Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal | View Issue

Cite this: *Chem. Sci.*, 2025, **16**, 18263

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

Received 18th March 2025
Accepted 31st August 2025

DOI: 10.1039/d5sc02113h

rsc.li/chemical-science

Molecule generation for target protein binding with hierarchical consistency diffusion model

Guanlue Li, †[a] Chenran Jiang, †[d] Ziqi Gao, [c] Yu Liu, [b] Chenyang Liu, [d] Jiean Chen, [d] Yong Huang *[b] and Jia Li *[c]

Effective generation of molecular structures that bind to target proteins is crucial for lead identification and optimization in drug discovery. Despite advancements in atom- and motif-wise models for 3D molecular generation, current methods often struggle with validity and reliability. To address these issues, we develop the Atom-Motif Consistency Diffusion Model (AMDiff), utilizing a joint-training paradigm for multi-view learning. This model features a hierarchical diffusion architecture that integrates both atom- and motif-views of molecules, allowing for comprehensive exploration of complementary information. By leveraging classifier-free guidance and incorporating topological features as conditional inputs, AMDiff ensures robust molecule generation across diverse targets. Compared to existing approaches, AMDiff exhibits superior validity and novelty in generating molecules tailored to fit various protein pockets. Case studies targeting protein kinases, including Anaplastic Lymphoma Kinase (ALK) and Cyclin-dependent kinase 4 (CDK4), demonstrate the capability in structure-based *de novo* drug design. Overall, AMDiff bridges the gap between atom-view and motif-view drug discovery and accelerating the development of target-specific molecules.

1 Introduction

In recent years, large-scale AI models have achieved remarkable breakthroughs, driving a surge of applications across various industries.[1,2] The pharmaceutical sector, in particular, has benefited significantly, with tools like AlphaFold revolutionizing protein structure prediction.[3] These advancements provide medicinal chemists with refined protein structures, accelerating structure-based drug design. Despite this progress, the chemical space of drug-like molecules remains vast and largely unexplored.[4,5] Traditional methods, such as virtual screening, are often inefficient, costly, and limited to known structures. AI-based models offer promising capabilities to effectively navigate this chemical space. Consequently, developing novel AI tools for end-to-end, structure-based drug discovery has become a crucial research direction.[6–8] In drug design, the interaction between a protein target and a ligand is often compared to the "Lock and Key" model, highlighting the necessity for precise structural complementarity.[9,10] The main challenge for medicinal chemists is to rapidly identify structurally novel and modifiable "keys" for the "lock"—the target protein. Advanced AI methodologies have facilitated this process by accelerating lead generation. However, *de novo* molecule generation, guided by the target pocket, remains insufficiently developed.[11,12] Unlike the straightforward "Lock and Key" analogy, this task requires processing 3D information within a flexible, continuous space. The limited data available often results in discontinuities and inaccuracies in atomic arrangements, leading to deviations from real-world atomic connectivity rules and bond lengths and angles that do not align with energy principles.

Existing models for *de novo* molecular generation often draw inspiration from real-life lead optimization strategies used in drug discovery, primarily chemical derivatization[13] and scaffold hopping.[14] Chemical derivatization involves a sequential approach, where molecules branch out from a known starting point. In contrast, scaffold hopping retains the molecule's overall 3D shape while altering atom connectivity. Recent tools like GraphBP[15] and FLAG[16] implement chemical derivatization by sequentially introducing specific atoms or motifs into a binding site. Meanwhile, ScaffoldGVAE[17] employs scaffold hopping by preserving side chains and modifying the main core. Additionally, novel frameworks, such as one-shot generation methods, present intriguing possibilities by creating entire molecular structures simultaneously.[18] TargetDiff[19] and

[a]*Data Science and Analytics, The Hong Kong University of Science and Technology (Guang Zhou), Guangzhou, 511400, China. E-mail: guanlueli@gmail.com*

[b]*Department of Chemistry, The Hong Kong University of Science and Technology, Hong Kong SAR, 999077, China. E-mail: yliuil@connect.ust.hk; yonghuang@ust.hk*

[c]*Division of Emerging Interdisciplinary Areas, The Hong Kong University of Science and Technology, Hong Kong SAR, 999077, China. E-mail: zgaoat@connect.ust.hk; jialee@ust.hk*

[d]*Pingshan Translational Medicine Center, Shenzhen Bay Laboratory, Shenzhen, 518118, China. E-mail: jiangcr@szbl.ac.cn; cyberyanglinox@outlook.com; chenja@szbl.ac.cn*

† Equal contributions.

DecompDiff[20] utilize this approach, employing diffusion models to generate molecules at the atom level in a single step. Regardless various strategies, these models typically use either individual atoms or motifs as building blocks for molecule construction. Atom-based *de novo* drug design methods[21–23] are not limited by predefined motif libraries, allowing exploration of vast chemical spaces and generation of highly diverse compounds. Yet, these methods are confronted with the validity of bond lengths and angles, which can result in the formation of structurally bizarre molecules. In contrast, motif-based approaches[16,24,25] utilize predefined libraries to assemble molecules, but the reliance on existing datasets and current chemical knowledge limits exploration of unknown chemical spaces. This restriction confines the potential to generate novel structures beyond the available fragments.

To balance novelty and validity, a hierarchical graph model can be used to generate molecules simultaneously at both the atom and motif levels. Several pioneering works have been inspired by this multi-granularity modeling. DrugGPS[21] incorporates an intrinsic two-level structure of the protein, featuring atom-level and residue-level encoders to learn sub-pocket prototypes, generating molecules motif-by-motif. Our development, HIGH-PPI[27] aims to establish a robust understanding of Protein–Protein Interactions (PPIs) by creating a hierarchical graph that includes both the PPI graph and the protein graph.

To achieve hierarchical performance at both the atom and motif levels, we introduce the Atom-Motif Consistency Diffusion Model (AMDiff), designed to efficiently generate high-quality 3D molecules for specific binding targets. AMDiff learns target information and constructs a graph structure incorporating topological details. At the ligand level, it employs a hierarchical diffusion approach, capturing both atom-view and motif-view of molecules to fully utilize available information. During molecular generation, we ensure that samples from both views are closely aligned in chemical space. The motif view provides insights into prior patterns, like aromatic rings, which the atom view might miss, while the atom view models diverse structures without being constrained by predefined motifs. The joint training approach leverages complementary information from different views, enhancing interaction during training. AMDiff employs the classifier-free guidance diffusion model in each view. We incorporate features extracted from binding sites as conditional inputs and train both conditional and unconditional diffusion models by randomly omitting the conditioning. This approach ensures balanced molecule synthesis across multiple targets. To enhance the coherence and connectivity of generated molecules, we incorporate persistent homology, a technique from topological data analysis (TDA). This method captures multiscale topological features from both molecules and binding sites. By integrating these topological features, we strengthen the structural characteristics of the generated molecules and refine binding site topology identification based on shape properties. We apply AMDiff to benchmark dataset and two kinase targets, demonstrating superior generation performance and effectiveness compared to other models. AMDiff exhibits exceptional performance when benchmarked to baseline methods, encompassing both atom- and motif-alone models, across diverse metrics. Further analysis on its robustness has verified that AMDiff can produce compounds tailored to varying pocket sizes.

2 Results

2.1 Atom-based and motif-based methods are complementary in target-aware molecule generation

Considering that molecular structures can be broken down into multiple levels of resolution, we aim to fully harness the potential of multi-level molecular structure generation. To this end, we propose AMDiff, a classifier-free hierarchical diffusion model designed for accurate and interpretable *de novo* ligand design through atom-motif consistency. Traditional ligand design strategies often utilize either individual atoms or motifs as foundational elements. Each granularity level offers distinct advantages and mechanisms for establishing interactions within the pocket environment, and they are not interchangeable. Depending solely on one resolution level may inadvertently overlook crucial structural patterns present at the other level. On one hand, using atoms as foundational units offers flexibility in accessing all possible coordinates in 3D space. However, it is challenging to produce reasonable structures due to the lack of necessary constraints to obey the fundamental electronic, steric, and geometric principles of chemistry. Atom-based methods often result in structural errors, such as unrealistic bond lengths and angles, which hinder proper ring formation. These models frequently generate thermodynamically unstable structures, like those with unstable fused cyclopropanes and seven-membered aromatic rings, due to the absence of Euclidean geometric constraints (Fig. 1, left panel). On the other hand, motif-based design builds a motif vocabulary from existing drug datasets and chemical knowledge, selecting suitable motifs to assemble final molecules. However, this approach faces limitations due to the restricted motifs contained in the vocabulary, which hinders access to structures that incorporate fragments beyond the existing motif repertoire, as demonstrated in Fig. 1(a), the right-hand panel. Additionally, incorrect integration between motifs can occur, such as improper linker construction or missing connections. These challenges are similar to those faced in atom-based generation approaches (for detailed comparison, refer to Table 2).

In the proposed model, AMDiff operates hierarchically, incorporating both atomic and motif views, as illustrated in Fig. 1(b). To connect these two views effectively, an interaction network is utilized. This network facilitates the exchange of complementary information between the atom-view and motif-view, enhancing the overall model performance. We establish ligand–protein interactions and cross-view interactions. Ligand–protein interactions are modeled through an equivariant graph neural network, which ensures that the generated molecules fit the target binding sites accurately by considering both geometric and chemical properties. Moreover, cross-view interactions are constructed to bridge the gap between atom-level precision and motif-level abstraction. Motifs interact with the target pocket, offering clustering information to the atom view, while the atom view provides detailed positioning information to the motif view. This bidirectional flow of

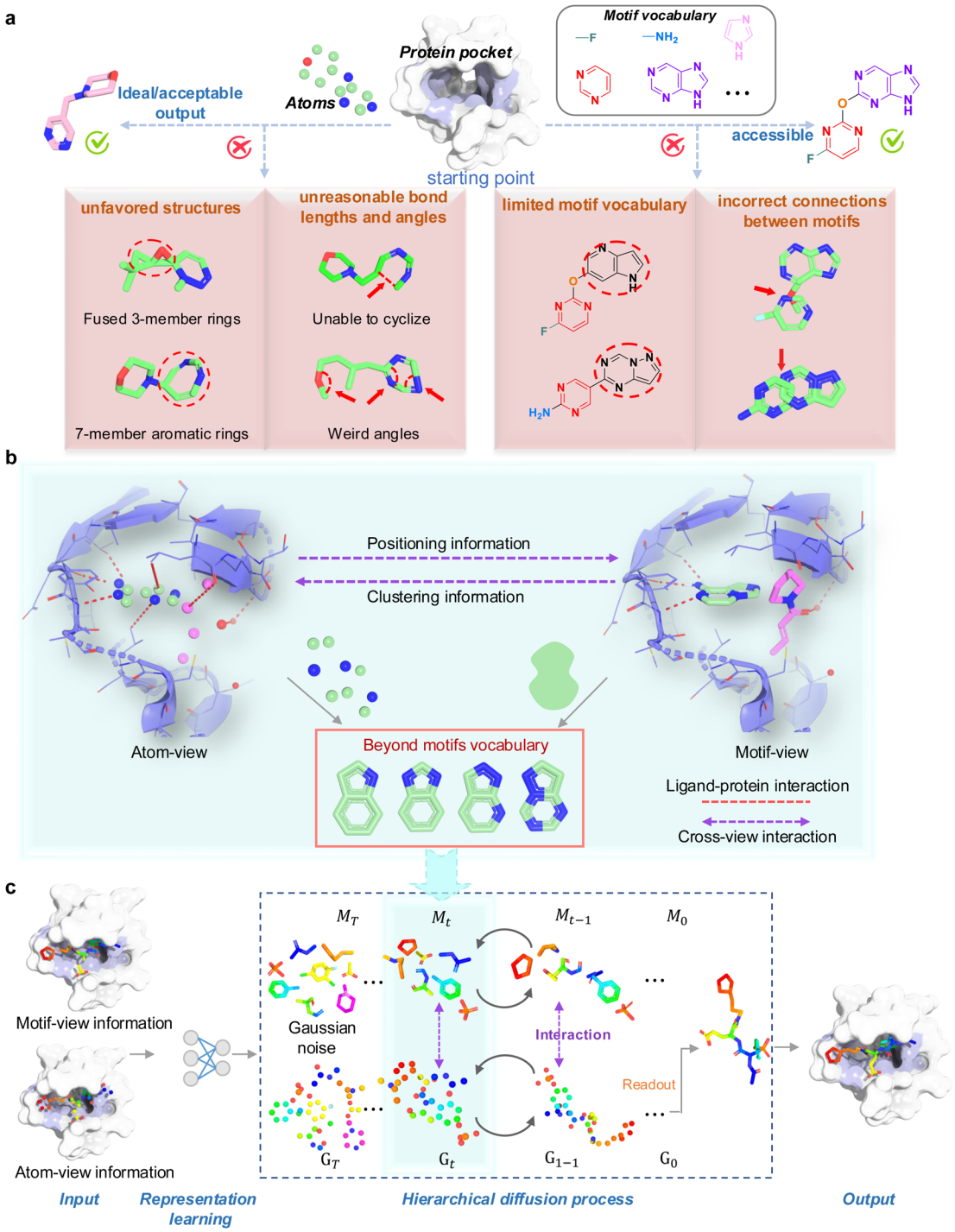

Fig. 1 (a) Ideal outputs and disadvantages of atom-based and motif-based methods for structure-based drug design. In atom-based methods (on the left), individual atoms serve as the fundamental units to construct highly diverse molecular structures. While these methods excel in generating variety, they often struggle to maintain coherence and realism in substructure formation, frequently leading to the creation of bonds with incorrect lengths and angles. Moreover, atom-based approaches can inadvertently produce unstable configurations such as three-membered rings. Conversely, motif-based methods (on the right) utilize predefined building blocks sourced from a motif vocabulary derived from existing datasets and chemical knowledge. However, these methods face limitations when desired motifs, such as 1*H*-pyrrolo[3,2-*b*]pyridine and pyrazolo[1,5-*a*][1,3,5]triazine, are absent from the vocabulary, potentially limiting structural diversity. Additionally, conflicts may arise in connecting different motifs, posing further challenges in generating cohesive structures. (b) The illustration of hierarchical-interaction information for ligand generation in this work. A ligand is decomposed into atoms and motifs, respectively. In the atom-view and motif-view, interaction details between the ligand and protein (represented by red dotted lines) are gathered using dedicated message passing networks. Additionally, for cross-view interactions (indicated by purple dashed line), facilitate the exchange of clustering and positioning information between the atom and motif views. (c) The AMDiff architecture is a diffusion-based model for hierarchical molecular generation. The AMDiff architecture is centered on a diffusion model that integrates atom-view and motif-view perspectives, each crucial for molecular generation. This

Table 1 Comparison of molecular properties between the baseline models and AMDiff targeting the CrossDocked[26] test set. The top three performing models are denoted by distinct colors, with the highest-performing model indicated by the darkest purple background color

Model	Validity ↑	Diversity ↑	Novelty ↑	QED ↑	SA ↑	Affinity ↓
liGAN	0.950 ± 0.021	0.488 ± 0.043	0.304 ± 0.016	0.411 ± 0.103	0.590 ± 0.089	−6.331 ± 1.108
AR	0.843 ± 0.012	0.583 ± 0.145	0.490 ± 0.177	0.434 ± 0.018	0.661 ± 0.099	−6.231 ± 1.660
Pocket2Mol	0.958 ± 0.002	0.612 ± 0.035	0.558 ± 0.102	0.454 ± 0.165	0.629 ± 0.134	−7.269 ± 2.239
GraphBP	0.974 ± 0.004	0.594 ± 0.014	0.547 ± 0.007	0.439 ± 0.090	0.512 ± 0.134	−7.381 ± 2.101
TargetDiff	0.957 ± 0.013	0.661 ± 0.025	0.629 ± 0.098	0.457 ± 0.175	0.606 ± 0.116	−7.367 ± 1.052
DecompDiff	0.967 ± 0.016	0.668 ± 0.026	0.624 ± 0.073	0.461 ± 0.154	0.624 ± 0.131	−7.324 ± 0.211
DiffBP	0.976 ± 0.016	0.659 ± 0.017	0.647 ± 0.038	0.492 ± 0.177	0.664 ± 0.162	−7.449 ± 0.162
FlAG	0.981 ± 0.013	0.604 ± 0.024	0.499 ± 0.129	0.463 ± 0.129	0.508 ± 0.150	−7.169 ± 2.019
AMDiff (ours)	0.989 ± 0.007	0.672 ± 0.013	0.663 ± 0.104	0.471 ± 0.209	0.681 ± 0.125	−7.466 ± 2.062
Testset	—	—	—	0.476 ± 0.206	0.727 ± 0.140	−7.502 ± 1.898

information ensures that the generated ligands not only fit the binding sites but also maintain structural coherence beyond the predefined motif vocabulary. A schematic view of the AMDiff architecture is shown in Fig. 1(c). The initial step in AMDiff involves obtaining representations of the protein and ligand through an embedding network. Subsequently, a denoising network predicts the state of ligand without noise. Each view employs a denoising process to predict the structure conditioned on binding sites, which includes a forward chain that perturbs data to noise and a reverse chain that converts noise back to data. In the atom-view, the model focuses on capturing the fine-grained details of atomic positions and interactions. This involves learning the precise atomic-level forces and positional information, providing a broader context that aids in forming reasonable molecular clusters and overall topology. In the motif-view, the model captures higher-level structural patterns, such as functional groups and larger molecular fragments, ensuring that the generated ligands are structurally coherent and chemically valid. By obtaining the persistence diagram and encoding it as topological fingerprints, AMDiff effectively captures the multi-scale topological features essential for accurate ligand generation, as detailed in Section 4.3.

We train our model on the CrossDocked dataset.[26] During the training phase, both atom-view and motif-view particles, along with their corresponding binding protein pockets, are input to the model. The protein pocket remains fixed as it serves as the conditional information. In the sampling stage, we initialize the data state by sampling from a standard normal distribution, $\mathrm{N}(0, \boldsymbol{I})$. Subsequent states are iteratively generated using $p_\theta(G_{t-1}|G_t,C)$, where C represents the condition. We evaluate AMDiff on the CrossDocked dataset, as well as on Anaplastic Lymphoma Kinase (ALK) and Cyclin-dependent kinase 4 (CDK 4) targets. We assess the performance of our model from two perspectives: (1) understanding the characteristic property distributions of ligands in different protein pockets. This entails learning the interaction patterns with protein pockets in order to achieve stronger binding. (2) Generating molecules for real-world therapeutic targets and exploring their interactions in the presence of mutated target proteins and varying pocket sizes.

2.2 AMDiff shows the best performance and generalization

We conduct a comprehensive evaluation of AMDiff's performance in generating molecular structures. Specifically, we select 100 protein targets and generate 100 molecules for each target, resulting in a total of 10 000 generated molecules. We compare the result with recent atom-based methods including LiGAN, AR, Pocket2Mol, GraphBP, DecompDiff, DiffBP and the motif-based FLAG. The evaluation metrics for molecule generation performance include Diversity, Novelty, QED, SA, and Affinity, as defined in Section 4.7.

In Table 1, we show the mean values with standard deviations of evaluation metrics. Generally, our method demonstrates best performance compared to the baseline methods. AMDiff achieves a 98.9% output validity, indicating its ability to accurately learn the chemical constraints of topological structures. The higher percentages of diversity and novelty in the generated molecules indicate that our model effectively explores the chemical space beyond the molecular structures present in the training dataset. We assess the affinity of the generated conformations of molecules by calculating the Vina docking value between the molecules and target proteins. Compared to the second-best method, AMDiff improves the QED ratio by 5.5%. Among the baseline methods, our model achieves the highest SA score. The results in Table 1 indicate that our model achieves an average affinity of -7.466 kcal mol^{-1}, demonstrating the model's capability to generate molecules with favorable binding affinity. Overall, our model exhibits improved performance compared to other methods.

model employs a conditional diffusion approach to recover noisy molecular structures and generate new ones through interactive denoising. In the atom-view, the model predicts atom types and positions, while in the motif-view, it constructs motif trees and generates predictions based on them. This architectural design fosters effective information exchange between views, providing valuable insights across various granularity levels in molecular structures.

To further evaluate AMDiff's capacity to accurately capture the distribution of training dataset and detect the distribution shifts with the test dataset, we conduct an additional analysis focusing on the physicochemical properties and topological structures of the generated molecules. We apply medicinal chemistry filters as described in Section 4.6 to exclude molecules with excessively high or low molecular weights, as well as those that are toxic or chemically infeasible. The distribution patterns of various key metrics, including docking score, molecular weight, QED, and SA, are presented in Fig. 2. The result shows that the generated molecules exhibit lower docking scores and smaller standard deviations compared to other baselines, indicating a higher affinity with target proteins. Regarding molecular weight, our model closely resembles that of known ligands, outperforming DiffBP and FLAG models. When comparing the QED distribution respectively, our model demonstrates mean values closer to the training and test sets. Models like AMDiff and FLAG tend to produce compounds with lower SA scores. This is primarily due to their propensity to generate more structurally intricate molecules, such as those with fused ring systems or polycyclic frameworks, which are inherently more challenging to synthesize. To better understand the chemical space occupied by the generated molecules, we assess the 3D shape distribution using NPR descriptors. As shown in Fig. 2(e), the molecules generated by our model tend towards a linear shape, slightly leaning towards the "planar" corner. This alignment with the training set suggests that our model produces molecules consistent with reference ligands. Conversely, FLAG model exhibit more divergent distributions, indicating a deviation from the expected shape characteristics, whose center of the distribution is positioned apart from the center of bioactive ligands. Although DiffBP has a distribution center closer to that of bioactive ligands, its shape distribution is more widely spread toward the planar region. Through above quantitative evaluations, AMDiff demonstrates excellent results by generating diverse and active results molecules, outperforming other methods in the drug design process.

We also calculate various bond angles and dihedral angle distributions for the generated molecules and compare them against the respective reference empirical distributions using Kullback–Leibler (KL) divergence. Bond angles and dihedral angles describe the spatial arrangement between three connected atoms and the rotation between planes defined by four sequential atoms, respectively. As depicted in Table 2, our model demonstrates lower KL divergence compared to all other atom-based baselines. Our model show competitive performance with FLAG, which operates on a motif-wise basis. The motif-wise models offer the advantage of predicting more precise angles within motifs by employing a strategy of combining predefined motifs from an established vocabulary. However, motif-based models also face the challenge of constructing cohesive connections between motifs. We additionally report statistics for aromatic angle distributions, where our model performs comparably to FLAG. The results highlight the

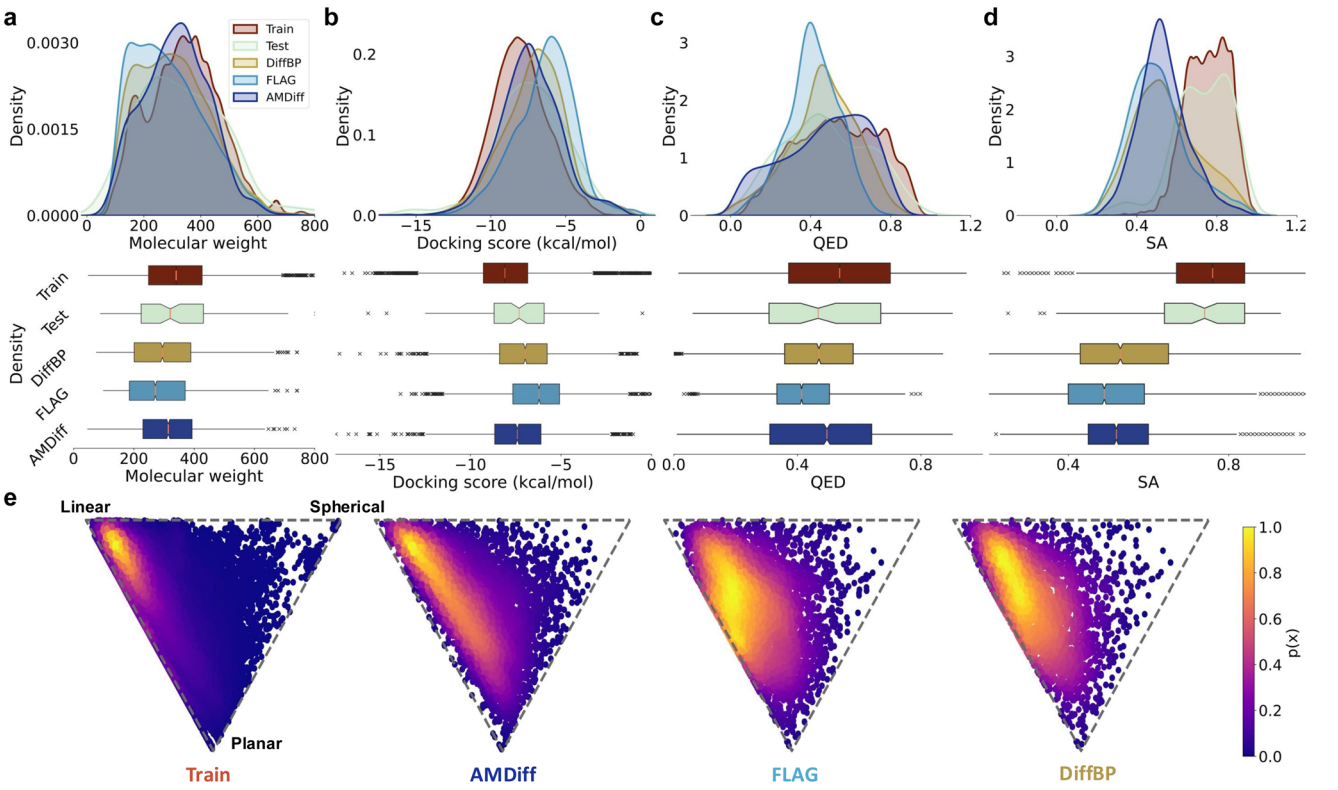

Fig. 2 Quantitative evaluations of the models targeting the CrossDocked[26] test set. (a–d) The distribution of the following metrics: (a) docking score; (b) molecular weight; (c) QED; (d) SA, comparing AMDiff (purple), DiffBP (yellow), FLAG (blue), train set (red), and test set (green) molecules. (e) Visualizes the 3D shape distribution of the generated molecules using NPR descriptors.

Table 2 The KL divergence measures the difference in bond angles and dihedral angles between the reference molecules and the generated molecules targeting the CrossDocked[26] test set. The lowercase letters represent the atoms in the aromatic rings. Lower values indicate that the models better capture the distribution of realistic structures. All the baseline models are atom-wise, with the exception of FLAG, which is a motif-based model. The top three performing models are denoted by distinct colors, with the highest-performing model indicated by the darkest purple background color

Model	CCC	CCO	CNC	CCN	CC=O	CCCC	CNCC	NCCN	CC=CC	CCCS	cccc	Cccc	ccco
liGAN	6.845	8.314	6.737	5.672	6.752	1.479	2.355	3.689	5.216	1.229	6.325	4.930	4.773
AR	1.973	1.857	2.54	2.361	2.743	1.271	0.947	1.417	3.575	0.854	8.267	8.311	5.801
Pocket2Mol	0.917	0.874	0.465	0.562	0.765	0.954	0.864	0.368	2.177	0.386	3.972	2.371	3.950
GraphBP	0.573	0.472	0.341	0.304	0.458	0.896	1.027	0.876	0.648	0.351	5.017	3.216	3.332
TargetDiff	0.396	0.417	0.324	0.339	0.312	0.510	0.542	0.461	0.493	0.326	0.950	1.294	0.758
DecompDiff	0.321	0.302	0.286	0.325	0.244	1.056	0.566	0.437	1.573	0.296	0.499	0.702	0.413
DiffBP	0.294	0.325	0.491	0.371	0.307	0.315	0.461	0.510	0.466	0.307	0.562	0.631	0.526
FLAG	0.251	0.318	0.196	0.287	0.216	0.396	0.658	0.547	0.336	0.317	0.311	0.482	0.339
AMDiff	0.306	0.297	0.257	0.294	0.205	0.471	0.431	0.348	0.292	0.208	0.413	0.422	0.368

effectiveness of AMDiff in capturing geometric characteristics and realistic substructures of bond angles and dihedral angles by utilizing atom and motif-views, thereby approaching the performance of motif-based models without relying on a predefined substructure.

2.3 AMDiff enables target-aware molecule generation for kinase targets

Our study focuses on harnessing the potential of AMDiff to predict novel small molecule binders to drug targets, aiming to expedite lead identification and optimization processes. Beyond examining the statistical performance metrics, our evaluation delves deep into the practical effectiveness of the model-generated molecules against real drug targets. In this aspect, we conduct a thorough assessment targeting two significant therapeutic kinases: ALK and CDK4. ALK, a receptor tyrosine kinase, plays a pivotal role in driving the progression of specific cancers like non-small cell lung cancer (NSCLC) and anaplastic large cell lymphoma (ALCL) when mutated or rearranged. Inhibitors of ALK disrupt the aberrant signaling pathways activated by mutated ALK proteins, effectively impeding cancer cell proliferation and viability. Notably, several ALK inhibitors, including crizotinib, ceritinib, alectinib, and lorlatinib, have been approved for treating NSCLC. On the other hand, CDK4, a key regulator of the cell cycle involved in the G1 to S phase transition, exerts control over cell proliferation by phosphorylating the retinoblastoma protein (Rb), leading to the release of E2F transcription factors. By inhibiting CDK4, drugs can halt the hyperphosphorylation of Rb, preserving its growth-suppressive function and arresting the cell cycle at the G1 phase. This mechanism underscores the potential of CDK4 inhibitors as treatments for cancers characterized by disrupted cell cycle regulation, such as breast cancer, melanoma, and certain sarcomas.

Following previous medicinal chemistry efforts,[28,29] we focus on the ATP-binding pockets of these two proteins, executed large-scale molecular generation, and systematically summarized and compared the performance of the generated molecules across various parameters. We employ AMDiff to generate 15 000 molecules and utilized a molecular filter, as described in Section 4.6, to identify high-quality candidates. Fig. 3(a)–(d) illustrates the affinity and drug-likeness properties of molecules targeting ALK (PDB id: 3LCS) produced by different methods. A bioactive ligand dataset serves as the reference, establishing a baseline for models capable of designing ALK-targeted active-like molecules.

Among the affinity prediction (docking score), AMDiff demonstrates the highest performance, indicating that our model has effectively learned the favorable molecular conformations within the target pockets. Additionally, we assess the molecular properties of the generated compounds. AMDiff achieves the highest QED and SA values, which closely resemble the distribution of active compounds. Fig. 3(e) displays the heatmaps of the docking score and QED value distributions for molecules generated by DiffBP, FLAG, and AMDiff. Each data grid is color-coded according to the corresponding SA score. Molecules generated using AMDiff exhibited higher docking scores compared to those generated using DiffBP and FLAG, and were in proximity to the docking scores of bioactive ligands. In addition to the aforementioned indicators, we also assess the spatial similarity between the molecular poses directly generated by models and those obtained after molecular docking. Specifically, we investigate the root-mean-square deviation (RMSD) between the generated conformations and docked structures, with detailed definitions provided in eqn (16) of the Methods section 4.7. As shown in Fig. 3(f), our model exhibits lower RMSD values, indicating minimal deviation from the optimal docked conformations. This demonstrates AMDiff's ability to generate conformations with minimal shifts, closely aligning with the docked poses.

To verify the capability of AMDiff in recognizing protein pockets, we further validated it using 3D visualization. Firstly, to visualize the generative process of the AMDiff, we showcase the gradual generative diffusion process of the model within the binding pocket of CDK4, as in Fig. 4(a). At nodes with time steps of 200, 500, 800, and 1000, atom-view and motif-view progressively capture the features of the protein pockets and guide the diffusion process. Through generation and interaction at these nodes, compound 1 is ultimately formed. AMC-Diff

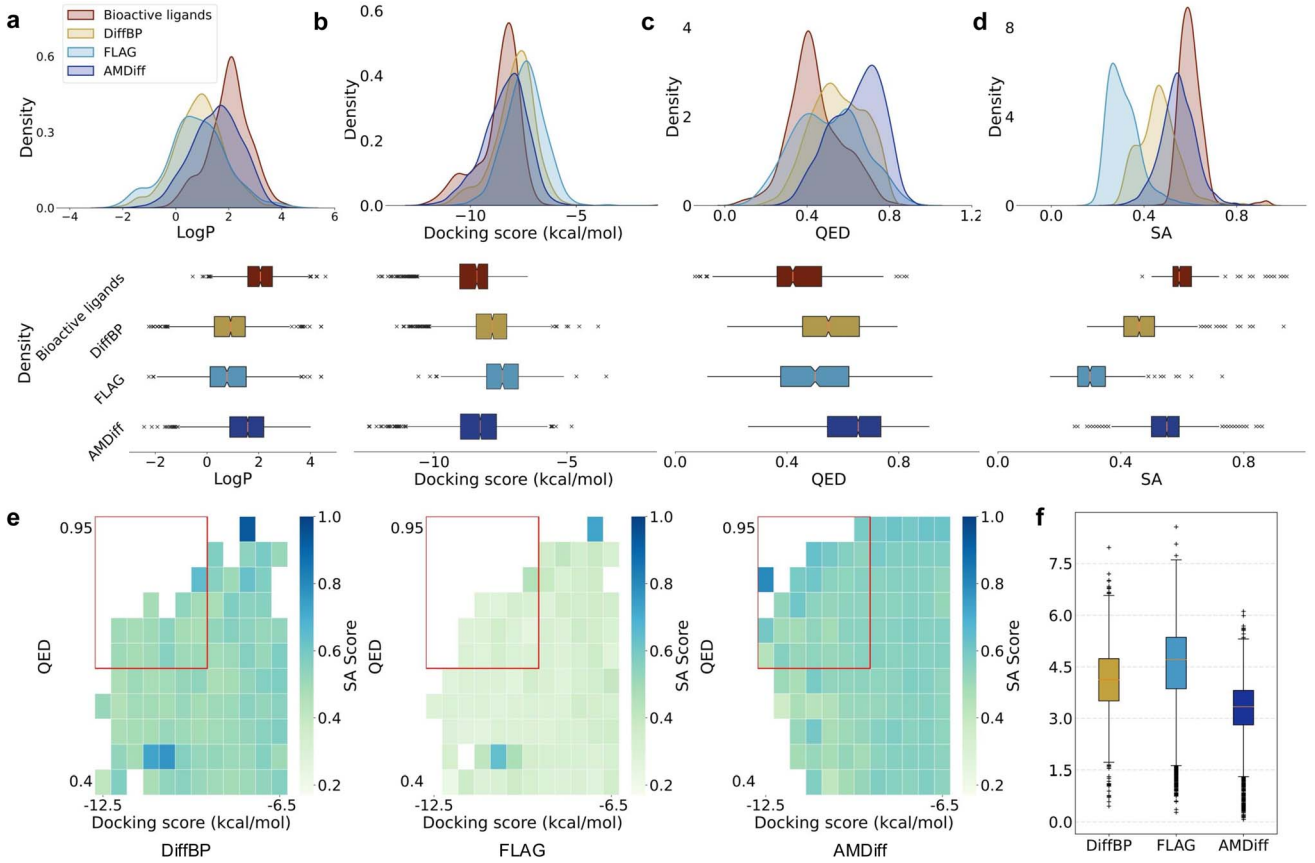

Fig. 3 Quantitative evaluations of the models targeting ALK (PDB id: 3LCS). The distributions of the following metrics were analyzed: (a) docking score; (b) molecular weight; (c) QED; (d) log *P*, comparing the performance of AMDiff (purple), Pocket2Mol (yellow), FLAG (green) models, molecules, and bioactive ligands (red). (e) The distribution of docking score, QED, and SA score for the generated samples was visualized. The drug-like region with QED ≥ 0.65 and docking score ≤ −8.5 (kcal mol^{-1}) is indicated with red boxes. (f) The RMSD was calculated to determine the conformational changes before and after the docking process.

demonstrates distinct generative processes in both atom- and motif-views, allowing for observations of atom or motif substitutions. However, following the cross-view interaction within the hierarchical diagram, the final molecule integrates the benefits of both views, resulting in a structure that is well-suited to the protein pocket. Then, based on similar molecular generation processes, we screen out other compounds from the generated molecules as potential CDK4 inhibitors, and analyze whether our model effectively learns the intricate microscopic interaction patterns within protein-ligand complexes. Key molecular descriptors are exhibited, including the quantitative estimate of QED, SA scores and top-1 docking score from AutoDcok Vina. The best conformation for each compound is also described in both 2D and 3D view. As shown in Fig. 4(b), most of these molecules exhibit interactions with the same amino acid residues. This suggests that the generated molecules are capable of fitting into the binding sites. Regarding pharmacophoric groups, AMDiff creates common important pharmacophore elements as the reference ligands. Specifically, compounds 1 and 2 form hydrogen bonds with Val96 and Glu144. Compound 3 forms hydrogen bonds with Val96, Glu144, and Asp99, as well as a pi–cation interaction with Asp99. Compound 4 forms hydrogen bonds with Val96, Asp97, and Asp158. The binding modes of the generated compounds align with the recognized binding patterns, demonstrating the target-aware ability of AMDiff to utilize known interactions while potentially uncovering novel ones.

2.4 AMDiff exhibits robustness in protein evolution, adapting to mutated proteins and multi-scale pocket sizes

Protein mutations are a natural part of evolution but can often lead to drug resistance during treatment, as they alter target interactions and modify pocket shape and size. Given AMDiff's performance in the gradual diffusion process, we explored its capability to address mutant proteins. The ALK protein mutation is a well-known challenge in drug development, as such mutations can significantly reduce therapeutic efficacy.

To address this, we generate ALK inhibitors against both wild-type ALK proteins from PDB bank (ALK^{WT}, PDB id: 3LCS) and Alphafold³ (AF-ALK^{WT}). We also design two mutant proteins based on AF-ALK^{WT} through site-directed mutagenesis: (i) AF-ALK^{G1202R}, where glycine (Gly) at position 1202 is replaced with arginine (Arg), and (ii) AF-ALK^{S1206Y}, where serine (Ser) at position 1206 is substituted with tyrosine (Tyr). Fig. 4(c) present the t-

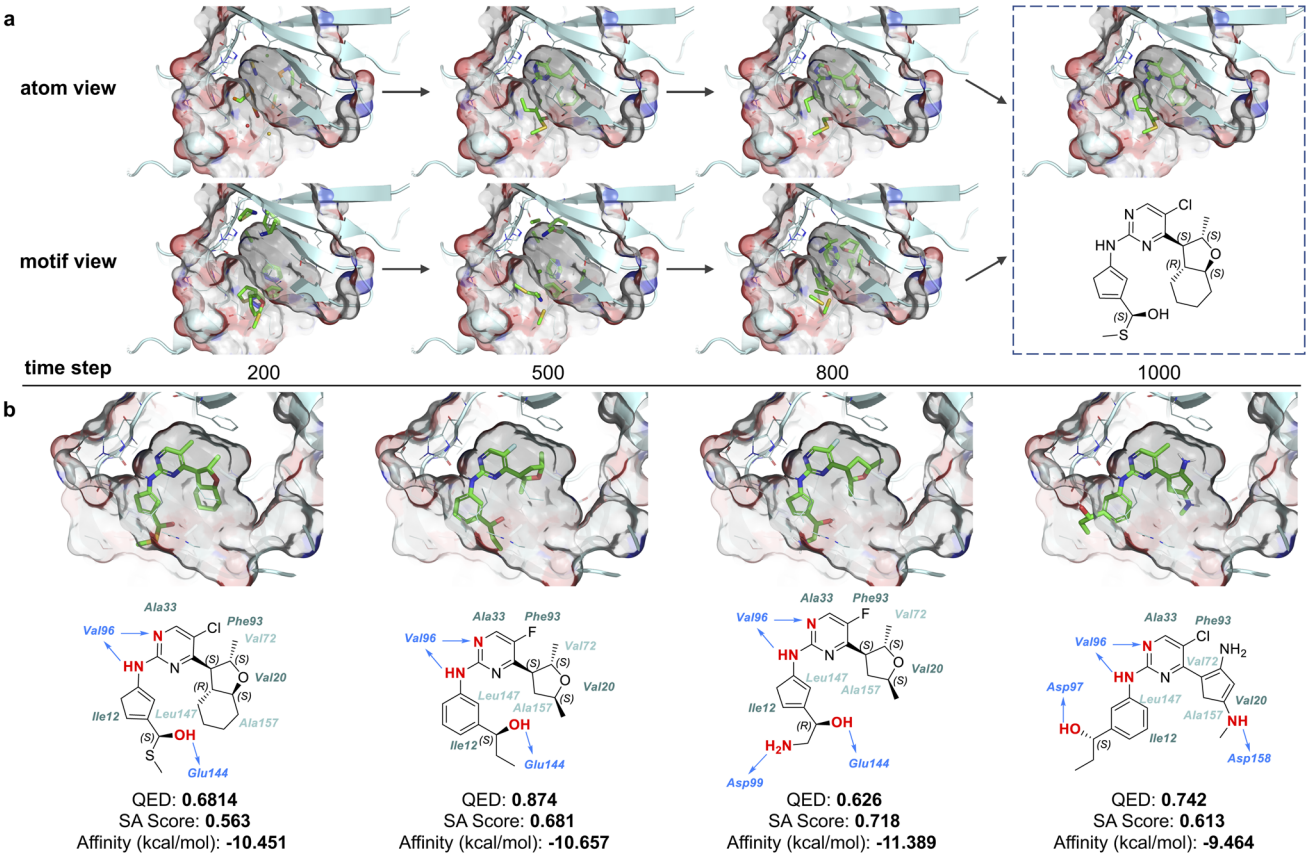

Fig. 4 Examples of molecules generated by AMDiff targeting CDK 4 (PDB id:7SJ3). (a) An example of a conditional design trajectory. At initial time steps, substructures progressively explore interactions with the pocket in both atom-view and motif-view. The trajectory gradually refines into a realistic molecule structure. (b) Molecules designed to target CDK 4 (PDB id:7SJ3), with molecular properties such as QED and SA score, as well as binding affinity and protein-ligand interaction analysis.

SNE visualization of the distributions of the USRCAT fingerprint[30] for molecules generated for these four proteins. The results reveal significant overlap in the chemical space of the generated molecules, yet there are distinct regions where the distributions do not overlap. This indicates that AMDiff can explore variations in local areas and align generated molecules with the target binding sites effectively. We further showcase the 3D-binding modes and molecular differences by AMDiff in Fig. 4(d). These molecules can form robust hydrogen bonds with Met1199 on AF-ALK^{WT}, AF-ALK^{G1202R} and AF-ALK^{S1206Y}. For mutations at positions 1202 and 1206, AMDiff recognizes steric hindrance and generates differentiated molecular structures in a targeted manner, coordinating energy loss at the global level. Given AMDiff's sensitivity to mutagenesis-induced differences during molecular generation, we further investigate its performance at different scales of protein pockets. Specifically, we assess ligands generated within ALK pocket sizes ranging from 4 Å to 30 Å. Fig. 5(a) compares the docking score, molecular weight, QED, and SA score of these molecules with those generated by the DiffBP and FLAG. Fig. 5(b) performs the 3D molecular spheres illustrating pocket fitness for molecules generated by AMDiff. The results indicate that AMDiff successfully generated viable molecules across all pocket scales. In contrast, FLAG exhibit limitations in generating normal molecules when the binding sites were too small (4 Å), suggesting these methods are constrained by their preset pocket boundaries. This could be attributed to the introduction of atom-view in AMDiff, which can construct smaller pieces when encountering larger hindrances, thus accommodating small pocket size.

2.5 Ablation study of AMDiff

We conducted an ablation study to evaluate the effectiveness of the hierarchical architecture, guidance diffusion, and topological features. We tested three variants of AMDiff: without motif-level (AMDiff w/o M), without classifier-free guidance (AMDiff w/o CG), and without topological features (AMDiff w/o TF). Table 3 shows the results on the CrossDocked test set. The experiments demonstrate that motif-view branches enhance molecular validity and SA by providing effective structural patterns. Additionally, both classifier-free guidance and topological features improve binding affinity by strengthening the integration of pocket and ligand representations.

3 Discussion

3.1 Hierarchical representation

Hierarchical organization is a fundamental characteristic of numerous biological structures, including proteins, DNA, and

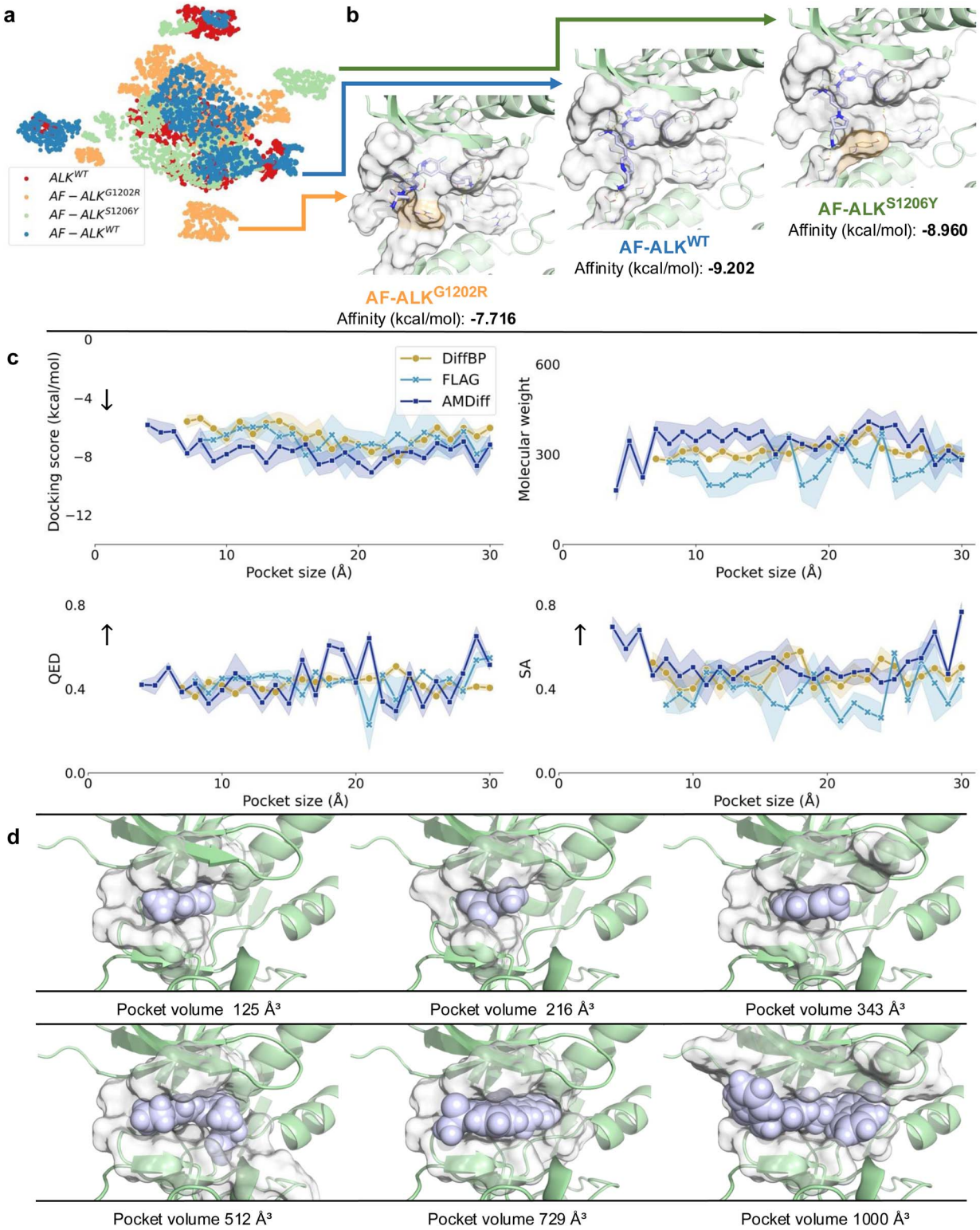

Fig. 5 (a) The distribution of molecules generated after mutating ALK (PDBID: 3LCS) is shown. The clustering results of USRCAT fingerprints for molecules targeting three mutations were visualized using t-SNE in two-dimensional space. ALK^{WT}: wild-type ALK proteins form PDB bank (PDB id: 3LCS). $AF-ALK^{WT}$: wild-type ALK proteins form Alphafold (PDB id: 3LCS). $AF-ALK^{G1202R}$: a substitution of the amino acid Gly with Arg at position 1202 in the protein sequence. $AF-ALK^{S1206Y}$: a substitution of the amino acid Ser with Tyr at position 1206 in the protein sequence. (b) Examples of molecules generated after modifying residues within the pocket of $AF-ALK^{G1202R}$, $AF-ALK^{S1206Y}$ and $AF-ALK^{WT}$. (c) Conditional generation of molecules for various pocket sizes targeting ALK (PDB ID: 3LCS). Comparison of key property performance when utilizing binding pockets of varying sizes, including docking score, molecular weight, SA and QED. (d) Visualization examples showcasing generated samples adjusted to match different pocket sizes. The molecular volumes are tailored to correspond with the given pocket volumes.

Table 3 The ablation study of AMDiff targeting the CrossDocked[26] test set. 'w/o M', 'w/o CG' and 'w/o TF' mean the variants of AMDiff without motif-view, classifier-free guidance and topological features respectively

Variant	Validity ↑	Diversity ↑	Novelty ↑	QED ↑	SA ↑	Affinity ↓
w/o M	0.896 ± 0.105	0.623 ± 0.089	0.576 ± 0.035	0.427 ± 0.073	0.603 ± 0.047	−7.349 ± 1.205
w/o CG	0.919 ± 0.061	0.598 ± 0.123	0.607 ± 0.156	0.459 ± 0.018	0.638 ± 0.104	−7.026 ± 1.050
w/o TF	0.937 ± 0.026	0.631 ± 0.078	0.618 ± 0.118	0.462 ± 0.063	0.652 ± 0.069	−7.109 ± 1.509
AMDiff	0.989 ± 0.007	0.672 ± 0.013	0.663 ± 0.104	0.479 ± 0.209	0.684 ± 0.125	−7.466 ± 2.062
Test set	—	—	—	0.476 ± 0.206	0.727 ± 0.140	−7.502 ± 1.898

molecules. Hierarchical representations have been successfully employed in various models, offering a more comprehensive understanding of biological systems at multiple scales. Protein structures can be represented at various levels of granularity, including the amino acid, backbone, and all-atom levels. ProNet[31] capture hierarchical relations among different levels and learn protein representations. HierDiff[32] employs a diffusion process to generate fragment representations instead of deterministic fragments in the initial step, followed by constructing atom-level 3D structures. HIGH-PPI[27] establish a robust understanding of Protein–Protein Interactions (PPIs) by creating a hierarchical graph that encompasses both the PPI graph and protein graph. By incorporating multiple levels of structural information, these models offer more nuanced and comprehensive representations of biological systems, potentially leading to improved predictions and deeper insights into biological functions and interactions. Similarly, in molecular domain, the Junction Tree Variational Autoencoder (JT-VAE)[33] introduces a hierarchical framework by clustering chemically meaningful substructures (*e.g.*, rings and functional groups) and assembling them into a junction tree. While this approach captures both local chemical motifs and their global connectivity, JT-VAE constructs the tree by applying a maximum spanning tree algorithm to the cyclic cluster graph, inherently introducing non-determinism. As a result, multiple valid junction tree decompositions may exist for the same molecule, leading to ambiguity in structure representation.

In this paper, we tackle the *de novo* ligand design problem from a hierarchical perspective, introducing a cross-view diffusion model that generates molecular structures at both the atomic and motif views. Our model excels in recognizing multi-level geometric and chemical interactions between ligands and target proteins. By capturing varying levels of granularity, we construct ligand–protein and cross-view interactions. Existing methods often neglect or inadequately utilize the hierarchical structure within molecules. Through empirical evaluation, our model, AMDiff, demonstrates its strong capability to generate valid molecules while maintaining the integrity of local segments. Furthermore, it exhibits robustness across diverse protein structures and pocket sizes.

3.2 Limitations and future work

3.2.1 Lack of support for dynamic protein structures. AMDiff does not currently account for the dynamic nature of protein conformations. In real-world applications, protein structures can undergo conformational changes, leading to shape shifts in pockets. Additionally, the formation of cryptic pockets expands the possibilities for drug discovery by enabling the targeting of proteins with multiple druggable binding sites. We recommend that future work considers the dynamics of protein structures, accounting for intrinsic or induced conformational changes.

3.2.2 Limited use of domain knowledge. It is essential to incorporate more domain knowledge into the model-building process. The integration of chemical and biomedical prior knowledge has proven effective in various tasks. Investigating the interactions between proteins and bioactive ligands, such as hydrogen bonds, salt bridges, pi–cation, and pi–pi stacking, is crucial. Furthermore, exploring the influence of pharmacophore elements during the generation of bioactive molecules would be valuable.

3.2.3 Absence of empirical evaluation. We evaluate the designed drug candidates using multiple metrics, it remains necessary to tightly collaborate with medicinal chemists and conduct wet-lab experiments for *in vitro* or *in vivo* validation of their effectiveness. The experimental results obtained can then be utilized to refine and improve the generative model.

3.2.4 Virtual screening remains unexplored. While AMDiff is currently designed as a generative framework for *de novo* ligand design, its hierarchical representations and topological features offer potential for extension into virtual scoring tasks. In principle, AMDiff could be adapted—similar to DiffDock's confidence scoring[34]—by introducing regression modules to predict binding scores directly from its learned structural features. However, several challenges remain, including the need for high-quality experimental data, the difficulty of accounting for solvent effects, and the risk of overfitting when using a unified framework for both generation and evaluation. At this stage, we recommend maintaining a practical separation between the generation and scoring processes. Specifically, we propose a combinatorial virtual screening workflow that integrates fast docking tools (*e.g.*, Vina-GPU[35]), advanced scoring models (*e.g.*, FeatureDock[36]), and high-precision methods (*e.g.*, FEP) to ensure both efficiency and accuracy across the screening pipeline.

3.2.5 Other modeling constraints. While AMDiff's hierarchical learning framework provides advantages in structural validity, geometric realism, and flexibility beyond fixed motif vocabularies, we acknowledge that its numerical improvements over strong baselines like FLAG and DiffBP are incremental. In particular, the atom-only variant of AMDiff currently underperforms compared to models such as DiffBP, which leverages

pre-trained modules to estimate atom count and molecular centers, improving spatial constraints but limiting generalizability to complex systems like protein–ligand assemblies. Similarly, DecompDiff introduces bond-level diffusion to enhance chemical validity, highlighting an important direction for future improvement. Extending AMDiff to integrate atom, bond, and motif levels within a unified hierarchical framework could further strengthen its ability to generate structurally coherent and chemically plausible molecules across diverse tasks.

4 Conclusions

In this work, we present AMDiff, a novel deep learning approach for *de novo* drug design. Our method focuses on generating 3D molecular structures within the binding site of a target protein. By employing the conditional diffusion model, AMDiff effectively learns the interactions across various binding pockets of the protein. One distinct feature of AMDiff is its hierarchical structure within the diffusion process. This architecture allows us to capture the 3D structure at both the atom and motif levels, overcoming the limitations of conventional atom- and motif-based molecular generation approaches. We incorporate a classifier-free guidance mechanism that provides interaction signals randomly during each denoising step, enabling the network to strengthen its guidance in identifying pocket structures.

To evaluate the performance of AMDiff, we conduct experiments aimed at designing potential hits for selected drug targets, using ALK and CDK4 as case studies. The results demonstrate that AMDiff can successfully generate drug-like molecules with novel chemical structures and favorable properties, exhibiting significant pharmacophore interactions with the target kinases. Additionally, AMDiff demonstrates high flexibility in generating 3D structures with various user-defined pocket sizes, further enhancing its utility. Overall, our work introduces a promising tool for structure-based *de novo* drug design, with the potential to significantly accelerate the drug discovery process.

5 Materials and methods

5.1 Experimental design

The protein can be represented as a set of atoms $P = \{(\boldsymbol{x}^{(i)},\boldsymbol{v}^{(i)})\}_{i=1}^{N_P}$ where N_P is the number of protein atoms, $\mathbf{x}\in\mathbb{R}^3$ represents the 3D coordinates of the atom, and $\mathbf{v}\in\mathbb{R}^F$ represents protein atom features such as amino acid types. The molecule with N atoms can be represented as $G = \{(\boldsymbol{x}^{(i)},\boldsymbol{v}^{(i)})\}_{i=1}^{N_G}$ where $\mathbf{x}\in\mathbb{R}^3$ is the atom coordinate, $\mathbf{v}\in\mathbb{R}^V$ is the one-hot atom type. $R = \{(\boldsymbol{x}^{(i)},\boldsymbol{v}^{(i)})\}_{i=1}^{N_R}$ is the atom 3D coordinates and features of binding pocket. A motif M_i is defined as a subgraph of a molecule G. Given a molecule, we extract its motifs $M = \{(\boldsymbol{x}^{(i)},\boldsymbol{w}^{(i)})\}_{i=1}^{N_M}$, where $\mathbf{x}\in\mathbb{R}^3$ represents the motif coordinates in 3D space, $\mathbf{w}\in\mathbb{R}^W$ denotes the motif IDs in the motif vocabulary. The motif vocabulary includes common molecular motifs extracted from the whole training set for molecule generation. The details about motif vocabulary construction can be found in Section 4.7. The goal is to develop a generative model, denoted as $p(G,M|R,P)$, that predicting the three-dimensional structure of ligands conditioned on the protein and binding pocket.

5.2 Classifier-free guidance diffusion model

We develop a conditional diffusion model for target-specific molecule generation. Our approach consists of a forward diffusion process and a reverse generative process. The diffusion process gradually injects noise to data, and the generative process learns to reconstruct data distribution from the noise distribution using a network parameterized by θ. The latent variable of every time step is represented as $p_\theta(G_t|(R,P))$, indicating that the predicted molecules are conditioned on the provided protein and binding sets. We can represent the diffusion process and the recovery process as follows:

$$\begin{aligned} q(G_{1:T}|G_0,(R,P)) &= \prod_{t=1}^{T} q(G_t|G_{t-1},(R,P)),\\ p_\theta(G_{0:T-1}|G_T,(R,P)) &= \prod_{t=1}^{T} p_\theta(G_{t-1}|G_t,(R,P)), \end{aligned} \tag{1}$$

where $G_1, G_2, \ldots, G_T$ is a sequence of latent variables of with the same dimensionality as the input data $G_0 \sim p(G_0|R,P)$. Following continuous diffusion model,[37] during the forward process, we add Gaussian noise to atom coordinates at each time step t according to the variance schedule $\beta_1, \ldots, \beta_T$:

$$q(x_t|x_0) = \mathrm{N}\Big(x_t;\sqrt{\overline{\alpha}_t}\,x_0,(1-\alpha_t)\boldsymbol{I}\Big),\quad \alpha_t := 1-\beta_t \quad \overline{\alpha}_t := \prod_{s=1}^{t}\alpha_s. \tag{2}$$

The model can be trained by optimizing the KL-divergence between $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)$ and $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$. The posterior distributions of the forward process, $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)$ can be represented as:

$$\begin{aligned} q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0) &= \mathrm{N}\big(\boldsymbol{x}_{t-1};\tilde{\mu}_t(\boldsymbol{x}_t,\boldsymbol{x}_0),\tilde{\beta}_t\boldsymbol{I}\big),\\ \tilde{\mu}_t(\boldsymbol{x}_t,\boldsymbol{x}_0) &:= \frac{\sqrt{\overline{\alpha}_{t-1}}\beta_t}{1-\overline{\alpha}_t}\boldsymbol{x}_0 + \frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}\boldsymbol{x}_t,\quad \tilde{\beta}_t := \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}\beta_t. \end{aligned} \tag{3}$$

In the denoising process, the transition distribution can be written as $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathrm{N}(\boldsymbol{x}_t-1;\mu_\theta(\boldsymbol{x}_t,t),{\sigma_t}^2\boldsymbol{I})$ and let ${\sigma_t}^2 = \beta_t$ experimentally. To reparametrize eqn (2), we define $\boldsymbol{x}_t(\boldsymbol{x}_0,\boldsymbol{\varepsilon}) = \sqrt{\overline{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon}\sim\mathrm{N}(0,\boldsymbol{I})$. Consequently, the final objective of the training process can be expressed as:

$$L_{\mathrm{pos}} = \mathbb{E}_{x_0,\varepsilon}\left[\frac{{\beta_t}^2}{2{\sigma_t}^2\alpha_t(1-\overline{\alpha}_t)}\|\boldsymbol{\varepsilon}-\boldsymbol{\varepsilon}_\theta\Big(\sqrt{\overline{\alpha}_t}\,\boldsymbol{x}_0 + \sqrt{1-\overline{\alpha}_t}\,\boldsymbol{\varepsilon},t,(R,P)\Big)\|^2\right]. \tag{4}$$

The representation of atom types can be achieved through a one-hot vector, denoted as $\boldsymbol{v}_t$. To predict the atom types in molecules, we utilize a discrete diffusion model.[38,39] At each timestep, a uniform noise term, β_t, is added to the previous timestep $\boldsymbol{v}_{t-1}$ across the K classes. This discrete forward transition can be expressed as follows:

$$q(\boldsymbol{v}_t|\boldsymbol{v}_{t-1}) = \mathrm{C}(\boldsymbol{v}_t; (1-\beta_t)\boldsymbol{v}_{t-1} + \beta_t/K), \tag{5}$$

where C is the categorical distribution. Through Markov chain, we can get:

$$q(\boldsymbol{v}_t|\boldsymbol{v}_0) = \mathrm{C}(\boldsymbol{v}_t|\overline{\alpha}_t\boldsymbol{v}_0 + (1-\overline{\alpha}_t)/K), \tag{6}$$

where $\alpha_t = 1 - \beta_t$ and $\overline{\alpha}_t = \prod_{\tau=1}^{t} \alpha_\tau$. Then the categorical posterior can be calculated as:

$$\begin{gathered} q(\boldsymbol{v}_{t-1}|\boldsymbol{v}_t, \boldsymbol{v}_0) = \mathrm{C}(\boldsymbol{v}_{t-1}|\boldsymbol{Q}(\boldsymbol{v}_t, \boldsymbol{v}_0)), \quad \boldsymbol{Q}(\boldsymbol{v}_t, \boldsymbol{v}_0) = \tilde{G} \Big/ \sum_{k=1}^{K} \tilde{G}_k, \\ \tilde{G} = [\alpha_t\boldsymbol{v}_t + (1-\alpha_t)/K] \odot [\overline{\alpha}_{t-1}\boldsymbol{v}_0 + (1-\overline{\alpha}_{t-1})/K]. \end{gathered} \tag{7}$$

For the training object of atom types, we can compute KL-divergence of categorical distributions:

$$L_{\text{type}} = \sum_k \boldsymbol{Q}(\boldsymbol{v}_t, \boldsymbol{v}_0)_k \log \frac{\boldsymbol{Q}(\boldsymbol{v}_t, \boldsymbol{v}_0)_k}{\boldsymbol{Q}(\boldsymbol{v}_t, \hat{\boldsymbol{v}}_0)_k}. \tag{8}$$

Motivated by the ability of guided diffusion models to generate high quality conditional samples, we apply classifier-free guided diffusion to the problem of pocket-conditional molecule generation. The feature of pocket provides a useful guidance signal to generation 3D structure in binding sites. During training process, the pocket in the diffusion model $\boldsymbol{\varepsilon}_\theta(G_t|R,P)$ is replaced with a null label $\varnothing$ with a fixed probability. During sampling, the output of the model is extrapolated further in the direction of $\boldsymbol{\varepsilon}_\theta(G_t|R,P)$ and away from $\boldsymbol{\varepsilon}_\theta(G_t|\varnothing,P)$:

$$\hat{\boldsymbol{\varepsilon}}_\theta(G_t|(R,P)) = \boldsymbol{\varepsilon}_\theta(G_t|(\varnothing,P)) + s\cdot(\boldsymbol{\varepsilon}_\theta(G_t(R,P)) - \boldsymbol{\varepsilon}_\theta(G_t|(\varnothing,P))). \tag{9}$$

Here, $s \geq 1$ is the guidance scale. By jointly train a conditional and an unconditional diffusion model, we combine the resulting conditional and unconditional score estimates to attain a trade-off between sample quality and diversity similar to that obtained molecules using classifier guidance. This methodology proves useful in obtaining a truncation-like effect across various proteins and sets of multi-druggable bindings.

5.3 Joint training for hierarchical diffusion

We build the Atom-Motif Consistency Diffusion Model (AMDiff), which is implemented in a joint training manner. This model incorporates both atom view and motif view to construct molecular structures. The atom view utilizes atoms as fundamental building blocks, allowing for the generation of highly diverse molecular structures. On the other hand, the motif view assembles subgraphs by leveraging fragments within a motif vocabulary, aiding in the learning of prior patterns. The interaction between these two views promotes transfer learning.

Each motif is treated as a rigid structure associated with a local coordinate frame. During the diffusion process, the motif view predicts both a motif ID (from a pre-defined vocabulary) and a Euclidean transformation consisting of a 3D translation vector. This transformation is applied to the idealized coordinates of the selected motif, thereby situating it in global space. Motif-level updates influence atom-level generation through a shared embedding space and synchronized updates in the joint diffusion network. Additional details are provided in SI C. For motif IDs, we adopt a discrete diffusion process governed by a uniform transition matrix, consistent with the atom-level scheme in AMDiff. Additionally, we introduce an alternative variant for comparison, AMDiff-Embedding Distance (AMDiff-ED), which employs transition matrices constructed from motif embedding similarities to guide the forward nosing process. Further details are provided in SI Section E.4.

We employ atom-view and motif-view diffusion models to generate feature representations in a latent space. To formulate the proposed AMDiff approach, we introduce a hierarchical diffusion model for the atom view and motif view, denoted by $\Phi(G_t,\theta_1)$ and $\Phi(M_t,\theta_2)$, respectively. These hierarchical diffusion networks update simultaneously, and the overall recovery process can be represented as:

$$(\hat{G}_0, \hat{M}_0) = \Phi_{\theta_1,\theta_2}(G_t, M_t, t, (R,P)). \tag{10}$$

The reverse network, denoted as $\Phi_{\theta_1,\theta_2}(G_t,M_t,t,(R,P))$, is constructed using equivariant graph neural networks (EGNNs).[40] We build the k-nearest neighbors graph between ligand atoms, motifs with the condition pocket and protein atoms, and using message passing to model the interaction between them:

$$\begin{gathered} \mathrm{Mes}_{i,j} = \phi_{\mathrm{Mes}}\left(\boldsymbol{h}_i^{\,l}, \boldsymbol{h}_j^{\,l}, \boldsymbol{e}_{ij}^{\,l}, t, \|\boldsymbol{x}_i - \boldsymbol{x}_j\|\right), \\ \boldsymbol{h}_i^{(l+1)} = \phi_{\mathrm{f}}\left(\boldsymbol{h}_i^{\,l}, \sum_{j\in \mathrm{N}(i)} \mathrm{Mes}_{ij}\right). \Delta\boldsymbol{x}_i^{\,l+1} = \phi_{\mathrm{x}}\left(\boldsymbol{h}_i^{(l+1)}\right), \\ \boldsymbol{x}_i^{(l+1)} = \boldsymbol{x}_i^{\,l} + \Delta\boldsymbol{x}_i^{\,l+1}. \end{gathered} \tag{11}$$

where $\boldsymbol{h}_i$ and $\boldsymbol{h}_j$ represent for the embeddings of vertices in the constructed neighborhood graph. The variable $\boldsymbol{e}_{ij}$ indicates the edge type, which is determined by the type of vertices connected by the edge. The variable $l \in (0,L)$ represents the index of the EGNN layer, where L denotes the total number of layers. $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ represent the coordinates of the nodes. The networks ϕ_{Mes}, ϕ_{f}, and ϕ_{x} are utilized for computing the message, feature embedding, and position, respectively. At each time step, the EGNN simultaneously updates atomic and motif positions through equivariant transformations, maintaining geometric invariance while ensuring cross-view interactions between atomic and motif-level representations. After get the position of every node, we calculate the atom type and motif id using the type prediction networks ϕ_{v} and ϕ_{w}:

$$\boldsymbol{v}_{\mathrm{G}} = \phi_{\mathrm{v}}(\boldsymbol{h}_{\mathrm{G}}{}^{L}), \quad \boldsymbol{w} = \phi_{\mathrm{w}}(\boldsymbol{h}_{\mathrm{M}}{}^{L}). \tag{12}$$

For final molecular integration, we employ an atom-view rather than motif-based assembly, enabling the construction of more flexible structures unconstrained by predefined motif vocabulary. In our implementation, motif rotational degrees prediction is omitted since the combination of motif positions and their vocabulary IDs adequately captures both the structural features and protein-ligand interaction patterns.

5.4 Topological features

The geometric and topological information existed in molecule structure is the essential clue to understand the interaction between drug and protein. However, traditional general GNNs models often struggle to capture the intricate structures of 3D objects, such as cycles and cavities. To address this limitation, we employ persistent homology, a mathematical tool utilized in topological data analysis,[41–43] to extract stable topological features from pocket and ligand point clouds.

We utilize filtration functions, denoted as f, to calculate the persistence diagrams (PD) $\mathrm{ph}(\boldsymbol{x}, f) = \{\boldsymbol{D}_0, \ldots, \boldsymbol{D}_l\}$, where $\boldsymbol{D}_l$ is the l-dimensional diagram and x is the point cloud of pockets or ligands. The resulting PD reflects the multi-scale summarized topological information. Then we calculate the normalized persistent entropy from the PD through the method introduced in:[44]

$$E(\boldsymbol{D}_l) = -\sum_{d\in \boldsymbol{D}_l} \frac{\mathrm{pers}(d)}{\mathrm{E}(\boldsymbol{D}_l)} \log_2\left(\frac{\mathrm{pers}(d)}{\mathrm{E}(\boldsymbol{D}_l)}\right), \quad \mathrm{E}(\boldsymbol{D}_l) = \sum_{d\in \boldsymbol{D}_l} \mathrm{pers}(d),$$
$$E_{\mathrm{norm}}(\boldsymbol{D}_l) = \frac{E(\boldsymbol{D}_l)}{\log_2(\mathrm{E}(\boldsymbol{D}_l))}. \tag{13}$$

where d represents the persistence of points in the diagram $\boldsymbol{D}_l$, and pers(d) denotes the lifetime. The normalized persistent entropy $E_{\mathrm{norm}}(\boldsymbol{D}_l)$ quantifies the entropy of the persistence diagram. We use an embedding function ϕ_{d} to map persistence diagrams into vector representations in high-dimensional space $\phi_{\mathrm{d}} : \{\boldsymbol{D}_1, \ldots, \boldsymbol{D}_l\} \to \mathbb{R}^{n' \times d}$, where n' is the dimension of vector. Consequently, we can get the topological fingerprints of ligands and pockets:

$$\boldsymbol{F}_{\mathrm{G}} = \phi_{\mathrm{d}}(E_{\mathrm{norm}}(\mathrm{ph}(\boldsymbol{x}_{\mathrm{G}}))), \quad \boldsymbol{F}_{\mathrm{R}} = \phi_{\mathrm{d}}(E_{\mathrm{norm}}(\mathrm{ph}(\boldsymbol{x}_{\mathrm{R}}))). \tag{14}$$

The topological fingerprints is incorporated to pocket and ligand representations. $\boldsymbol{F}_{\mathrm{G}}$ and $\boldsymbol{F}_{\mathrm{R}}$ are concatenated with geometric and chemical features of the ligand and pocket to form the initial input representations. Then the features are also fed into the global context encoder, allowing the model to learn how topological patterns influence atomistic interactions. To further preserve critical global structures—such as rings or cavities—topological coherence constraints are incorporated during optimization. These constraints are implemented as regularization terms in the loss function, comparing the persistence diagram of the predicted molecule to that of the ground truth.

5.5 Training strategy

We construct a hierarchical structure consisting of atom-view and motif-view representations during the diffusion process. To train our model, we adopt a joint training approach that provides supervision for each view and their parameters updated simultaneously by a joint loss. In the atom-view, we employ two types of losses as introduced in 4.2: the position loss ($L_{\mathrm{a_{pos}}}$) and the atom type loss ($L_{\mathrm{a_{type}}}$). Similarly, in the motif-view, we incorporate the motif position loss ($L_{\mathrm{m_{pos}}}$) and the motif id loss ($L_{\mathrm{m_{id}}}$). Consequently, the final loss is determined as a weighted sum of the coordinate loss and type loss, expressed as follows:

$$L = L_{\mathrm{a_{pos}}} + \lambda_1 L_{\mathrm{a_{type}}} + L_{\mathrm{m_{pos}}} + \lambda_2 L_{\mathrm{m_{id}}}. \tag{15}$$

Here, λ_1 and λ_2 are parameters used to control the contribution of the loss terms in the overall loss function.

5.6 Medicinal chemistry filters

We utilize filters to refine our search process for preliminary candidates with favorable medicinal chemistry and structural novelty. First, we utilize structural filters to eliminate structures with problematic substructures. These substructures encompass promiscuously, reactive substructures (like pan-assay interference compounds, PAINS), pharmacokinetically unfavorable substructures (Brenk substructures),[45,46] and other alerts. We employ these filters to screen compounds and exclude those with potentially toxic structures and compounds containing undesirable groups. Second, we employ property filters to further refine our selection. These filters aim to eliminate compounds that are unlikely to exhibit optimal molecular properties, thus enhancing the quality of our candidate pool. Finally, we leverage the Tanimoto similarity metric to assess the resemblance of our compounds to bioactive molecules. Specifically, we calculate Tanimoto similarity scores to identify compounds with a high degree of similarity to known ALK ligands. By prioritizing compounds that demonstrate significant Tanimoto similarity to these reference ligands, we increase the likelihood of identifying candidates with potential targeted activity against ALK.

5.7 Statistical analysis

Following the works of[47] and,[16] we utilize the CrossDocked dataset[26] to train and evaluate our model. This dataset comprises a comprehensive collection of 22.5 million docked protein binding complexes. We drop all complexes that cannot be processed by the RDKit and filter out complexes with a binding pose Root Mean Square Deviation (RMSD) greater than 2 Å and a sequence identity lower than 40%. This filtering step aimed to remove complexes displaying significant structural deviations and low sequence similarity. After applying these filters, we obtain a subset of 100 000 complexes for training our model, while reserving an additional 100 proteins for testing purposes. For real-word therapeutic targets, we download the protein structures form PDB,[48] and the active molecules corresponding to these targets were downloaded from BindingDB.[49] To extract structural motifs from the training set, we follow.[16] The motif vocabulary construction process involves decomposing molecules at rotatable bonds (bonds whose cutting creates two valid components with ≥2 atoms each). Substructures that appear more frequently than threshold τ in the training set are selected as motifs. If a substructure doesn't meet the frequency threshold, it's further broken down into smaller rings and bonds. The final vocabulary size can be controlled by τ. Our implementation uses

a vocabulary of 520 motifs. Detailed training procedures and hyperparameter settings are provided in the SI D.

In this work, we apply the widely used metrics of deep generative models to evaluate the performance of our method. (1) Validity measures the percentage of generated molecules that successfully pass the RDKit sanitization check. (2) Uniqueness is assessed by calculating the percentage of unique structures among the generated outputs. (3) Diversity considers the proportion of unique scaffold structures generated, as well as the internal diversity values calculated for 2D and 3D structures using the Morgan fingerprint and USRCAT, respectively. (5) Novelty measures the proportion of unique scaffold structures generated that are not present in either the test set or the known ligand set. (6) Molecular properties. We report the distributions of several important 2D and 3D molecular properties, including molecular weight (MW), QED, Synthetic Accessibility (SA), log P, normalized principal moment of inertia ratios (NPR1 and NPR2). SA is a widely used computational metric that estimates how easily a molecule can be synthesized based on its structural features. These distributions are compared to those of the train and test set to assess the model's ability to learn important molecular properties. (6) Affinity: the average binding affinity of the generated molecules, which is assumed to be characterized by the docking score. The value in kcal mol^{-1} is estimated by AutoDock Vina. (7) RMSD stands for root-mean-square deviation, which is a metric used to measure the dissimilarity between different conformations of the same molecule. A smaller RMSD value indicates a higher degree of similarity between the conformations. The formula for calculating RMSD is as follows:

$$\mathrm{RMSD}\left(\hat{\boldsymbol{R}}, \boldsymbol{R}\right) = \left(\frac{1}{n}\sum_{i=1}^{n}\left\|\hat{\boldsymbol{R}}_i - \boldsymbol{R}_i\right\|^2\right)^{\frac{1}{2}}. \quad (16)$$

Here, n represents the number of atoms, $\hat{\boldsymbol{R}}$ represents the conformation of the generated molecule, and $\hat{\boldsymbol{R}}_i$ represents the Cartesian coordinates of the i-th atom. We use the RMSD calculation to assess the conformational changes that occur before and after the docking process.

We conducted a comparative analysis with several baselines: liGAN[50] is a method that combines a 3D CNN architecture with a conditional variational autoencoder (VAE), enabling the generation of molecular structures using atomic density grids. AR[47] introduces an auto-regressive sampling scheme to estimates the probability density of atom occurrences in 3D space. The atoms are sampled sequentially from the learned distribution until there is no room within the binding pocket. Pocket2Mol[51] design a new graph neural network capturing both spatial and bonding relationships between atoms of the binding pockets, then samples new drug candidates conditioned on the pocket representations from a tractable distribution. GraphBP[15] use a flow model to generate the type and relative location of new atoms. It also obtains geometry-aware and chemically informative representations from the intermediate contextual information. TargetDiff[19] is a 3D SE(3)-equivariant diffusion model that jointly generates atomic coordinates and atom types in a non-autoregressive manner for target-aware molecular design. DecompDiff[20] is an end-to-end diffusion-based method that utilizes decomposed priors and validity guidance to generate atoms and bonds of 3D ligand molecules. FLAG[16] constructs a motif vocabulary by extracting common molecular fragments from the dataset. The model selects the focal motif, predicts the next motif type, and attaches the new motif. DiffBP[52] proposed a diffusion model that generates molecular 3D structures by simultaneously denoising both atomic element types and 3D coordinates using an equivariant network architecture.

Author contributions

Conceptualization: Guanlue Li, Chenran Jiang. Methodology: Guanlue Li, Chenran Jiang, Jia Li. Investigation: Guanlue Li, Yu Liu, Jiean Chen. Visualization: Guanlue Li, Chenran Jiang. Supervision: Yong Huang, Jia Li. Writing—original draft: Guanlue Li, Chenran Jiang. Writing—review & editing: Chenran Jiang, Yong Huang, Jia Li, Ziqi Gao.

Conflicts of interest

There are no conflicts to declare.

Data availability

The train and test dataset CorssDocked is available at the dataset link. Instructions and data loaders are provided *via* the associated GitHub repository. Case studies targeting protein kinases, including Anaplastic Lymphoma Kinase (ALK) and Cyclin-dependent kinase 4 (CDK4), can be found on the RCSB Protein Data Bank (PDB) website at **https://www.rcsb.org/**.

The SI file provides a detailed description of the data collection and preprocessing steps, together with illustrations of the network architecture, motif representation, model training procedure, and hyperparameter settings. It also reports extended results across several dimensions, including performance on multiple targets, sensitivity to motif vocabulary size, evaluation on the PoseBusters test, and analyses of discrete diffusion transition matrices. Finally, it includes a comprehensive description of the algorithm used in this work. Supplementary information is available. See DOI: **https://doi.org/10.1039/d5sc02113h**.

Code availability: the source code for model architecture is publicly available on our GitHub repository **https://github.com/guanlueli/AMDiff**.

Acknowledgements

This work was supported by Guangdong S&T "1+1+1" Joint Funding Program C019.

References

1 J. G. Cumming, A. M. Davis, S. Muresan, M. Haeberlein and H. Chen, *Nat. Rev. Drug Discovery*, 2013, **12**, 948–962.

2 T. Hou and J. Wang, *Expert Opin. Drug Metab. Toxicol.*, 2008, **4**, 759–770.
3 J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, *et al.*, *Nature*, 2021, **596**, 583–589.
4 R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams and A. Aspuru-Guzik, *ACS Cent. Sci.*, 2018, **4**, 268–276.
5 P. G. Polishchuk, T. I. Madzhidov and A. Varnek, *J. Comput.-Aided Mol. Des.*, 2013, **27**, 675–679.
6 F. Imrie, T. E. Hadfield, A. R. Bradley and C. M. Deane, *Chem. Sci.*, 2021, **12**, 14577–14589.
7 M. Xu, T. Ran and H. Chen, *J. Chem. Inf. Model.*, 2021, **61**, 3240–3254.
8 C. Isert, K. Atz and G. Schneider, *Curr. Opin. Struct. Biol.*, 2023, **79**, 102548.
9 A. Tripathi and V. A. Bankaitis, *Journal Of Molecular Medicine And Clinical Applications*, 2017, **2**, 04–05.
10 W. Xie, F. Wang, Y. Li, L. Lai and J. Pei, *J. Chem. Inf. Model.*, 2022, **62**, 2269–2279.
11 D. M. Anstine and O. Isayev, *J. Am. Chem. Soc.*, 2023, **145**, 8736–8750.
12 J. C. Fromer and C. W. Coley, *Patterns*, 2023, **4**, 03–04.
13 B. A. Garcia, S. Mollah, B. M. Ueberheide, S. A. Busby, T. L. Muratore, J. Shabanowitz and D. F. Hunt, *Nat. Protoc.*, 2007, **2**, 933–938.
14 H. Zhao, *Drug discovery today*, 2007, **12**, 149–155.
15 M. Liu, Y. Luo, K. Uchino, K. Maruhashi and S. Ji, *International Conference on Machine Learning*, 2022, pp. 13912–13924.
16 Z. Zhang, Y. Min, S. Zheng and Q. Liu, *The Eleventh International Conference on Learning Representations*, 2022.
17 C. Hu, S. Li, C. Yang, J. Chen, Y. Xiong, G. Fan, H. Liu and L. Hong, *J. Cheminf.*, 2023, **15**, 91.
18 Y. Zhu, Y. Du, Y. Wang, Y. Xu, J. Zhang, Q. Liu and S. Wu, *Learning on Graphs Conference*, 2022, pp. 47–1.
19 J. Guan, W. W. Qian, X. Peng, Y. Su, J. Peng and J. Ma, *International Conference on Learning Representations*, 2023.
20 J. Guan, X. Zhou, Y. Yang, Y. Bao, J. Peng, J. Ma, Q. Liu, L. Wang and Q. Gu, *International Conference on Machine Learning*, 2023, pp. 11827–11846.
21 O. Zhang, T. Wang, G. Weng, D. Jiang, N. Wang, X. Wang, H. Zhao, J. Wu, E. Wang, G. Chen, *et al.*, *Nat. Comput. Sci.*, 2023, 1–11.
22 O. Zhang, J. Zhang, J. Jin, X. Zhang, R. Hu, C. Shen, H. Cao, H. Du, Y. Kang, Y. Deng, *et al.*, *Nat. Mach. Intell.*, 2023, **5**, 1020–1030.
23 H. Lin, Y. Huang, O. Zhang, S. Ma, M. Liu, X. Li, L. Wu, J. Wang, T. Hou and S. Z. Li, *Chem. Sci.*, 2025, **16**, 1417–1431.
24 A. S. Powers, H. H. Yu, P. Suriana, R. V. Koodli, T. Lu, J. M. Paggi and R. O. Dror, *ACS Cent. Sci.*, 2023, **9**, 2257–2267.
25 S. Wills, R. Sanchez-Garcia, T. Dudgeon, S. D. Roughley, A. Merritt, R. E. Hubbard, J. Davidson, F. von Delft and C. M. Deane, *J. Chem. Inf. Model.*, 2023, **63**, 3423–3437.
26 P. G. Francoeur, T. Masuda, J. Sunseri, A. Jia, R. B. Iovanisci, I. Snyder and D. R. Koes, *J. Chem. Inf. Model.*, 2020, **60**, 4200–4215.
27 Z. Gao, C. Jiang, J. Zhang, X. Jiang, L. Li, P. Zhao, H. Yang, Y. Huang and J. Li, *Nat. Commun.*, 2023, **14**, 1093.
28 P. Pan, H. Yu, Q. Liu, X. Kong, H. Chen, J. Chen, Q. Liu, D. Li, Y. Kang, H. Sun, *et al.*, *ACS Cent. Sci.*, 2017, **3**, 1208–1220.
29 C. Jiang, Y. Ye, W. Kang, J. Yang, Z. He, Q. Cao, C. Lian, Y. Xing, Q. Yang and J. Zhao, *J. Med. Chem.*, 2025, **68**, 1499–1510.
30 A. M. Schreyer and T. Blundell, *J. Cheminf.*, 2012, **4**, 1–12.
31 L. Wang, H. Liu, Y. Liu, J. Kurtin and S. Ji, *International Conference on Learning Representations*, 2023.
32 B. Qiang, Y. Song, M. Xu, J. Gong, B. Gao, H. Zhou, W.-Y. Ma and Y. Lan, *International Conference on Machine Learning*, 2023, pp. 28277–28299.
33 W. Jin, R. Barzilay and T. Jaakkola, *International conference on machine learning*, 2018, pp. 2323–2332.
34 G. Corso, H. Stärk, B. Jing, R. Barzilay and T. Jaakkola, *International Conference on Learning Representations*, 2023.
35 J. Ding, S. Tang, Z. Mei, L. Wang, Q. Huang, H. Hu, M. Ling and J. Wu, *J. Chem. Inf. Model.*, 2023, **63**, 1982–1998.
36 M. Xue, B. Liu, S. Cao and X. Huang, *npj Drug Discovery*, 2025, **2**, 4.
37 J. Ho, A. Jain and P. Abbeel, *Adv. Neural Inf. Process. Syst.*, 2020, **33**, 6840–6851.
38 E. Hoogeboom, D. Nielsen, P. Jaini, P. Forré and M. Welling, *Adv. Neural Inf. Process. Syst.*, 2021, **34**, 12454–12465.
39 J. Austin, D. D. Johnson, J. Ho, D. Tarlow and R. Van Den Berg, *Adv. Neural Inf. Process. Syst.*, 2021, **34**, 17981–17993.
40 V. G. Satorras, E. Hoogeboom and M. Welling, *International conference on machine learning*, 2021, pp. 9323–9332.
41 Z. Meng and K. Xia, *Sci. Adv.*, 2021, **7**, eabc5329.
42 X. Liu, H. Feng, Z. Lü and K. Xia, *Briefings Bioinf.*, 2023, **24**, bbad046.
43 D. D. Nguyen, Z. Cang and G.-W. Wei, *Phys. Chem. Chem. Phys.*, 2020, **22**, 4343–4367.
44 A. Myers, E. Munch and F. A. Khasawneh, *Phys. Rev. E*, 2019, **100**, 022314.
45 J. B. Baell and G. A. Holloway, *J. Med. Chem.*, 2010, **53**, 2719–2740.
46 R. Brenk, A. Schipani, D. James, A. Krasowski, I. H. Gilbert, J. Frearson and P. G. Wyatt, *ChemMedChem*, 2008, **3**, 435–444.
47 S. Luo, J. Guan, J. Ma and J. Peng, *Adv. Neural Inf. Process. Syst.*, 2021, **34**, 6229–6239.
48 J. L. Sussman, D. Lin, J. Jiang, N. O. Manning, J. Prilusky, O. Ritter and E. E. Abola, *Acta Crystallogr., Sect. D:Biol. Crystallogr.*, 1998, **54**, 1078–1084.
49 T. Liu, Y. Lin, X. Wen, R. N. Jorissen and M. K. Gilson, *Nucleic Acids Res.*, 2007, **35**, D198–D201.
50 M. Ragoza, T. Masuda and D. R. Koes, *Chem. Sci.*, 2022, **13**, 2701–2713.
51 X. Peng, S. Luo, J. Guan, Q. Xie, J. Peng and J. Ma, *International Conference on Machine Learning*, 2022, pp. 17644–17655.
52 H. Lin, Y. Huang, O. Zhang, S. Ma, M. Liu, X. Li, L. Wu, J. Wang, T. Hou and S. Z. Li, *Chem. Sci.*, 2025, **16**, 1417–1431.