
DUD-E (8)	AutoDock Vina	0.662	0.667	0.074	0.039	0.143	0.080	9.321	4.505	7.981	3.388	3.899	3.625
	Glide SP	**0.697**	**0.701**	**0.192**	**0.099**	**0.289**	**0.209**	**22.644**	**16.553**	**16.616**	**12.460**	**5.874**	**5.440**
	Interformer	0.612	0.612	0.034	0.028	0.075	0.067	3.580	3.369	4.259	3.729	2.679	2.506
	Interformer_Energy	0.680	0.660	0.121	0.054	0.215	0.162	16.295	13.097	12.763	9.911	5.011	4.310
	DiffBindFR_MDN	0.609	0.596	0.029	0.024	0.041	0.031	1.654	1.100	1.591	0.996	1.874	1.732
	DiffBindFR_Smina	0.314	0.298	0.012	0.012	0.009	0.003	0.564	0	0.465	0	0.261	0.190
	DynamicBind	0.560	0.562	0.025	0.020	0.035	0.012	2.212	0	1.560	0	1.253	0.534
	SurfDock	0.648	0.622	0.104	0.051	0.207	0.143	15.862	10.241	12.312	8.145	4.646	3.686
	GAABind	0.568	0.577	0.024	0.022	0.037	0.028	2.058	1.170	1.810	1.328	1.375	1.181
	KaramDock	0.627	0.611	0.030	0.028	0.052	0.034	2.454	1.399	2.729	1.892	1.846	1.442
	KaramDock_aligned	0.681	0.672	0.052	0.044	0.104	0.088	4.820	4.004	5.060	3.418	3.965	3.779
	KaramDock_FF	0.656	0.612	0.036	0.035	0.059	0.036	3.381	0.895	3.138	2.021	2.297	2.357
	QuickBind	0.492	0.466	0.053	0.025	0.073	0.015	0.809	0	1.790	0.500	1.763	0.932

[a] The best result is emphasized by bold formatting, while the second-ranked result is underlined.

Typically, these models first sample ligand conformations *via* a diffusion process and then rank them using a separate SF. This decoupled approach contrasts with traditional methods, where SFs actively guide the conformational search in real time, potentially compromising the intermolecular reasonableness of the generated poses. In traditional and hybrid methods like Interformer-Energy, SFs dynamically steer the search process and impose penalties for steric clashes, ensuring realistic binding interactions—a factor validated by Interformer-Energy's superior performance, outstripping the rescoring-only Interformer-PoseScore. The two-step nature of current diffusion models may thus result in conformations that lack optimal intermolecular validity.

The poor chemical and physical validity of ligands generated by regression-based methods likely arises from their direct prediction of atomic coordinates in 3D space—a challenging task—or from predicting ligand–protein distance matrices, combined with an overreliance on RMSD as a loss function, which overlooks critical physical constraints. Notably, the distance-matrix-based method GAABind demonstrates superior interaction recovery compared to direct coordinate regression approaches. This suggests that distance-matrix predictions may better capture both long–range interactions and short-range geometric constraints.[38] However, GAABind's reliance on a subsequent geometric reconstruction step to derive ligand conformations from distance matrices introduces additional computational overhead and risks geometric errors, undermining its efficiency compared to end-to-end methods.

In terms of computational efficiency, regression-based methods outperform diffusion models, which in turn surpass traditional search methods, whether augmented with AI SFs or not.[39] Consequently, regression-based methods, with their rapid processing and moderate performance, are well-suited for the initial coarse screening in ultra-large-scale VS campaigns, as demonstrated by QuickBind and supported by Gu *et al.* studies.[23] In large-scale VS, these methods can swiftly identify potential active compounds from vast chemical libraries, providing a foundation for subsequent refined screening and experimental validation.

A critical limitation across most methodologies is the inadequate incorporation of protein flexibility. Proteins are not static but exhibit intrinsic flexibility, undergoing conformational changes upon ligand binding *via* mechanisms such as induced fit (ligand-driven receptor adaptation) or conformational selection (preferential binding to low-population conformers).[40] This flexibility ranges from local side-chain rotations to optimize interactions, loop rearrangements to modulate pocket accessibility, to large-scale domain shifts that redefine binding interfaces—crucial for affinity, specificity, and entropy-enthalpy balance in flexible targets like GPCRs or intrinsically disordered proteins.[12,41] Neglecting these dynamics introduces systematic errors in pose prediction and affinity

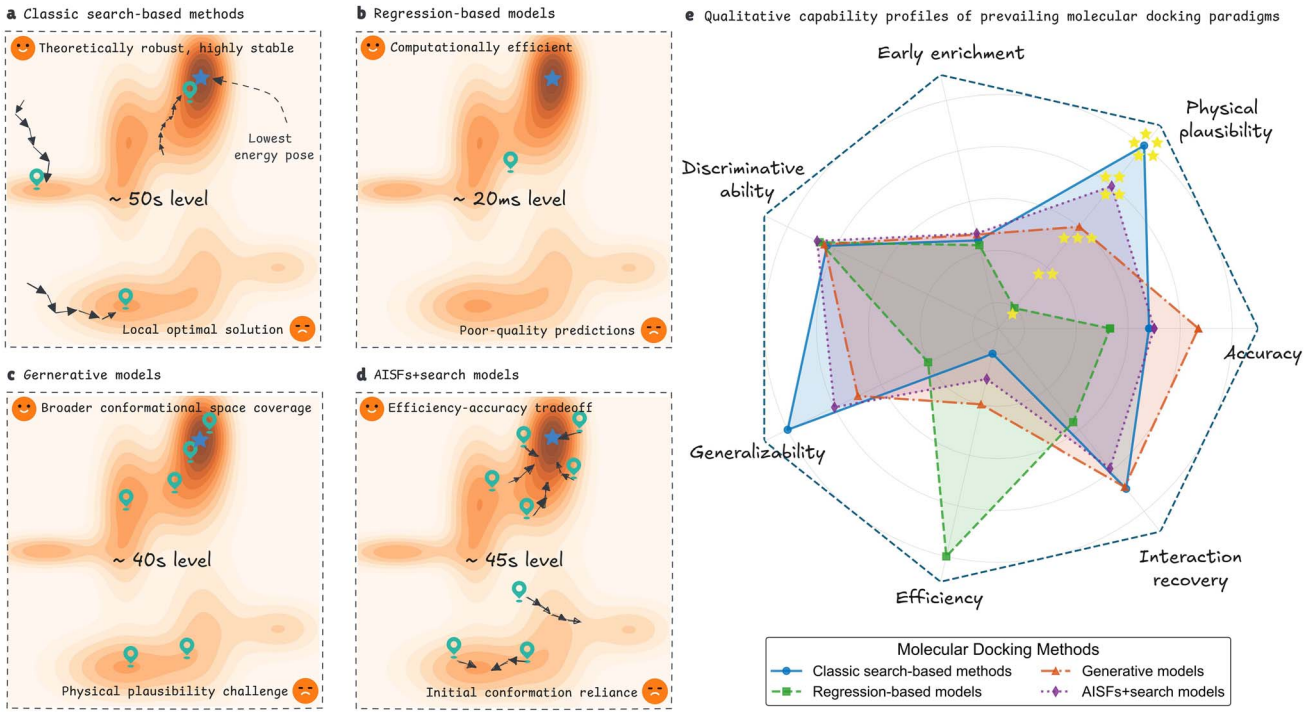

Fig. 5 Conceptual comparison and qualitative performance summary of docking paradigms. (a–d) Two-dimensional projections of conformational density distributions for ligand–protein binding(ref DiffDock[37]). (a) Classical search-based methods, (b) regression-based methods, (c) generative methods, and (d) classical search with AI-driven scoring function methods (e) Radar chart qualitatively evaluating the performance of representative methods from the four paradigms (Glide SP, SurfDock, Interformer-Energy and KarmaDock) using seven metrics: early enrichment, physical plausibility, accuracy, interaction recovery, discrimination ability, generalizability, and efficiency. Scores are illustrative, ranging from low (center) to high (periphery).

estimation,[42,43] particularly for malleable binding pockets, as evidenced by the poor enrichment for GPCRs and cytochrome P450 enzymes in VS evaluations. Among evaluated methods, only DynamicBind and DiffBindFR explicitly address protein flexibility. DiffBindFR employs diffusion to refine side-chain orientations for local adjustments, enhancing interaction fidelity. In contrast, DynamicBind leverages equivariant geometric diffusion to predict ligand-specific backbone and domain motions, enabling the capture of cryptic pockets in apo or unbound structures. Other methods, such as SurfDock, implicitly account for flexibility *via* surface-informed features but may fall short in scenarios requiring extensive backbone remodeling.

Finally, we summarize the performance of representative methods from four docking categories across seven key metrics: discriminative ability, early enrichment, physical plausibility, accuracy, interaction recovery, efficiency, and generalizability (Fig. 5e).

To enhance the performance of diffusion models in molecular docking, future research should focus on refining confidence modules and integrating more advanced, precise SFs to guide sampling toward more efficient and realistic outcomes. Traditional search methods augmented with AI-based or classical SFs could benefit from cutting-edge diffusion-based sampling techniques and high-precision AI-physics hybrid SFs, leveraging GPU architectures for efficient conformational searches and accurate affinity predictions. For regression-based methods, incorporating physical constraints and predicting ligand translations, rotations, and internal torsion angles could enhance the physical plausibility of the predicted poses. Across all methods, explicit joint modeling of ligand and protein flexibility or implicit incorporation *via* coarse-grained priors promises to elevate docking fidelity. These advancements will bolster the role of DL in drug discovery, providing robust support for the evolution of molecular docking technologies.

Conclusions

This study provides a comprehensive, multidimensional evaluation of molecular docking methods, spanning traditional physics-based approaches (Glide SP, AutoDock Vina), DL-based strategies—including generative diffusion models (SurfDock, DiffBindFR, DynamicBind), regression-based models (QuickBind, GAABind, KarmaDock), and hybrid methods (Interformer)—and their performance across diverse benchmark datasets. By assessing binding pose prediction, physical validity, interaction recovery, VS efficacy, and generalization across protein sequence, ligand topology, and binding pocket similarity, our findings elucidate the strengths, limitations, and practical utility of these approaches in accelerating lead discovery and optimization.

The strengths of this study lie in its rigorous, multidimensional approach, evaluating docking methods across diverse datasets and metrics that reflect real-world drug discovery

needs. By considering not only pose prediction accuracy but also physical validity, VS performance, and generalization, we provide a holistic assessment that bridges theoretical insights with practical applications. However, limitations must be acknowledged. The reliance on specific benchmark datasets (*e.g.*, Astex, PoseBusters, DockGen) may not fully capture the complexity of all drug discovery scenarios, particularly those involving non-small-molecule modalities like peptides. Additionally, while our VS analysis on DEKOIS2.0 and DUD-E provides valuable insights, the reduced protein sets used to test generalization may amplify variability, warranting further validation with larger, more diverse datasets.

Future research should prioritize several directions to advance molecular docking technologies. First, enhancing the physical plausibility of DL-based methods—through integrated scoring in diffusion models or physics-informed regression frameworks—while modeling protein flexibility, could bridge the gap between efficiency and accuracy. Second, expanding training datasets to encompass greater diversity in protein sequences, binding pockets, and ligand chemistries may improve DL generalization, reducing overfitting and enhancing adaptability. Third, prospective studies applying these methods to real-world VS campaigns or lead optimization efforts would validate their practical utility and guide further refinement. Finally, extending evaluations to emerging modalities, such as biologics or protein–protein interactions, could broaden the applicability of these tools in modern drug discovery.

In conclusion, this study underscores the evolving landscape of molecular docking, where traditional reliability meets DL-driven innovation. While DL methods offer transformative potential in speed and pattern recognition, their current limitations in generalization and physical validity highlight the need for hybrid strategies that synergize data-driven and mechanistic approaches. By addressing these challenges, the field can develop robust, versatile docking tools that enhance the efficiency and success of drug discovery, paving the way for the next generation of therapeutic breakthroughs.

Materials and methods

Dataset

Astex diverse set. Introduced in 2007, a meticulously curated collection of 85 high-quality protein–ligand complexes from the Protein Data Bank (PDB)[44] forms the Astex diverse set. This diverse assembly, featuring drug-like ligands and proteins vital to pharmaceutical and agrochemical sectors, underpins computational drug discovery, especially molecular docking.

PoseBusters benchmark set. The PoseBusters benchmark set is a meticulously curated collection of 428 high-quality, publicly available protein–ligand crystal complexes sourced from the PDB. This diverse dataset exclusively includes complexes released since 2021, ensuring no overlap with the PDBbind General Set v2020.[45] Each complex features unique proteins and drug-like ligands, making it ideal for evaluating molecular docking and related methods.

DockGen. The DockGen dataset, introduced by Corso *et al.* in 2024,[26] is a challenging benchmark for molecular docking, featuring 189 diverse single-ligand protein–ligand complexes with unique binding pockets not found in PDBBindv2020 before 2019. Derived from the Binding MOAD database[46] and filtered using ECOD[47] classification, it excludes complexes with multiple ligands, metals, or large molecules (>60 heavy atoms), ensuring chemical diversity. With 189 complexes, DockGen tests generalization to novel binding sites, making it a rigorous single-ligand benchmark for molecular docking. As the original ligand data in PDB format lacked bond information, we processed the complex structures using Schrödinger 2024's Protein Preparation Wizard[48] module.

DEKOIS2.0. Serving as a vital benchmark dataset for VS, DEKOIS 2.0 features 81 diverse protein targets, each with 40 active compounds and 1200 decoys (30 decoys per active). Built from BindingDB bioactivity data,[49] it enhances the original DEKOIS methodology by improving physicochemical matching-now incorporating molecular weight, log *P*, hydrogen bond donors/acceptors, rotatable bonds, charged states, and aromatic rings—and eliminating latent actives in the decoy set (LADS) to reduce bias. Decoys are selected for low fingerprint similarity to actives, ensuring robust evaluation.

DUD-E. DUD-E stands as a cornerstone benchmark for molecular docking, encompassing 22 886 active ligands across 102 diverse protein targets, including GPCRs and ion channels. Each active, sourced from ChEMBL[50] and clustered by Bemis–Murcko frameworks,[51] is paired with 50 topologically dissimilar, property-matched decoys from ZINC.[52] Matching properties include molecular weight, log *P*, hydrogen bond donors/acceptors, rotatable bonds, and net charge. Due to the significant computational resources required for assessing all methods across the entire dataset, we focus on 26 representative targets from distinct protein families, as listed in the original DUD-E publication, which still demands substantial computational effort. In subsequent analyses of virtual screening generalization, we maximized the number of generalized targets by analyzing pocket similarity with the DL training set across the entire DUD-E dataset, identifying 8 targets for further generalization performance evaluation.

Docking method

We conducted a comprehensive evaluation of nine molecular docking methods, encompassing two traditional physics-based approaches (Glide SP[8] and AutoDock Vina[9]) and seven DL-based methods. The DL methods included three generative diffusion models (DynamicBind,[12] DiffBindFR,[16] and SurfDock[13]), three regression-based models (QuickBind,[24] GAABind,[11] and KarmaDock[14]), and a hybrid approach (Interformer[15]) that integrates traditional conformational search with an AI-driven scoring function. Additionally, we assessed variants of the DL methods: DiffBindFR-Smina and DiffBindFR-MDN (selecting top-ranked poses using Smina and a Mixture Density Network, respectively), KarmaDock FF and KarmaDock aligned (optimizing ligand conformations with the MMFF94 force field and aligned RDKit ligand conformations), and Interformer-Energy and Interformer-PoseScore (selecting top-ranked poses using energy functions and pose scoring, respectively). The selection

criterion for these deep learning methods was their training on the PDBBindv2020 dataset. Additionally, beyond predicting molecular binding conformations, these methods are capable of estimating the binding affinity of ligand molecules. All docking approaches were implemented following their official guidelines using default parameters, retaining 40 docking poses per method, as detailed in SI Section 1.1 (Docking protocols).

Interaction recovery evaluation

Protein–ligand interaction detection *via* Protein–Ligand Interaction Fingerprint (PLIF) libraries[53] is sensitive to protonation states, which dictate whether interactions are ionic or hydrogen bonds. Classical docking methods model hydrogens explicitly but often infer interactions from heavy-atom geometries, while deep learning methods typically use only heavy atoms. To ensure equitable evaluation, we processed top1 ranked poses and protein structures using the Protein Preparation Wizard[48] in Schrödinger 2024, assigning protonation states, adding explicit hydrogens, optimizing hydrogen-bond networks, and minimizing systems with the OPLS3e force field, constraining heavy atoms. This optimizes hydrogen-bond networks for accurate interaction detection.

PLIFs were computed using ProLIF (v2.0.3), focusing on hydrogen and halogen bonds, π-stacking, cation–π, π–cation, and ionic interactions, excluding non-specific hydrophobic and van der Waals contacts. Custom distance thresholds were set at 3.7 Å (hydrogen bonds), 5.5 Å (cation–π), and 5.0 Å (ionic interactions), with other parameters at ProLIF defaults. The PLIF computation was performed using a modified version of the *plif_analysis.ipynb* script provided by Dreyer *et al.*[19].

Similarity-based generalization analysis

To assess the generalization performance of our models, we quantified the similarity between the test set and the PDBbind v2020 training dataset using multiple metrics, and the data in the test set with similarity above the threshold were excluded, implemented as follows:

(1) Ligand similarity: we computed ligand similarity using RDKit topological fingerprints, which comprehensively encode molecular structural features such as atom connectivity, bond types, and ring systems. These fingerprints, generated *via* the Morgan algorithm, enable a detailed Tanimoto coefficient-based comparison, capturing subtle differences in chemical scaffolds and functional groups to identify potential overlaps with training data.

(2) Protein sequence similarity: protein sequence similarity was evaluated with MMseqs2,[54] a high-performance tool optimized for large-scale sequence analysis. This method employs a sensitive *k*-mer-based indexing and iterative alignment strategy, offering rapid yet precise similarity scores (*e.g.*, *via* BLAST-like bit scores) across protein sequences. By detecting evolutionary relationships and conserved domains, it effectively flags test set proteins closely related to those in the PDBbind v2020 training set.

(3) Protein pocket similarity: binding pocket similarity was measured using USalign,[55] computing the Template Modeling Score (TM-score) between heavy atoms within 10 Å of the ligand in each test set protein and PDBbind v2020 protein. The TM-score, ranging from 0 to 1 (1 indicating identical structures), quantifies structural similarity. Higher TM-scores suggest greater similarity to the closest training complex, potentially highlighting training bias in test performance.

These metrics collectively enable a robust evaluation of model generalization by identifying potential overlaps between training and test datasets.

Evaluation metrics

Docking power. Performance was quantified using three metrics: (1) the percentage of top1 ranked ligand conformations with RMSD $\leq$ 2 Å relative to the crystal structure (RMSD $\leq$ 2 Å), (2) physical validity as determined by the PoseBusters package (PB-valid), and (3) their intersection (RMSD $\leq$ 2 Å & PB-valid), reflecting overall docking success.

Virtual screening power. ROC-AUC quantifies the overall discriminative power of docking method to distinguish active compounds from inactive ones across all possible ranking thresholds. It represents the probability that a randomly selected active compound is ranked higher than a randomly selected inactive compound. A ROC-AUC value of 1.0 indicates perfect discrimination, while 0.5 reflects random performance.

PRC-AUC measures the balance between precision (the fraction of predicted actives that are true positives) and recall (the fraction of true actives correctly identified) across ranking thresholds. This metric is particularly informative in imbalanced datasets, where active compounds are significantly outnumbered by inactives, providing insight into docking method's ability to maintain high precision while recovering true actives.

BEDROC evaluates early recognition performance by emphasizing the ranking of active compounds at the top of the list. Using an exponential weighting function, BEDROC assigns higher importance to early-ranked actives, with the parameter α set to 80.5 (corresponding to 80% of the score concentrated in the top 2% of the ranked list). BEDROC ranges from 0 to 1, with higher values indicating superior early enrichment.

Enrichment factors assess the ability to prioritize active compounds within the top 0.5%, 1%, and 5% of the ranked list relative to random selection. EF is calculated as the ratio of the fraction of actives in the specified top percentage to the fraction of actives in the entire dataset. $EF_{0.5\%}$ evaluates performance at a highly stringent cutoff, $EF_{1\%}$ at a slightly broader threshold, and $EF_{5\%}$ at a more inclusive cutoff, with higher values indicating stronger enrichment of actives in the early ranks.

Author contributions

Y. Li and J. Yi designed the study, performed experiments and drafted the manuscript; H. Li and K. Li assisted with modeling and analysis; F. Kang, and Y. Deng handled data processing; C. Wu and X. Fu, provided technical support; D. Jiang and D. Cao supervised the project, provided funding, and were responsible for project administration; Y. Li, J. Yi, D. Jiang and D. Cao revised the manuscript.

Conflicts of interest

There are no conflicts to declare.

Data availability

All datasets used in this study are publicly available. Benchmark datasets for evaluating docking performance, including Astex diverse set, Posebusters benchmark set and DockGen dataset, are accessible at **https://zenodo.org/record/8278563**, **https://zenodo.org/records/10656052**. Benchmark datasets for virtual screening, including DEKOIS2.0 and DUD-E, are accessible at **https://zenodo.org/records/8131256** and **https://dude.docking.org**. The PDBbind dataset is available at **https://www.pdbbind.org.cn**. Additional result files and processed data used in this study, including benchmark datasets processed by Schrödinger, are available on our GitHub repository at **https://github.com/liyue9129/DeepDockingData**. All evaluated molecular docking methods were implemented using publicly available source code from their official repositories.

Supplementary information is available. See DOI: **https://doi.org/10.1039/d5sc05395a**.

Acknowledgements

This work was supported by National Natural Science Foundation of China (22173118, 22307112, 82304316), Young Scientists Fund of Natural Science Foundation of Hunan Province of China (2025JJ60651), Noncommunicable Chronic Diseases-National Science and Technology Major Project [2023ZD0507104]. We acknowledge Haikun Xu, and the High-Performance Computing Center of Central South University for support.

Notes and references

1 J. P. Hughes, S. Rees, S. B. Kalindjian and K. L. Philpott, Principles of early drug discovery, *Br. J. Pharmacol.*, 2011, **162**, 1239–1249.
2 S. Simoens and I. Huys, R&D Costs of New Medicines: A Landscape Analysis, *Front. Biomed.*, 2021, **8**, 760762.
3 M. Schlander, K. Hernandez-Villafuerte, C.-Y. Cheng, J. Mestre-Ferrandiz and M. Baumann, How Much Does It Cost to Research and Develop a New Drug? A Systematic Review and Assessment, *PharmacoEconomics*, 2021, **39**, 1243–1269.
4 J. M. Paggi, A. Pandit and R. O. Dror, The Art and Science of Molecular Docking, *Annu. Rev. Biochem.*, 2024, **93**, 389–410.
5 A. V. Sadybekov and V. Katritch, Computational approaches streamlining drug discovery, *Nature*, 2023, **616**, 673–685.
6 D. Jiang, C.-Y. Hsieh, Z. Wu, Y. Kang, J. Wang, E. Wang, B. Liao, C. Shen, L. Xu, J. Wu, D. Cao and T. Hou, InteractionGraphNet: A Novel and Efficient Deep Graph Representation Learning Framework for Accurate Protein–Ligand Interaction Predictions, *J. Med. Chem.*, 2021, **64**, 18209–18232.
7 D. Jiang, Z. Ye, C.-Y. Hsieh, Z. Yang, X. Zhang, Y. Kang, H. Du, Z. Wu, J. Wang, Y. Zeng, H. Zhang, X. Wang, M. Wang, X. Yao, S. Zhang, J. Wu and T. Hou, MetalProGNet: a structure-based deep graph model for metalloprotein–ligand interaction predictions, *Chem. Sci.*, 2023, **14**, 2054–2069.
8 R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley and J. K. Perry, Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy, *J. Med. Chem.*, 2004, **47**, 1739–1749.
9 J. Eberhardt, D. Santos-Martins, A. F. Tillack and S. Forli, AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings, *J. Chem. Inf. Model.*, 2021, **61**, 3891–3898.
10 J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis, Highly accurate protein structure prediction with AlphaFold, *Nature*, 2021, **596**, 583–589.
11 H. Tan, Z. Wang and G. Hu, GAABind: a geometry-aware attention-based network for accurate protein–ligand binding pose and binding affinity prediction, *Briefings Bioinf.*, 2024, **25**, bbad462.
12 W. Lu, J. Zhang, W. Huang, Z. Zhang, X. Jia, Z. Wang, L. Shi, C. Li, P. G. Wolynes and S. Zheng, DynamicBind: predicting ligand-specific protein-ligand complex structure with a deep equivariant generative model, *Nat. Commun.*, 2024, **15**, 1071.
13 D. Cao, M. Chen, R. Zhang, Z. Wang, M. Huang, J. Yu, X. Jiang, Z. Fan, W. Zhang, H. Zhou, X. Li, Z. Fu, S. Zhang and M. Zheng, SurfDock is a surface-informed diffusion generative model for reliable and accurate protein–ligand complex prediction, *Nat. Methods*, 2024, **22**, 310–322.
14 X. Zhang, O. Zhang, C. Shen, W. Qu, S. Chen, H. Cao, Y. Kang, Z. Wang, E. Wang, J. Zhang, Y. Deng, F. Liu, T. Wang, H. Du, L. Wang, P. Pan, G. Chen, C.-Y. Hsieh and T. Hou, Efficient and accurate large library ligand docking with KarmaDock, *Nat. Comput. Sci.*, 2023, **3**, 789–804.
15 H. Lai, L. Wang, R. Qian, J. Huang, P. Zhou, G. Ye, F. Wu, F. Wu, X. Zeng and W. Liu, Interformer: an interaction-aware model for protein-ligand docking and affinity prediction, *Nat. Commun.*, 2024, **15**, 10223.
16 J. Zhu, Z. Gu, J. Pei and L. Lai, DiffBindFR: an SE(3) equivariant network for flexible protein–ligand docking, *Chem. Sci.*, 2024, **15**, 7926–7942.
17 X. Zhang, C. Shen, H. Zhang, Y. Kang, C.-Y. Hsieh and T. Hou, Advancing Ligand Docking through Deep Learning: Challenges and Prospects in Virtual Screening, *Acc. Chem. Res.*, 2024, **57**, 1500–1509.
18 M. Buttenschoen, G. M. Morris and C. M. Deane, PoseBusters: AI-based docking methods fail to generate

physically valid poses or generalise to novel sequences, *Chem. Sci.*, 2024, **15**, 3130–3139.
19 D. Errington, C. Schneider, C. Bouysset and F. A. Dreyer, Assessing interaction recovery of predicted protein-ligand poses, *J. Cheminf.*, 2025, **17**, 76.
20 A. Mullard, When can AI deliver the drug discovery hits?, *Nat. Rev. Drug Discovery*, 2024, **23**, 159–161.
21 C. Deane and M. Mokaya, A virtual drug-screening approach to conquer huge chemical libraries, *Nature*, 2022, **601**, 322–323.
22 I. Wallach, D. Bernard, *et al*, AI is a viable alternative to high throughput screening: a 318-target study, *Sci. Rep.*, 2024, **14**, 7526.
23 S. Gu, C. Shen, X. Zhang, H. Sun, H. Cai, H. Luo, H. Zhao, B. Liu, H. Du and Y. Zhao, Benchmarking AI-powered docking methods from the perspective of virtual screening, *Nat. Mach. Intell.*, 2025, 1–12.
24 W. Treyde, N. Bouatta, S. C. Kim and M. AlQuraishi, QuickBind: A Light-Weight And Interpretable Molecular Docking Model, *arXiv*, 2024, preprint, arXiv: 2410.16474.v16471, DOI: **10.48550/arXiv.2410.16474**.
25 M. J. Hartshorn, M. L. Verdonk, G. Chessari, S. C. Brewerton, W. T. Mooij, P. N. Mortenson and C. W. Murray, Diverse, high-quality test set for the validation of protein–ligand docking performance, *J. Med. Chem.*, 2007, **50**, 726–741.
26 G. Corso, A. Deng, B. Fry, N. Polizzi, R. Barzilay and T. Jaakkola, Deep confident steps to new pockets: Strategies for docking generalization, *arXiv*, 2024, preprint, arXiv: 2402.18396.v18391, DOI: **10.48550/arXiv.2402.18396**.
27 M. R. Bauer, T. M. Ibrahim, S. M. Vogel and F. M. Boeckler, Evaluation and Optimization of Virtual Screening Workflows with DEKOIS 2.0 – A Public Library of Challenging Docking Benchmark Sets, *J. Chem. Inf. Model.*, 2013, **53**, 1447–1462.
28 M. M. Mysinger, M. Carchia, J. J. Irwin and B. K. Shoichet, Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking, *J. Med. Chem.*, 2012, **55**, 6582–6594.
29 C. Shen, Y. Hu, Z. Wang, X. Zhang, J. Pang, G. Wang, H. Zhong, L. Xu, D. Cao and T. Hou, Beware of the generic machine learning-based scoring functions in structure-based virtual screening, *Briefings Bioinf.*, 2021, **22**, bbaa070.
30 P. A. Williams, J. Cosme, D. M. Vinkovic, A. Ward, H. C. Angove, P. J. Day, C. Vonrhein, I. J. Tickle and H. Jhoti, Crystal structures of human cytochrome P450 3A4 bound to metyrapone and progesterone, *Science*, 2004, **305**, 683–686.
31 K. Haga, A. C. Kruse, H. Asada, T. Yurugi-Kobayashi, M. Shiroishi, C. Zhang, W. I. Weis, T. Okada, B. K. Kobilka, T. Haga and T. Kobayashi, Structure of the human M2 muscarinic acetylcholine receptor bound to an antagonist, *Nature*, 2012, **482**, 547–551.
32 R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge and F. A. Wichmann, Shortcut learning in deep neural networks, *Nat. Mach. Intell.*, 2020, **2**, 665–673.
33 S. Moon, W. Zhung and W. Y. Kim, Toward generalizable structure-based deep learning models for protein–ligand interaction prediction: Challenges and strategies, *Wiley Interdiscip. Rev.:Comput. Mol. Sci.*, 2024, **14**, e1705.
34 A. Mastropietro, G. Pasculli and J. Bajorath, Learning characteristics of graph neural networks predicting protein–ligand affinities, *Nat. Mach. Intell.*, 2023, **5**, 1427–1436.
35 E. Perola and P. S. Charifson, Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding, *J. Med. Chem.*, 2004, **47**, 2499–2510.
36 M. Vieth, J. D. Hirst and C. L. Brooks, Do active site conformations of small ligands correspond to low free-energy solution structures?, *J. Comput.-Aided Mol. Des.*, 1998, **12**, 563–572.
37 G. Corso, H. Stärk, B. Jing, R. Barzilay and T. Jaakkola, Diffdock: Diffusion steps, twists, and turns for molecular docking, *arXiv*, 2022, preprint, arXiv:2210.01776, DOI: **10.48550/arXiv.2210.01776**.
38 J. Xu, Distance-based protein folding powered by deep learning, *Proc. Natl. Acad. Sci. U. S. A.*, 2019, **116**, 16856–16865.
39 Z. Yang, J. Ji, S. He, J. Li, T. He, R. Bai, Z. Zhu and Y. S. Ong, Dockformer: A transformer-based molecular docking paradigm for large-scale virtual screening, *arXiv*, 2024, preprint, arXiv:2411.06740, DOI: **10.48550/arXiv.2411.06740**.
40 S. J. Teague, Implications of protein flexibility for drug discovery, *Nat. Rev. Drug Discovery*, 2003, **2**, 527–541.
41 H. Mazal, H. Aviram, I. Riven and G. Haran, Effect of ligand binding on a protein with a complex folding landscape, *Phys. Chem. Chem. Phys.*, 2018, **20**, 3054–3062.
42 J. A. Erickson, M. Jalaie, D. H. Robertson, R. A. Lewis and M. Vieth, Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy, *J. Med. Chem.*, 2004, **47**, 45–55.
43 K. W. Lexa and H. A. Carlson, Protein flexibility in docking and surface mapping, *Q. Rev. Biophys.*, 2012, **45**, 301–343.
44 H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, The protein data bank, *Nucleic Acids Res.*, 2000, **28**, 235–242.
45 Z. Liu, Y. Li, L. Han, J. Li, J. Liu, Z. Zhao, W. Nie, Y. Liu and R. Wang, PDB-wide collection of binding data: current status of the PDBbind database, *Bioinformatics*, 2015, **31**, 405–412.
46 S. Wagle, R. D. Smith, A. J. Dominic, D. DasGupta, S. K. Tripathi and H. A. Carlson, Sunsetting Binding MOAD with its last data update and the addition of 3D-ligand polypharmacology tools, *Sci. Rep.*, 2023, **13**, 3008.
47 R. D. Schaeffer, Y. Liao, H. Cheng and N. V. Grishin, ECOD: new developments in the evolutionary classification of domains, *Nucleic Acids Res.*, 2017, **45**, D296–D302.
48 G. Madhavi Sastry, M. Adzhigirey, T. Day, R. Annabhimoju and W. Sherman, Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments, *J. Comput.-Aided Mol. Des.*, 2013, **27**, 221–234.
49 T. Liu, Y. Lin, X. Wen, R. N. Jorissen and M. K. Gilson, BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities, *Nucleic Acids Res.*, 2007, **35**, D198–D201.

50 A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani and J. P. Overington, ChEMBL: a large-scale bioactivity database for drug discovery, *Nucleic Acids Res.*, 2011, **40**, D1100–D1107.
51 G. W. Bemis and M. A. Murcko, The properties of known drugs. 1. Molecular frameworks, *J. Med. Chem.*, 1996, **39**, 2887–2893.
52 J. J. Irwin and B. K. Shoichet, ZINC^{-} a free database of commercially available compounds for virtual screening, *J. Chem. Inf. Model.*, 2005, **45**, 177–182.
53 C. Bouysset and S. Fiorucci, ProLIF: a library to encode molecular interactions as fingerprints, *J. Cheminf.*, 2021, **13**.
54 M. Steinegger and J. Söding, MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets, *Nat. Biotechnol.*, 2017, **35**, 1026–1028.
55 C. Zhang, M. Shine, A. M. Pyle and Y. Zhang, US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes, *Nat. Methods*, 2022, **19**, 1109–1115.