Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal | View Issue

Cite this: *Chem. Sci.*, 2024, **15**, 19452

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

Received 11th July 2024
Accepted 11th October 2024

DOI: 10.1039/d4sc04620j

rsc.li/chemical-science

FragGen: towards 3D geometry reliable fragment-based molecular generation†

Odin Zhang, ‡[a] Yufei Huang,‡[b] Shichen Cheng, ‡[a] Mengyao Yu,‡[a] Xujun Zhang,[a] Haitao Lin,[b] Yundian Zeng,[a] Mingyang Wang,[a] Zhenxing Wu,[a] Huifeng Zhao,[a] Zaixi Zhang, [c] Chenqing Hua,[d] Yu Kang, [a] Sunliang Cui, *[a] Peichen Pan, *[a] Chang-Yu Hsieh *[a] and Tingjun Hou *[a]

3D structure-based molecular generation is a successful application of generative AI in drug discovery. Most earlier models follow an atom-wise paradigm, generating molecules with good docking scores but poor molecular properties (like synthesizability and drugability). In contrast, fragment-wise generation offers a promising alternative by assembling chemically viable fragments. However, the co-design of plausible chemical and geometrical structures is still challenging, as evidenced by existing models. To address this, we introduce the Deep Geometry Handling protocol, which decomposes the entire geometry into multiple sets of geometric variables, looking beyond model architecture design. Drawing from a newly defined six-category taxonomy, we propose FragGen, a novel hybrid strategy as the first geometry-reliable, fragment-wise molecular generation method. FragGen significantly enhances both the geometric quality and synthesizability of the generated molecules, overcoming major limitations of previous models. Moreover, FragGen has been successfully applied in real-world scenarios, notably in designing type II kinase inhibitors at the ~nM level, establishing it as the first validated 3D fragment-based drug design algorithm. We believe that this concept-algorithm-application cycle will not only inspire researchers working on other geometry-centric tasks to move beyond architecture designs but also provide a solid example of how generative AI can be customized for drug design.

Introduction

Despite the emergence of a plethora of novel modalities in the past decade, designing druggable molecules that target functional proteins remains the most effective treatment option. Empowered by the rapid advancement of artificial intelligence (AI)-aided drug design (AIDD),[1] our ability to discover suitable organic-molecule-based drug candidates has been dramatically enhanced. The ambitious endeavor of computer-aided drug discovery primarily bifurcates into two streams: virtual screening, which involves sifting through existing molecular libraries,[2] and molecular generation, which entails crafting molecules from scratch.[3] The former, essentially a classification task, has seen significant development over the past decade in the AI landscape, exemplified by advancements in scoring functions.[4] On the other hand, the latter has been synergized with the language and graph generation methods, leading to SMILES-based[5] and graph-based molecular generation models,[6] bringing in fresh computational perspectives to drug discovery. Despite the progress in AIDD, the absence of any AI-designed drugs passing regulatory approval highlights the formidable challenge of data-driven drug design. A key issue is data sparsity, a domain-specific obstacle that does not severely affect other fields like image or language processing where extensive data is available. In drug discovery, limited datasets are common due to the high costs and complexity of drug development, confidentiality in pharmaceutical research, and the vastly complex functioning principles of biological systems.[7] Data scarcity restricts the potential and applicability of many advanced AI models that have previously been proven successful in data-rich environments. Thus, external assistance,

[a]*College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, Zhejiang, China. E-mail: slcui@zju.edu.cn; panpeichen@zju.edu.cn; kimhsieh@zju.edu.cn; tingjunhou@zju.edu.cn*

[b]*Zhejiang University, Hangzhou 310058, Zhejiang, China*

[c]*Anhui Province Key Lab of Big Data Analysis and Application, University of Science and Technology of China, Hefei, Anhui, China*

[d]*Montreal Institute for Learning Algorithms, McGill University, Montreal, QC, Canada*

† Electronic supplementary information (ESI) available: Part S1. The detailed architectures of several models. Part S2. Additional results of retrospective studies on three well-studied targets. Part S3. Ablation study of geometry handling protocols in FragGen. Part S4. Synthesis routes and molecular characterization of validated compounds. Fig. S1. Fragment decomposition of crystal ligand and FragGen's top generated molecules. Fig. S2. Illustration of ablation studies. Table S1. The Top5 molecules mean binding energies and drug-like properties across three well-studied targets; Table S2. The ablation results of three geometry handling protocols in FragGen. See DOI: https://doi.org/10.1039/d4sc04620j

‡ Equivalent authors.

particularly in the form of physical constraints, becomes crucial to mitigate this intrinsic challenge by introducing prior knowledge to restrain the solution space. The rapid development and impressive performance of AlphaFold[8] and other structure-related models[9] underscore the efficacy of this approach. Concurrently, there is a growing emphasis on structure-based methodologies in both virtual screening and molecular generation, opening up new frontiers and challenges, such as binding conformation prediction[10] and pocket-aware molecular generation.[11]

In the realm of 3D pocket-aware molecular generation, recent years have witnessed the emergence of many promising models like LiGAN,[12] Pkt2Mol,[13] DiffBP,[14] ResGen,[15] *etc.*, which have manifested varying degrees of success in generating potentially superior ligands with a lower binding energy (as estimated by docking scores) than the reference ligands. However, a closer inspection on the generated ligands, particularly before any post-processing, reveals two critical limitations of most existing models. Firstly, the generated molecular conformations often appear distorted, which is noted in the outputs of GraphBP[16] and DiffBP (Fig. 1). Secondly, there is a tendency to produce molecules with multi-fused rings to fill the cavity of protein pockets, which is observed in the outputs of Pkt2Mol and ResGen (Fig. 1). While these generated structures may induce stronger interactions with protein pockets, they either look physically implausible or the complex structure poses significant challenges in synthesis and often results in toxic properties, thus actually distancing them from ideal drug candidates. Fragment-wise molecular generation offers a solution by assembling a molecule from synthesizable fragments as basic elements, as illustrated in previous Reinforcement-Learning-based methods such as DeepFMPO.[17] However, the only existing generative implementation of this approach, *i.e.*, FLAG,[18] encounters significant challenges with geometry handling as illustrated in Fig. 1. The error in each fragment generation step accumulates, ultimately causing the collapse of the molecular structure. Therefore, there is a pressing need for a reliable fragment-wise deep generative model in structure-based drug design (SBDD).

Rendering smooth geometries is a central focus of the computational study of physical reality, not just for 3D molecular generation but across almost all geometry-centric application domains. For instance, in molecular conformation generation, researchers[19] have adopted the distance-then-geometry protocol first to generate distance matrices and then deduce Cartesian coordinates by optimizing randomly initialized conformations under the distance constraint. However, the non-uniqueness in mapping under-specified distance matrices to Cartesian coordinates often introduces additional errors, leading to geometric distortions. Subsequent research[20,21] has explored force-field optimization or end-to-end Cartesian coordinate prediction to enhance a deep learning model's capability to generate accurate geometry. In addition to efforts on the direct generations of plausible molecular conformations, deep learning has also concurrently made significant advancements on the front of molecular docking. Early models, such as TANKBind,[22] extended the idea of distance-then-geometry protocol to protein–ligand binding conformation prediction. However, the incorporation of protein nodes into these models introduced a formidable challenge: a significant increase in redundant degrees of freedom, which led to unsatisfactory geometries. Then researchers delved into the end-to-end solutions, directly predicting the Cartesian coordinates, as pioneered by EquiBind.[23] KarmaDock[24] further advanced this protocol by employing a recycling mechanism, emulating the classical geometry optimization, and finally raising the successful rate of docking by about 50%. Yet, all these methods still struggle with the generation of unrealistic local structures,

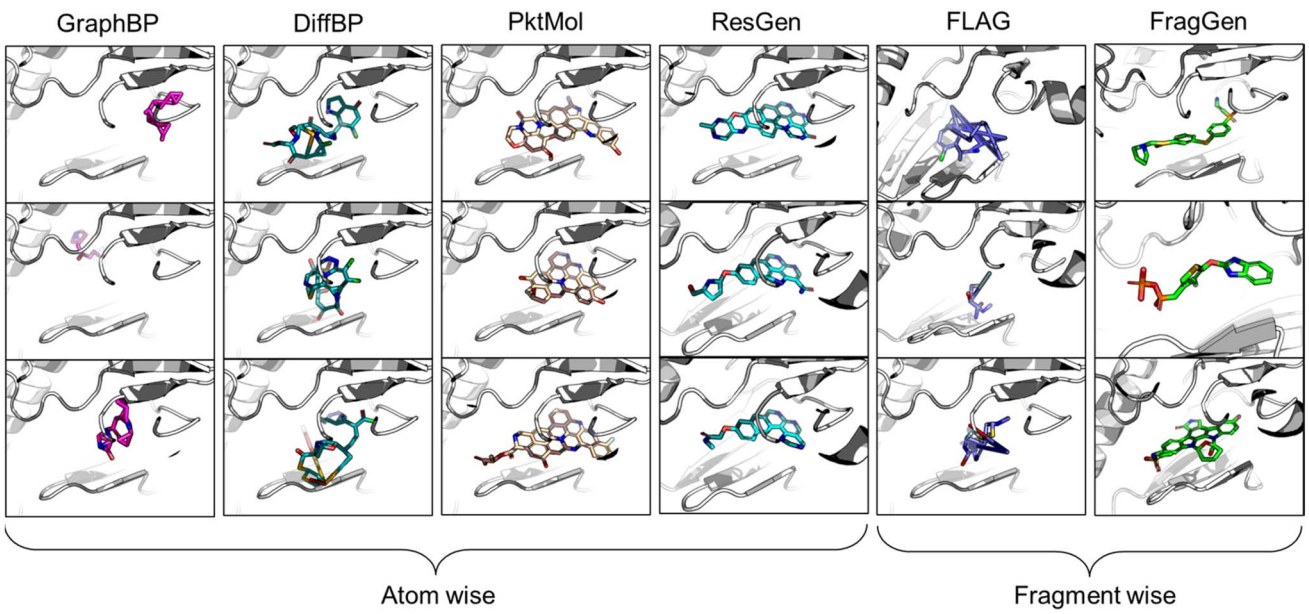

Fig. 1 Visualized molecules generated by different methods. All models are performed without force-field optimization.

such as non-coplanar aromatic rings and excessively long chemical bonds, necessitating post-processing steps like geometry optimization or alignment corrections. DiffDock[25] represents a different technical approach, focusing on tuning constrained variables like overall translation, orientation, and torsion angles in order to simplify the morphing of molecular conformations. DiffDock's idea works well as it improves the state of geometric plausibility of deep-learning-based generations, though its generated ligands may still encounter clashes with protein pocket residues.

The challenges in correctly handling geometry with deep learning models are twofold: the inherent symmetries in geometric variables (illustrated in Fig. 2A) and in which way the geometry is constructed. The first aspect, symmetry considerations, like SE(3)-invariance/equivariance, has been thoroughly addressed. Many works have concentrated on enhancing the feature extraction capability of models while enforcing adherence to the necessary equivariance or invariance principles. For example, the transformation of Cartesian coordinates should comply with roto-translational equivariance, which is mathematically expressed as $Rf(x) + t = f(Rx + t)$, where R and t represents the rotation matrix and translation vector, respectively, f denotes the neural network function. However, the second aspect, the high-level geometric handling protocol, has not received as much attention compared to the development of symmetry-focused architectural designs, as exemplified by models such as EGNN,[26] SchNet,[27] and Geodesic-GNN.[28] While computational scientists, (when first entering into a new field such as drug design) would tend to tinkle with model architectures in order to attain better performance under the existing practices (for instance, a given geometric protocol), it is crucial to recognize that the protocol itself should also be re-assessed if a substantial breakthrough is the goal. The selected protocol sets the performance boundary of a model and significantly dictate the outcome. Therefore, we advocate that a thorough review and re-thinking of existing geometric handling protocols are imperative.

In light of these observations, we first review and summarize six protocols that could be used in 3D molecular generation, highlighting their respective challenges and discussing their usages in other molecular geometry-centric problems, like molecular conformation generation and docking problems. Building on this foundation, we propose a hybrid approach that employs multiple protocols and effectively draws upon the unique strength of each one to achieve an optimal performance in 3D molecular generation, as highlighted in Fig. 2C. This

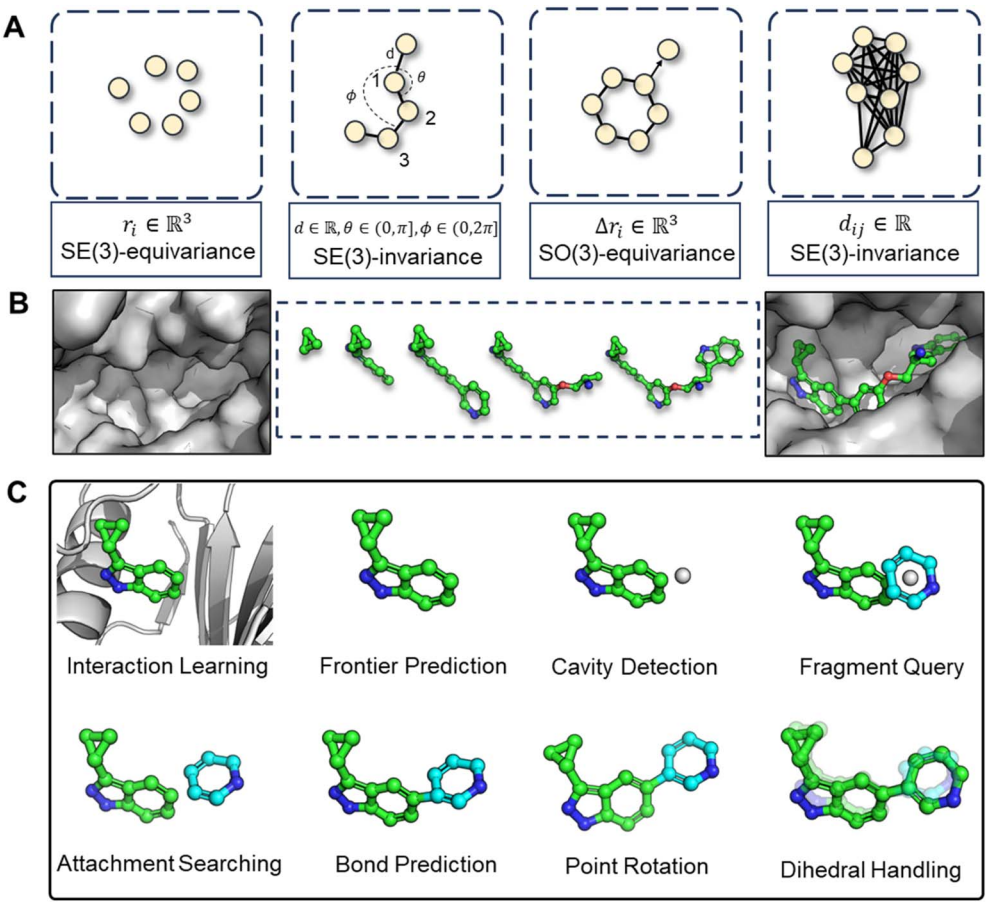

Fig. 2 (A) Illustration of symmetry requirements for various geometric variables. (B) Structure-aware and fragment-wise molecular generation (C). Workflow of our proposed combined geometry handling protocol, which is specifically designed for 3D fragment-wise molecular generation.

novel strategy led to the development of the first geometry-reliable and fragment-wise molecular deep generation, FragGen as presented in Fig. 2B. It achieves state-of-the-art performance in our reported experiments and validates our argument on the need to re-formulate the geometry handling protocol. Furthermore, we grounded our algorithmic development into real-world drug design campaigns, successfully designing potent type II inhibitors (75.9 nM) targeting the leukocyte receptor tyrosine kinase. To our best knowledge, this is the first successful application of 3D fragment-based molecular generation methods. This concept-algorithm-application work not only serves as a SOTA drug design tool but also enriches the discourse on geometric handling protocols, complementing symmetric neural network design and offering a blueprint for model development for other geometry-related fields.

Results and discussions

Analysis of geometry handling protocols

The continuing advancement of structural predictions for various biomolecules, exemplified by AlphaFold, has drawn the AI community's attention onto structure-based drug design, where accurately modeling molecular geometry plays a pivotal role in estimating drug–target interactions. In this context, we meticulously examine six universal geometry handling protocols, as depicted in Fig. 3, underscoring the unique challenges each of them encounters in the context of pocket-aware 3D molecular design.

The Internal Coordinate protocol, which initially determines four atomic orders before predicting bond lengths, angles, and dihedral angles, often leads to distorted molecular conformations. This protocol is adopted by the GraphBP method (Fig. 3), whose errors have been found to predominantly arise from incorrect determination of the initial topological order, which is inherently difficult to determine within protein pockets. Unlike structure-free models like G-SphereNet,[29] where topological orders naturally follow generation trajectories in the ligand-only scenarios, the application of Internal Coordinate protocol in pocket-aware context struggles in the more complex environments, such as the protein pockets. In contrast, the Cartesian coordinate approach, which involves probabilistic learning directly on 3D coordinates, lacks local structural constraints. This often results in the accumulation of errors at each atomic position, leading to implausible geometries, such as non-coplanar rings or benzene rings with unequal bond lengths (Fig. 3). This challenge is prevalent in diffusion model-based methods like DiffBP and DiffSBDD,[30] which generate molecules in one shot. The Relative Vector protocol, predicting coordinate vector differences between atoms, appears more robust. Ensuring that the predicted 3D vector satisfies SE(3)-equivariance, this method effectively confines the degrees of freedom to bond lengths, thereby minimizing the impact of prediction errors on overall geometry. Methods like Pocket2Mol and ResGen, which employ this protocol, have achieved more rational generation of conformations. However, they still face challenges, particularly in generating multi-fused ring molecules that, while favoring stronger protein pocket interactions, are complex and difficult to synthesize.

The GeomGNN approach, utilized in KarmaDock, leverages equivariant graph neural networks to learn atomic forces, followed by a coarse coordinate update ($x_i = x_{i-1} + F_i$). This protocol benefits from straightforward training and inference, as it avoids complex transitions between different coordinate descriptions. Our implementation in the 3D molecular generation problem, resulting in FragGen-GNN, demonstrates this advantage. However, it also exhibits limitations in achieving precise atom localization. GeomOPT, a classical method for determining next atom or fragment coordinates, theoretically avoids local structure implausibility through force-field interactions involving bond angles and dihedrals. Despite its potential, this protocol faces significant limitations, including lengthy optimization times and a tendency for structures to become trapped in local minima, leading to twisted molecular structures, as shown in Fig. 3. Distance Geometry, another recognized approach used by models in conformation generation, such as ConfGF[31] and SDEGen,[20] circumvents equivariance demands in neural network design by modeling interatomic distances. This reduces model construction complexity but suffers from an overabundance of degrees of freedom, making it impossible to uniquely determine 3D coordinates from a distance matrix. Consequently, even with a perfectly predicted distance matrix, accurate reconstruction of original Cartesian coordinates remains elusive, often resulting in distorted conformations, as seen with the FLAG method (Fig. 3).

While ongoing advancements in model architecture design strive for improved performance, they do not directly address the inherent challenges of each geometry protocol summarized above. Recognizing this lack of algorithmic development on an equally important issue that contributes to the overall quality of generated conformations, this work sets out to improve the existing protocol and propose a combined strategy which integrates insights emerged from our systematic investigation on the pros and cons of each existing protocol.

More specifically, the combined strategy works as follows. We first utilize the Relative Vector protocol for sub-pocket detection, determining suitable locations for subsequent fragment assembly. Upon predicting the next fragment type, its geometry is decomposed into local conformation, rotation around a point (connected atom), and rotation around an axis (connected bond). Traditional methods and deep learning approaches generally perform well for local fragment geometries. For rotations around a point, we apply hybrid orbital theory constraints,[32] such as the consistent bond angles in standard SP3 hybridization (*e.g.*, 109.5° in methane), to guide the molecular assembly with chemical initialization founded on rigorous theoretical insights. Finally, for rotation around an axis, we directly predict dihedral angles using von Mises loss, more details can be found in method part. This decoupling of complex fragment-wise generation geometry has led to an effective solution, with subsequent experiments providing strong validation of our approach.

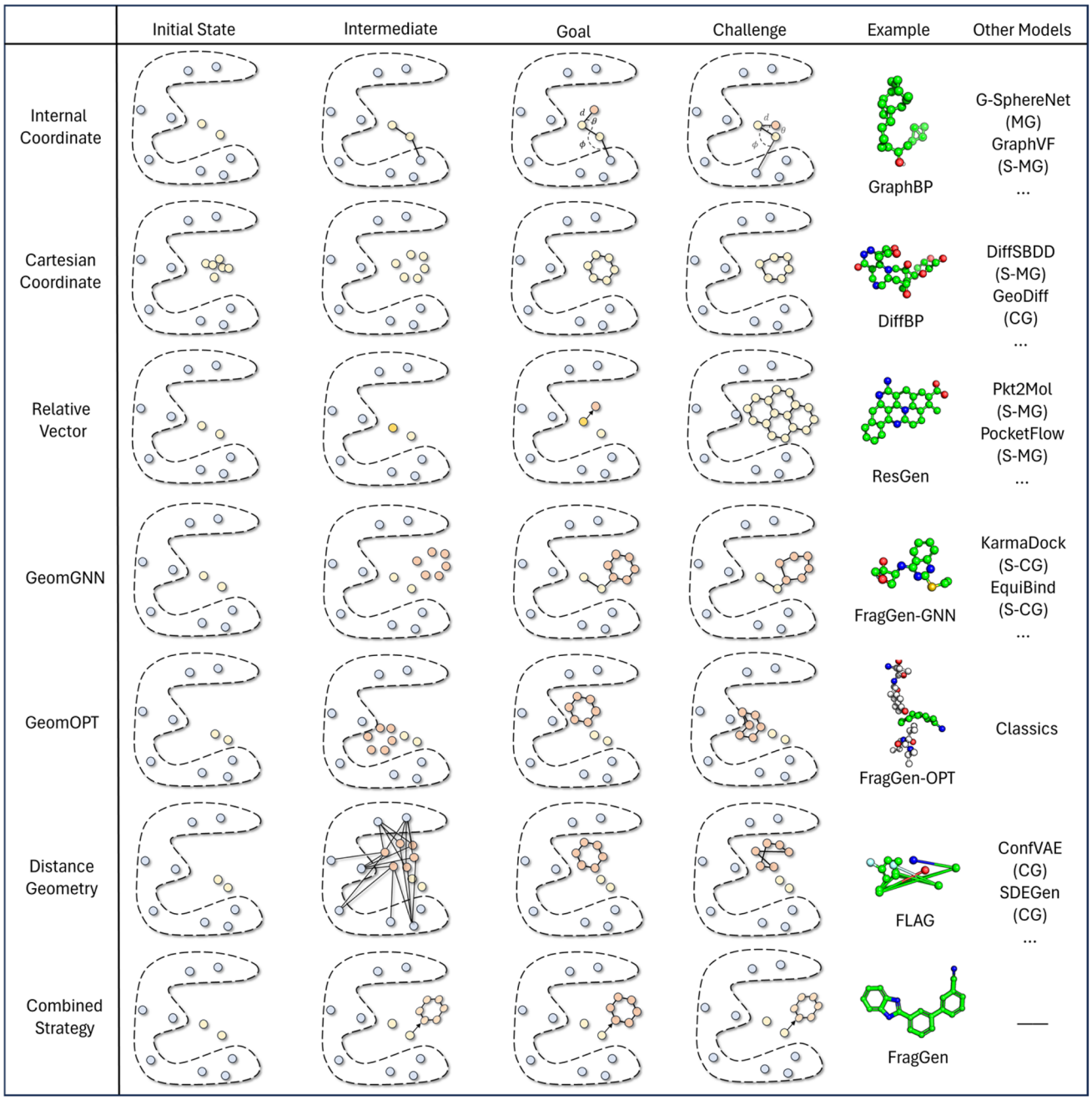

Fig. 3 This figure presents a comparative illustration of workflows, challenges, objectives, and implementations across different geometry handling protocols. The 'example' column focuses on applications within the field of 3D molecular generation, while the 'other models' column spans a broader range of geometry-centric topics. Key abbreviations include MG: Molecular Generation (without structures), S-MG: Structure-based Molecular Generation, CG: Conformation Generation (without structures), and S-CG: Structure-based Conformation Generation (also known as Docking).

Performance of FragGen on the CrossDock benchmark

Leveraging our novel geometry handling protocol, we developed FragGen, a structure-based, fragment-wise molecular generation method. Its efficacy was rigorously tested using the widely recognized CrossDock dataset,[33] a benchmark in previous atom-wise molecular generation research.[12–16] The evaluation involved calculating the Vina Score with AutoDock Vina[34] to gauge the ligand's binding affinity to its target protein. The Hit Pocket refers to the ratio of binding pockets where a molecular generation method produces molecules that bind tighter than a reference molecule. Additionally, other critical metrics are also included, such as the Quantitative Estimation of Drug-likeness (QED),[35] Synthetic Accessibility (SA),[36] Lipinski's Rule of Five,[37] and the octanol–water partition coefficient (Log P), to characterize the properties of the molecules generated. Notably, SA emerged as a crucial metric in contrasting atom-wise and

Table 1 The mean binding energies and drug-like properties for Top1/5 molecules

	Test set	GraphBP	DiffBP	Pocket2Mol	ResGen	FLAG	FragGen
Top1							
Vina score (↓)	−7.158	−9.332	−9.237	−9.247	−9.622	−8.954	−9.926
Hit pocket	—	87.07%	9.42%	92.10%	93.15%	87.14%	96.15%
QED (↑)	0.531	0.560	0.479	**0.562**	0.536	0.552	0.541
SA (↑)	0.730	0.464	0.411	0.341	0.307	0.565	0.740
Lipinski (↑)	4.684	4.821	4.734	4.921	**4.958**	4.955	4.871
Log *P*	0.947	1.552	0.452	0.8249	1.891	0.746	0.154
Top5							
Vina score (↓)	−7.158	−8.515	−8.723	−8.924	−9.343	−8.188	−9.654
QED (↑)	0.531	**0.563**	0.492	0.571	0.546	0.522	0.573
SA (↑)	0.730	0.478	0.433	0.346	0.316	0.582	0.717
Lipinski (↑)	4.684	4.776	4.788	4.931	4.953	4.975	4.859
Log *P*	0.947	1.430	0.457	0.758	1.646	0.451	1.273

fragment-wise methodologies, with the latter typically yielding higher SA due to the assembly of existing commercial fragments. Our baseline models included four atom-wise molecular generation approaches (GraphBP, DiffBP, Pkt2Mol, and ResGen) and one fragment-wise model FLAG, the only open-source model of its kind. The performance metrics for each model are detailed in Table 1.

From the results in Table 1, FragGen outperforms other methods in Vina Score, ranking as follows: FragGen > ResGen > Pkt2Mol > GraphBP > DiffBP > FLAG. FragGen leads with a Vina Score 2.5 kcal mol^{-1} higher than the test set average, translating to over 100-fold increase in binding affinity based on the thermodynamic principles.[15] This significant boost in binding potency is almost enough to elevate a ligand from μM IC_{50} to nM IC_{50}. Furthermore, FragGen excels in generating high-quality ligands with superior chemical and geometric structures. As illustrated in Fig. 1, atom-wise methods like GraphBP and DiffBP often yield distorted molecular geometries, with some GraphBP-generated molecules even straying out of the target pockets. These flawed geometries stem from the limitations of the Internal coordinate and Cartesian coordinate protocols, where the latter necessitates predefined topological atomic orders, and the former lacks local structural constraints to guide the generative process. In contrast, ResGen and Pkt2Mol, employing the Relative Vector protocol, achieve more accurate and visually rational molecular geometries. FLAG and FragGen, both fragment-wise approaches, turn out to give outputs that sits on opposite ends of the Vina Score spectrum (FLAG: ∼−8.9 *vs.* FragGen: ∼−9.9), a testament to their geometry handling capabilities. FLAG, based on Distance Geometry, often struggles with ill-structured molecules due to the challenges in mapping an extensive number of pairwise distances to Cartesian coordinates. Conversely, FragGen employs a sophisticated geometry handling approach, decomposed into four geometric variables and effectively managed through a blend of chemical knowledge and end-to-end learning. To be more specific, the four geometric components in FragGen are Cavity detection, Bond linking, Chemical initialization, and Dihedral handling, which are comprehensively explained in the Method section.

Regarding molecular properties, FragGen achieves the highest scores in QED and SA on the Top-5 results, underscoring the chemical viability of its generated molecules. These impressive results stem from two key factors: the inherent nature of the fragment-wise protocol and the advantages of a robust geometry handling approach. The fragment-wise protocol inherently guarantees better synthesizability, as it typically decomposes molecules into a set of existing fragments, also explaining FLAG's relatively high SA score. In contrast, atom-wise methods like Pkt2Mol and ResGen often generate molecules that completely fill the cavity of protein pockets, resulting in lower QED and SA scores. This tendency has contributed to the hesitancy among medicinal chemists to integrate previous molecular generation methods into their workflows. In summary, the advancements of FragGen in terms of Vina Score, QED, and SA indicate that geometric accuracy plays a crucial role in enhancing chemical plausibility, as the geometry of the current molecular state influences the structure of the subsequent fragment. For real-world applications, FragGen also establishes it as a valuable tool in drug discovery, particularly for generating easily synthesizable samples.

Performance of FragGen on well-studied pharmaceutical targets

To demonstrate FragGen's applicability in real-world scenarios, we evaluated its performance on several well-studied pharmaceutical targets. These targets, with well-characterized active sites and numerous experimentally discovered inhibitors, provide a suitable testing ground. Unlike the CrossDock benchmark, this experiment included two additional molecule sets: active (experimentally validated molecules serving as positive controls) and random (randomly selected chemical moieties from the GEOM-Drug set,[38] serving as negative controls). The Vina Score and molecular properties, akin to those used in the CrossDock experiment, are detailed in Table S1.† Fig. 4A illustrates the binding potency distribution of

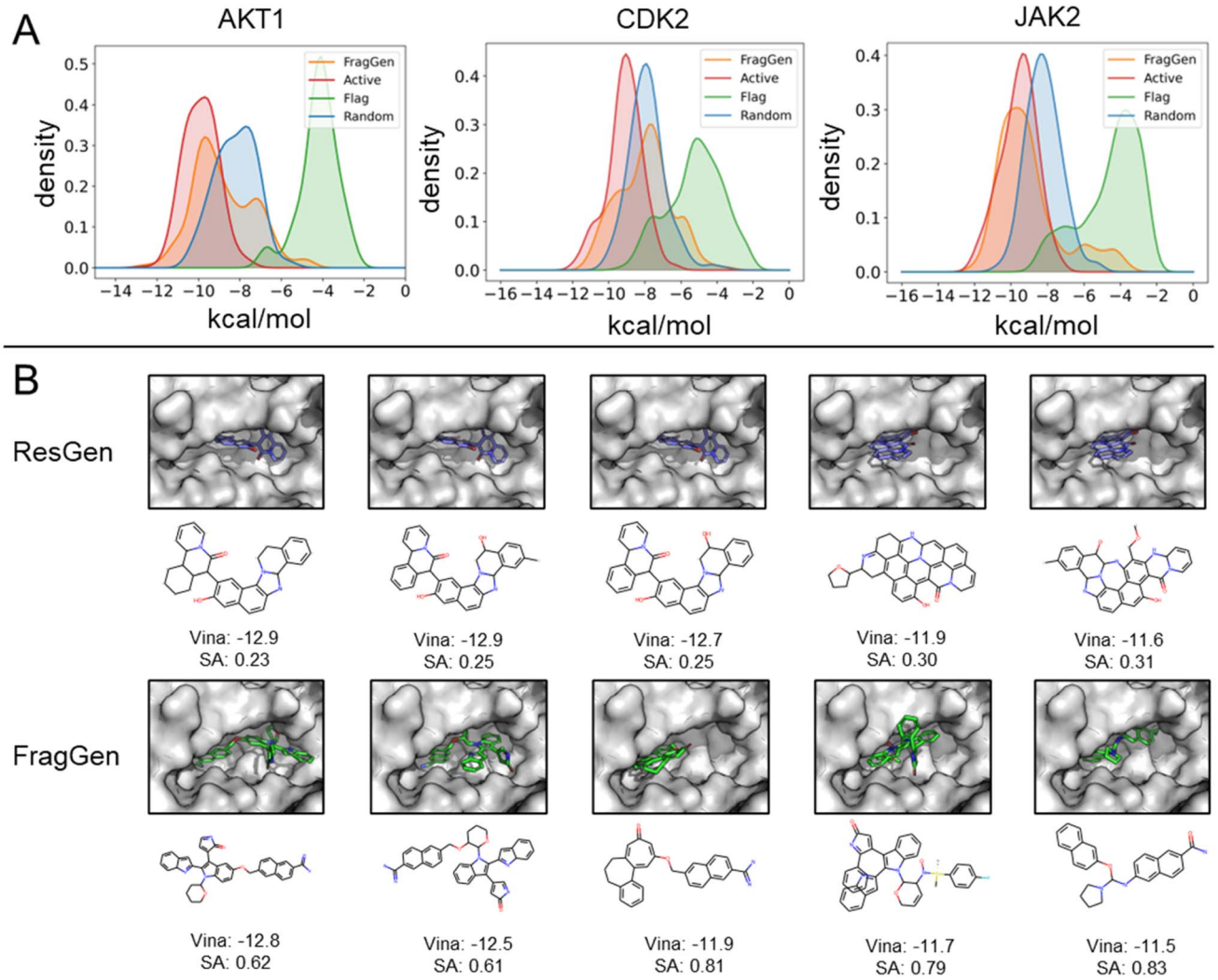

Fig. 4 (A) Distribution of binding potency (Vina Score) for FragGen and its counterpart across three well-studied targets. (B) Comparative visualization of the top 5 molecules in terms of binding potency, highlighting differences between the atom-wise (ResGen) and fragment-wise (FragGen) approaches.

FragGen-generated molecules (in orange) in comparison to the fragment-based counterpart, FLAG (in green). Notably, FragGen's distribution aligns more closely with the Active molecules, while FLAG aligns with the random set. This result again highlights the advantage of a rational geometry protocol in fragment-wise molecular generation, where accurate geometries lead to a better energy match with the binding protein.

From Table S1,† it is evident that ResGen, a state-of-the-art (SOTA) atom-wise molecular generation method, scores highly in terms of binding potency on targets like AKT1 and CDK2, with FragGen closely following. Despite this, we assert FragGen's superiority, as illustrated in Fig. 4B. While ResGen's top-generated molecules exhibit strong binding potency, they compromise on synthesizability and drugability. In contrast, FragGen's molecules not only achieve comparable binding potency to the top-Active molecules (with a marginal ~0.4 kcal-mol^{-1} difference) but also maintain the highest chemical accessibility, making them more favorable for chemists. This is further supported by the SA comparison in Table S1,† where FragGen outperforms other models.

Applying FragGen to design type II inhibitors of LTK with wet-lab validations

Kinases, essential enzymes in cellular signaling, play a critical role in various physiological processes, including cell growth, differentiation, and metabolism. As a result, numerous kinase inhibitors have been developed and approved for the treatment of diseases such as cancer, cardiovascular disorders, and inflammation.[39] Traditional kinase inhibitors, known as type I inhibitors, target the ATP-binding sites in the active conformations of kinases, offering therapeutic benefits but facing limitations in selectivity and resistance issues. In contrast, type II inhibitors, like sorafenib, target an additional allosteric site, the DFG-out pocket, potentially enabling more selective and less toxic treatments.[40] Despite the advantages of type II inhibitors, existing computational tools, such as quantitative structure–

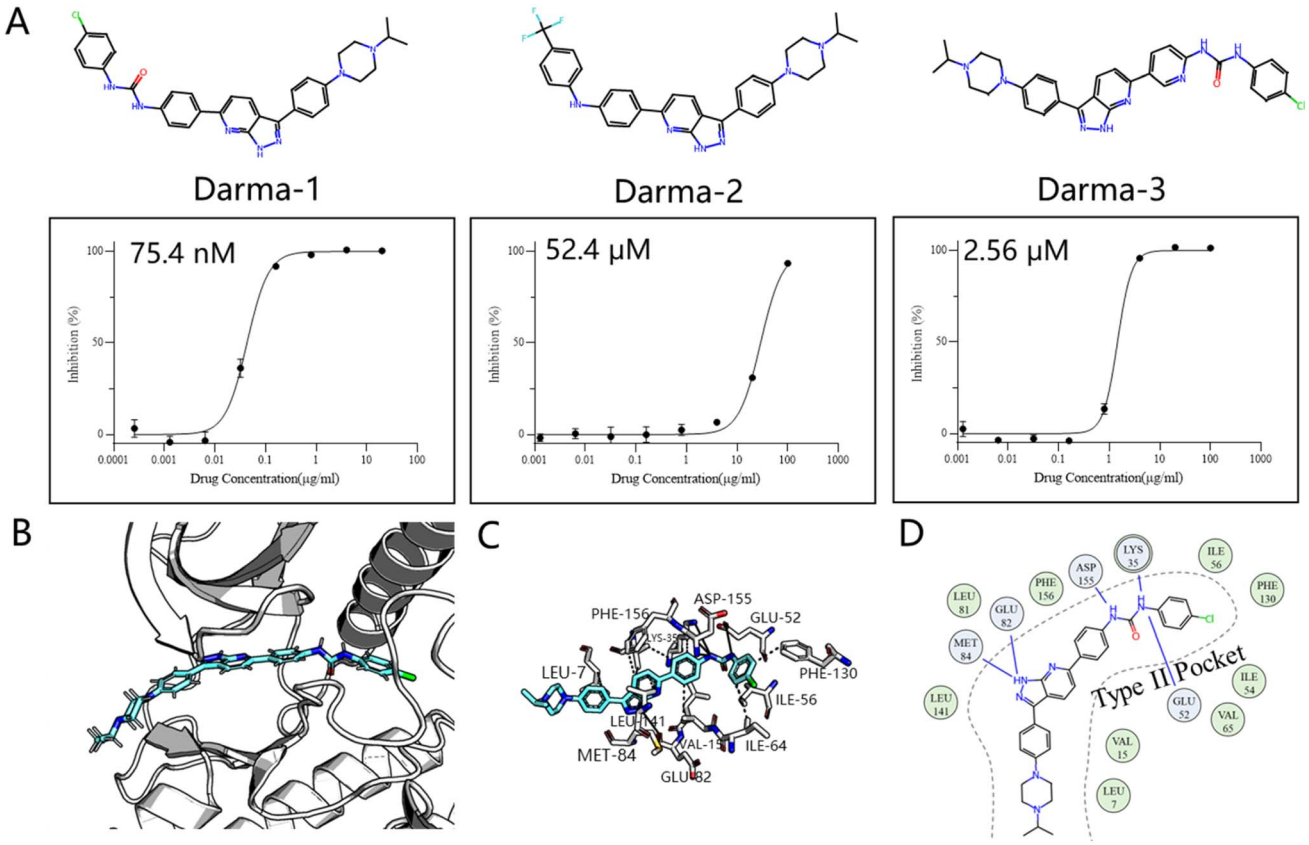

Fig. 5 (A) Structures of the three synthesized compounds designed by FragGen and their inhibitory activity (IC_{50}) against Ba/F3-CLIP1-LTK cells. (B) The binding conformation of Darma-1 in LTK DFG-out model. (C) 3D protein–ligand interactions analyzed by PLIP.[47] (D) 2D visualization of protein–ligand interactions, where the green represents hydrophobic interaction and the blue denotes hydrogen bond interaction.

activity relationship (QSAR) and docking screening,[41,42] fall short in designing potent molecules beyond the known chemical space, limiting the scope of discovering novel therapeutic agents. Therefore, the current molecular generation methods are ideal for filling this gap.

We chose the LTK as the validation system, a promising kinase target for treating non-small cell lung cancer according to the recent study.[43] This choice differs from previous retrospective studies, not only because it was validated through wet experiments rather than a controversial docking metric, but also because it is a novel target with few inhibitors designed for it. Inspired by the historical drug development of PDGFRβ target, which designs type II inhibitors based on the type I framework,[44] we developed an AI-powered structure-based workflow using FragGen. Specifically, we first built the LTK DFG-out homology model based on the anaplastic lymphoma kinase (ALK)[45] protein, owing to their high sequence similarity. Then we docked a previously reported type I inhibitor[46] of ALK into the LTK model, aiming to anchor the molecule at the pocket I region by retaining the head hinge-binding moiety. Starting with the anchored structure, FragGen was utilized to explore the chemical space targeting type II pocket. Within 10 minutes, FragGen proposed 97 chemical candidates. Subsequently, four filtering criteria were applied to narrow down the candidates: (1) number of hydrogen donors <5; (2) number of hydrogen acceptors <10; (3) $2 < \mathrm{Log}\,P < 5$; (4) and number of rotatable bonds <10. Out of this group, 10 molecules satisfied these conditions. Among them, three were chosen for further investigation based on synthesis feasibility as recommended by organic chemists (Fig. 5A). Details on the synthetic routes and molecular characterization are provided in the ESI.† Bioassays demonstrated high affinities for LTK, with Darma-1 exhibiting notable potency at 75.4 nM. The other two candidates showed affinities of 52.4 μM and 2.56 μM, respectively, highlighting FragGen's ligand design capability within protein pockets. The successful design of potent type II inhibitors may be attributed to FragGen's sophisticated handling of geometries. To illustrate this point, we analyzed the binding mode of the directly generated Darma-1 compound in Fig. 5B–D. It is evident that the generated compound forms comprehensive physical interaction with the type II pocket, like three hydrogen bonds with the ASP-155, LYS-35, and GLU-52 residues. Molecular generation models would lose practical utility if the generated geometries are not as reasonable as those proposed by FragGen no matter how promising the docking metric/ADMET metric they score: improper conformations will disrupt the interaction between proteins and ligands, diminishing the credibility of the generated samples.

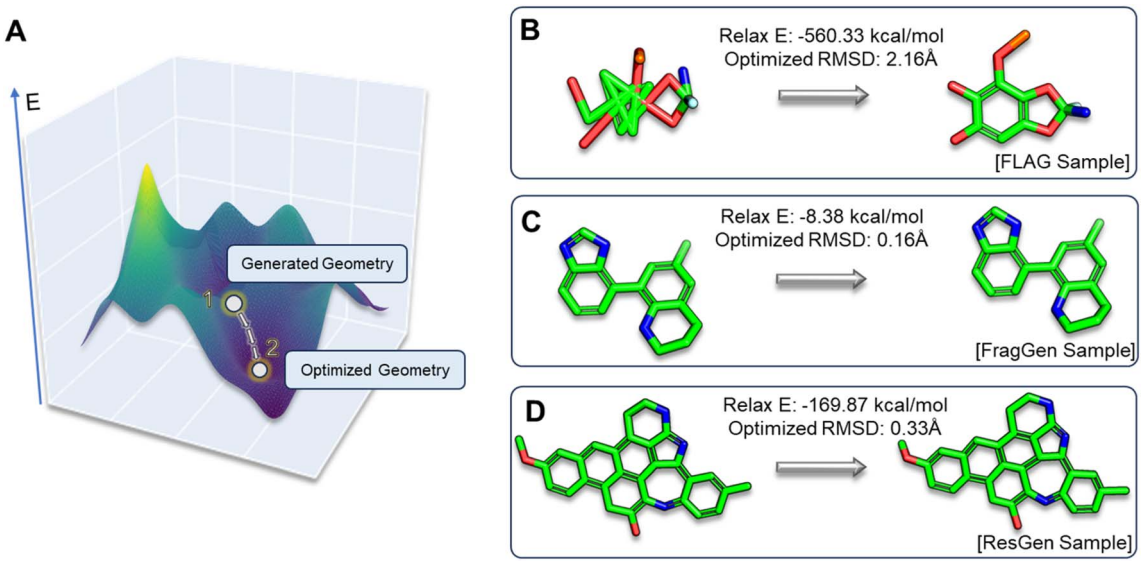

Fig. 6 (A). Visualization depicting the positions of generated and optimized geometries within the energy landscape. (B–D). Three case studies showcasing Relax E and OptRMSD metrics, each illustrating distinct scenarios encountered in FLAG, FragGen, and ResGen.

Geometric plausibility of generated molecules

In the realm of 3D molecular generation, many models rely on resort to geometry optimization to rectify distortions in generated molecules, essentially obscuring the limitations of deep learning methods in co-designing molecules with accurate geometries. Recognizing that previous experiments have only been able to indirectly and qualitatively address these geometric challenges, we introduce two novel metrics to gain a more detailed and quantitative assessment: relaxation energy (Relax E) and optimized root mean square deviation (OptRMSD). Specifically, the generated molecules undergo force field optimization, then the energy released and RMSD between the directly generated and optimized molecules are calculated, as shown in Fig. 6A.

Table 2 presents the results for Relax E and OptRMSD. Notably, in the realm of OptRMSD, certain models exhibit superior performance. However, it is crucial to acknowledge that OptRMSD inherently exhibits a preference for multi-ring structures. This is due to the fact that larger aromatic systems, with their more rigid frameworks, are less prone to conformational alterations, a phenomenon illustrated in Fig. 6D. Consequently, the lower OptRMSD scores observed in models like ResGen and Pkt2Mol, which are predisposed to generating multi-ring molecules, align with expectations. In contrast, FragGen distinguishes itself by achieving an OptRMSD score below 1 Å, underscoring its proficiency in creating structurally coherent molecules. When considering Relax E, a metric less biased towards multi-ring structures, a different picture emerges. Multi-ring structures, as shown in Fig. 6C and D, tend to release more energy following force-field optimization, even when they exhibit similar OptRMSD values to simpler molecules. In this context, FragGen again demonstrates superior performance, effectively aligning with our earlier assessments of its geometric accuracy. Conversely, the fragment-wise method FLAG, along with models like DiffBP and GraphBP that are prone to generating distorted conformations, give less favorable results in this metric.

OptRMSD is RMSD(R_i,R_e), and Relax E is $E_e - E_i$, where R_i,-R_e,E_i,E_e denote the initial and ending conformations and energy, respectively.

Ablation study of geometry handling protocols in FragGen

In the 3D molecular generation task, four of the six protocols in Fig. 3, Internal Coordinate, Cartesian Coordinate, Relative Vector, and Distance Geometry, have been instantiated by works like GraphBP, DiffBP, ResGen, and FLAG, respectively.

Table 2 The results of OptRMSD and Relax E across different methods

Case	GraphBP	DiffBP	Pkt2Mol	ResGen	FLAG	FragGen
OptRMSD	1.359 ±0.722	1.158 ±2.378	0.499 ±0.404	**0.465 ±0.319**	1.379 ±0.855	0.878 ±1.010
Relax E	−83.22 ±288.5	−100.9 ±235.1	−46.76 ±40.05	−54.33 ±45.21	−387.1 ±481.9	**−40.26 ±71.45**

In addition to these, we have integrated the GeomGNN and GeomOPT protocols into FragGen, creating two more versions of FragGen thereby providing a comprehensive analysis of each protocol within the context of 3D molecular generation. The results of this ablation within FragGen are detailed in Table S2.†

Table S2† reveals that molecules generated using the GeomGNN protocol exhibit the highest binding propensity. However, this favorable binding tendency comes at a cost to their synthesizability, which is approximately 24% lower compared to the other protocols. This reduction in synthesizability can be attributed to the compromise in local structural rationality while the model attempts to fill the protein pocket cavity (as depicted in Fig. S1A†) without explicitly considering the overall synthetic feasibility of the molecules. On the other hand, the GeomOPT approach shows a marked improvement in synthesizability, but the molecules generated under this protocol demonstrate a reduced binding tendency. This is primarily due to the geometric conformations becoming trapped in local minima within the protein structure during the generation process, leading to suboptimal molecule–protein interactions, as illustrated in Fig. S1A.† The Combined Strategy, which synergizes the physical constraints and the strengths of both Relative Vector and Internal Coordinates, emerges as a robust approach. It not only facilitates realistic molecule generation but also ensures a potent binding affinity to target proteins. The molecules produced under this strategy not only exhibit a higher binding tendency, outperforming all baseline methods (both atom-level and fragment-level) as shown in Table 1, but also demonstrate the highest level of synthesizability among all the protocols. This underscores the effectiveness and rationality of the molecular structures generated through this comprehensive protocol.

Conclusion

In this study, we aimed to address the frequently encountered issues of implausible chemical and geometric structures generated by many 3D molecular generative models. This journey began with a meticulous identification and analysis of six geometry handling protocols, each with its unique strengths and shortcomings. After acquiring the insights on the problems associated with existing approaches, we proposed developed FragGen, a hybrid strategy tailored for structure-based fragment-wise molecular generation. Experiments across the recognized benchmark and pharmaceutically relevant targets demonstrate that FragGen-generated molecules exhibit the highest binding potency (as estimated with docking scores) and synthesizability, meeting the practical demands of real-world drug discovery efforts. Our detailed geometric analysis and ablation study demonstrate that FragGen effectively coordinates the intricate interplay between molecular geometry and protein pocket structure, highlighting the crucial role of our proposed hybrid strategy in combining various geometry handling techniques to achieve FragGen's remarkable success. Finally, we successfully employed FragGen to design potent LTK type II inhibitors, showcasing its practical utility and completing the final step in the concept-algorithm-application chain. In summary, by integrating insights from different geometry handling protocols and tailoring them to the specific needs of fragment-wise molecular generation, FragGen has proven to be a robust tool for structure-based drug design. We believe the next step to advance FragGen is to realize objective optimization functionality. Specifically, utilizing a Reinforcement Learning approach could steer FragGen towards generating molecules that are more efficacious according to predefined objective functions.

Methods

Protein–ligand interaction learning module

To fully perceive the protein–ligand interaction, we first construct the protein–ligand graph and then apply the geometric message passing framework to them. This framework is described in the following formula:

$$\left(n'_{\mathrm{p}_i}, \vec{n}'_{\mathrm{p}_i}\right) = \mathrm{Emb}\left(n_{\mathrm{p}_i}, \vec{n}_{\mathrm{p}_i}\right),$$

$$\left(n'_{\mathrm{l}}, \vec{n}'_{\mathrm{l}_i}\right) = \mathrm{Emb}\left(n_{\mathrm{l}_i}, \vec{n}_{\mathrm{p}_i}\right),$$

$$\left(h_i, \overrightarrow{h_i}\right) = \mathrm{GeomEncoder}\left(n_{\mathrm{l}_i}, n_{\mathrm{p}_i}, \vec{n}_{\mathrm{l}_i}, \vec{n}_{\mathrm{p}_i}, \mathrm{e}_{ij}, \vec{\mathrm{e}}_{ij}\right).$$

where n_{p} and n_{l} denote the node features of proteins and ligands; $\rightarrow$ signifies the vector features; e_{ij} is the edge features between nodes i and j; h_i refers to the hidden features of the protein–ligand graph. Emb is the embedding layer, which maps the raw features of protein and ligand to the corresponding spaces with the same dimension. GeomEncoder is composed of several interaction layers based on geometric equivariant networks. The detailed architectures of Emb and GeomEncoder can be found in Part 1, ESI.†

Frontier prediction

To autoregressively generate the subsequent fragment, it is crucial to predict the frontier atom within the existing ligands. Notably, at the initial stage, there are no ligand atoms present, so the frontier is chosen from among the protein atoms. The probability of selecting the frontier from either ligand atoms or protein atoms can be simplified and represented as follows:

$$\left(n_{\mathrm{f}_i}, \overrightarrow{n_{\mathrm{f}_i}}\right) = f\left(\mathrm{SL}_{\mathrm{f}_1}(h_i), \mathrm{VL}_{\mathrm{f}_1}\left(\overrightarrow{h_i}\right)\right),$$

$$p_{\mathrm{f}_i} = \sigma\left(\mathrm{SL}_{\mathrm{f}_3}\left(\left\|\overrightarrow{n_{\mathrm{f}_i}}\right\|_2 + f\left(\mathrm{SL}_{\mathrm{f}_2}\left(n_{\mathrm{f}_i}\right)\right)\right)\right).$$

where p_{f_i} is the focal probability of node i; σ is the sigmoid function; SL and VL denote scalar layers and vector layers,[48] respectively. n_{f_i}, $\overrightarrow{n_{\mathrm{f}_i}}$ are intermediate scalar and vector features.

Cavity detection

Once the frontier has been established, the next step is to predict the cavity where the subsequent fragment can be optimally positioned. This prediction of the next cavity is

accomplished using a mixture density network, which is implemented as follows:

$$\left(r_i, \overrightarrow{r_i}\right) = \mathrm{GVP}_{\mathrm{r}}\left(\mathrm{SL}_{x1}(h_i), \mathrm{VL}_{x1}\left(\overrightarrow{h_i}\right)\right),$$
$$\left(w_i, \overrightarrow{w_i}\right) = \mathrm{GVP}_{w}\left(\mathrm{SL}_{x2}(h_i), \mathrm{VL}_{x2}\left(\overrightarrow{h_i}\right)\right),$$
$$\left(\Sigma_i, \overrightarrow{\Sigma_i}\right) = \mathrm{GVP}_{\Sigma}\left(\mathrm{SL}_{x3}(h_i), \mathrm{VL}_{x3}\left(\overrightarrow{h_i}\right)\right),$$
$$\overrightarrow{x_i} = \overrightarrow{x_{\mathrm{a}i}} + \sum_{k=1}^{K} w_i{}^k \overrightarrow{r_i}{}^k.$$

where $\overrightarrow{r_i}$ is the predicted relative vector, w_i and Σ_i are the factor and variance of the i-th component of the mixture Gaussian density, respectively, $\overrightarrow{x_{ai}}$ is the coordinate of the focal atom, and $\overrightarrow{x_i}$ is the detected cavity coordinate. GVP is the geometric vector perceptron,[49] which can be found in Part 1, ESI.†

Fragment query

Once the next cavity is identified, we can begin to search for suitable fragments that can be placed within it. It is important that the placement adheres to the principles of geometry and energy matching, which requires a thorough understanding of the local cavity environment. To achieve this, we gather detailed information about the cavity. This data is then integrated with the frontier features to facilitate an informed query for the appropriate fragment placement:

$$y_{\mathrm{m_{ij}}}, \vec{y}_{\mathrm{m_{ij}}} = \mathrm{GeomMessage}\left(h_{\mathrm{i}}, \overrightarrow{h_{\mathrm{i}}}, \mathrm{e_{ij}}, \overrightarrow{\mathrm{e_{ij}}}\right),$$
$$y_{h_{\mathrm{i}}}, \vec{y}_{h_{\mathrm{i}}} = \sum_{k=1}^{\mathrm{j}} \left(y_{\mathrm{m}_{ik}}, \vec{y}_{\mathrm{m}_{ik}}\right),$$
$$p_{y_{\mathrm{i}}} = \sigma\left(\mathrm{SL}_{t2}\left(\left\|\vec{y}_{h_{\mathrm{i}}}\right\|_2 + f\left(\mathrm{SL}_{t1}\left(y_{h_{\mathrm{i}}}\right)\right)\right)\right).$$

where $y_{\mathrm{m_{ij}}}, \vec{y}_{\mathrm{m_{ij}}}$ are the message between i, cavity node, and j, the K nearest neighborhoods of node i. $y_{h_{\mathrm{i}}}, \vec{y}_{h_{\mathrm{i}}}$ are clustered type hidden features on the cavity node i, and $p_{y_{\mathrm{i}}}$ is the probability of the next fragment type. GeomMessage is the message block that makes cavity node i blended with its pocket environment.

Attachment selection

The key difference between atom-wise and fragment-wise generation lies in the uncertainty associated with selecting the appropriate atom within a predicted fragment for connection and determining its subsequent geometry. Methods like FLAG addresses this challenge by pre-storing fragments with annotated connection points. While effective, this approach significantly increases the size of the fragment database and lacks elegance. In contrast, FragGen directly addresses this challenge using a Graph Attention Network (GAT),[50] a two-dimensional approach, to extract chemical information from the upcoming fragment. Additionally, a geometric network is applied to the frontier node to gather geometric information, such as the influence of existing molecular states and their interaction with protein pockets on the selection of the attachment point. This innovative approach is operationalized as follows:

$$h_{\mathrm{a_i}}, \vec{h}_{\mathrm{a_i}} = \mathrm{GVP}_{\mathrm{atta}}\left(h_{\mathrm{i}}, \overrightarrow{h_{\mathrm{i}}}\right),$$
$$h'_{\mathrm{f_j}} = \mathrm{GAT}\left(h_{\mathrm{f_j}}, \mathrm{e_{f_j}}\right),$$
$$y_{\mathrm{cr}}^{\mathrm{emb}}, y_{\mathrm{nx}}^{\mathrm{emb}} = \mathrm{Embed}(y_{\mathrm{cr}}, y_{\mathrm{nx}}),$$
$$h'_{\mathrm{a_j}} = \left(h'_{\mathrm{f_j}} \| y_{\mathrm{cr}}^{\mathrm{emb}} \| y_{\mathrm{nx}}^{\mathrm{emb}} \| h_{\mathrm{a_i}}\right),$$
$$p_{\mathrm{a_j}} = \sigma\left(\mathrm{MLP}\left(h'_{\mathrm{a_j}}\right)\right).$$

where $h_{\mathrm{a_i}}, \vec{h}_{\mathrm{a_i}}$ are the hidden features of i-th node's connected atom, *i.e.*, focal atom; and $h'_{\mathrm{f_j}}$ are the hidden feature of next fragment's atom j; $h_{\mathrm{f_j}}$,$\mathrm{e_{f_j}}$ are atom and edge features within the next fragment, respectively; y_{cr},y_{nx} are the current and next fragment types, respectively, and $y_{\mathrm{cr}}^{\mathrm{emb}}$, $y_{\mathrm{nx}}^{\mathrm{emb}}$ are their corresponding embeddings; $h'_{\mathrm{a_j}}$ is the concatenated feature of j-th atom in next fragment, and $\|$ is the concatenate operation; and $p_{\mathrm{a_j}}$ is the probability of the attachment of j-th node in the next fragment.

Bond linking

After identifying the next attachment atom, the subsequent variable to predict is the covalent bond. While many molecular generation methods, such as DiffSBDD, determine bonding relationships using empirical rules, FragGen takes a direct prediction approach that is both valence- and geometry-aware. The reason for incorporating geometric considerations is that the local pocket environment may favor certain types of interactions, such as the formation of π–π stacking interactions. At the same time, valence constraints guide bond prediction, ensuring that the cumulative valence from forming bonds does not exceed the valence capacity determined by the valence states of the two connected atoms. These principles are operationalized as follows:

$$h_{\mathrm{b_i}}, \vec{h}_{\mathrm{b_i}} = \mathrm{GVP}_{\mathrm{bond}}\left(h_{\mathrm{i}}, \overrightarrow{h_{\mathrm{i}}}\right),$$
$$h_{d_{\mathrm{ij}}},\ h_{\mathrm{nx}} = \mathrm{MLP}\left(d_{\mathrm{ij}}, n_{\mathrm{nx}}\right),$$
$$y_{\mathrm{cr}}^{\mathrm{emb}}, y_{\mathrm{nx}}^{\mathrm{emb}} = \mathrm{Embed}(y_{\mathrm{cr}}, y_{\mathrm{nx}}),$$
$$h_{\mathrm{valen}} = \mathrm{MLP}(\mathrm{valen_{cr}} \| \mathrm{valen_{nx}}),$$
$$p_{\mathrm{b_{ij}}} = \sigma\left(\mathrm{MLP}\left(h_{\mathrm{b_i}} h_{d_{\mathrm{ij}}} \| y_{\mathrm{cr}}^{\mathrm{emb}} \| y_{\mathrm{nx}}^{\mathrm{emb}} \| h_{\mathrm{valen}}\right)\right).$$

where $h_{\mathrm{b_i}}, \vec{h}_{\mathrm{b_i}}$ are the features of bonded atom, *i.e.*, focal atom; d_{ij} is the distance between focal node i and cavity node j; n_{nx} is the bonded atom of next fragment; $\mathrm{valen_{cr}}$ and $\mathrm{valen_{nx}}$ are valence of current and next bonded atoms, respectively; h_{valen} is the concatenated feature of valence information; and the $p_{\mathrm{b_{ij}}}$ is the probability of bond type between the current and next bonded atoms i and j.

Chemical initialization

As mentioned earlier, the geometry of the next fragment can be divided into four components. For the local geometries and rotation around the point, the former can be effectively achieved by the DL approach or a classical approach, as exemplified in the SDEGen,[20] and the latter benefits from an end-to-end approach. In our novel approach, we integrate knowledge from hybrid orbital theory, which has been instrumental in elucidating molecular conformations, into our prediction process. To illustrate this, consider a methane fragment; it naturally adopts a tetrahedral structure, thereby fixing the rotation around the point. When predicting the conformation of such a fragment, we first identify its connection to the existing molecule *via* a predicted bond. This involves defining a vector from the focal atom to the next attachment point (the to-be-aligned vector) and another from the focal atom to a designated pocket node (the target vector). We then compute a rotation matrix that aligns these vectors. This matrix is applied to rotate the fragment's conformation, initially set in a vacuum, to establish the initial geometry of the next fragment. The computation of this matrix proceeds as follows:

$$\boldsymbol{a}_{\text{norm}} = \frac{\boldsymbol{a}}{||\boldsymbol{a}||}$$

$$\boldsymbol{b}_{\text{norm}} = \frac{\boldsymbol{b}}{||\boldsymbol{b}||}$$

$$\boldsymbol{v} = \boldsymbol{a}_{\text{norm}} \times \boldsymbol{b}_{\text{norm}}$$

$$c = \boldsymbol{a}_{\text{norm}} \cdot \boldsymbol{b}_{\text{norm}}$$

$$[\boldsymbol{v}]_x = \begin{bmatrix} 0 & -v_z & v_y \\ v_z & 0 & -v_x \\ -v_y & v_x & 0 \end{bmatrix}$$

$$R_{\text{ab}} = I + [v]_x + [v]_x{}^2 \frac{1}{1+c}$$

$$r_{\text{f}}^{'} = R_{\text{ab}} r_{\text{f}}$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ are to-be-aligned and the target vectors, respectively, R_{ab} is the rotation matrix from vector $\boldsymbol{a}$ to $\boldsymbol{b}$, r_{f} is the fragment conformation generated in vacuum, and $r_{\text{f}}^{'}$ is the initialized fragment conformation.

Dihedral handling

For the next geometric variable, rotation around an axis, we employ a direct prediction method. This approach leverages both the geometric information of the connected atoms and the global characteristics of the ligands. The primary objective is to minimize the overall energy while simultaneously avoiding spatial clashes. The process of handling dihedral angles is executed as follows:

$$\&\, doublehyphen; 230pt\left(h_i^{\text{tor}}, \overrightarrow{h_i^{\text{tor}}}\right) = \text{GeomEncoder}\left(n_{\text{l}}, n_{\text{p}}, \vec{n}_{\text{l}}, \vec{n}_{\text{p}}, \text{e}_{\text{ll,pp,pl}}\vec{\text{e}}_{\text{ll,pp,pl}}\right),$$

$$\&\, doublehyphen; 335pth_{\text{mol}} = \sum_{i=1}^{N} h_i^{\text{tor}},$$

$$\&\, doublehyphen; 282pt\theta = \text{MLP}(h_{\text{a}}||h_{\text{b}}||h_{\text{mol}}),$$

$$R(\mathbf{u},\theta) = \begin{bmatrix} \cos(\theta) + u_x{}^2(1-\cos(\theta)) & u_x u_y(1-\cos(\theta)) - u_z\sin(\theta) & u_x u_z(1-\cos(\theta)) + u_y\sin(\theta) \\ u_y u_x(1-\cos(\theta)) + u_z\sin(\theta) & \cos(\theta) + u_y{}^2(1-\cos(\theta)) & u_y u_z(1-\cos(\theta)) - u_x\sin(\theta) \\ u_z u_z(1-\cos(\theta)) - u_y\sin(\theta) & u_z u_y(1-\cos(\theta)) + u_x\sin(\theta) & \cos(\theta) + u_z{}^2(1-\cos(\theta)) \end{bmatrix},$$

$$\&\, doublehyphen; 324ptr_{\text{f}}^{''} = R(\mathbf{u},\theta) r_{\text{f}}^{'}.$$

where $n_{\text{l}}, n_{\text{p}}, \vec{n}_{\text{l}}, \vec{n}_{\text{p}}, \text{e}_{\text{ll,pp,pl}}, \vec{\text{e}}_{\text{ll,pp,pl}}$ are the node and edge features of ligand and protein, ll,pp,pl denotes edge within ligands, within proteins, and between them, respectively; h_{mol} is the summation of ligand features; h_{a} and h_{b} are the features of current and next bonded atom, respectively, *i.e.*, focal atom and the next attachment atom; θ is the predicted dihedral angle; $R(\mathbf{u},\theta)$ is the rotation around the predicted bond vector ($r_{\text{a}} - r_{\text{b}}$); $r_{\text{f}}^{'}$ is the initialized fragment conformation; and $r_{\text{f}}^{''}$ is the final predicted fragment conformation.

Loss function

The total loss function is:

$$\begin{aligned} \mathscr{L} = & -\frac{1}{n}\left(\sum_{i=1}^{n} f_i \cdot \log p_{\text{f}_i} + (1-f_i)\cdot\log\left(1-p_{\text{f}}\right)\right) \\ & -\frac{1}{m}\left(\sum_{i=1}^{m} a_j \cdot \log p_{a_j} + \left(1-a_j\right)\cdot\log\left(1-p_{a_j}\right)\right) \\ & -\log\sum_{k=1}^{K} w_i{}^{(k)}\mathcal{N}\left(x_i{}^{(k)} + r_{a_i}, \Sigma_i{}^{(k)}\right) \\ & -\sum_{i=1}^{n} y_i \log p_{y_i} - \sum_{j=1}^{n} \text{b}_{ij}\log p_{\text{b}_{ij}} \\ & -\log\left(\frac{\text{e}^{\kappa\cos(\theta-\mu)}}{2\pi I_0(\kappa)}\right). \end{aligned}$$

where f_i and p_{f_i} are the frontier atom label and prediction, respectively, and n is the total number of the existing ligand/protein atoms; a_j and p_{a_j} are the attachment atom label and

preiction, respectively, and m is the number of the next fragment atoms; $x_i^{(k)}$, $w_i^{(k)}$, $\Sigma_i^{(k)}$ are the k-th component of the relative vector, coefficient, and variance in the cavity detection module, respectively, and K is the number of components; y_i and p_{y_i} are predicted fragment label and prediction, respectively; b_{ij} and $p_{b_{ij}}$ are predicted bond label and prediction, respectively. The final term is the von-mises loss, aiming to evaluate how close are two angles. In this loss, μ and θ are dihedral angle label and prediction, respectively, κ is the concentration parameter, a higher value means a more peaked distribution, and the I_0 is the modified Bessel function of order 0.

Cell culture

Ba/F3 cells (ACC 300) were purchased from DSMZ, and 293T cells (SCSP-502) were purchased from National Collection of Authenticated Cell Cultures. Ba/F3 cells are cultured in RPMI M Medium 1640 (U21-279b, YOBIBIO) with 10% FBS (F8318, Sigma-Aldrich) and 10 ng ml^{-1} IL-3(90143ES10, Yeasen). 293T cells are cultured in DMEM (U21-265B, YOBIBIO) with 10% FBS. All growth media are supplemented with 1% Penicillin–Streptomycin–Glutamine (10378016, Gibco). Cell cultures are maintained in culture flasks in 5% CO_2 atmosphere at 37 °C.

Transformation of Ba/F3-CLIP1-LTK cell line

pMD2.G (DB00002) and pCMVR 8.74 (P4872) were purchased from Miaoling Biology. CLIP1-LTK fusion genes are generated based on cDNAs of human-derived CLIP1 and LTK genes using pLV vector. The full-length plv-CLIP1-LTK plasmids were constructed and packaged by VectorBuilder. 293T cells are co-transfected with pLV-CLIP1-LTK, pMD2.G and pCMVR 8.74 to produce retrovirus particles. The viral supernatants are collected and concentrated following the instructions of Lenti-X Concentrator (631231, Takara). Ba/F3 cells are subsequently transfected with the virus and selected with 2 μg ml^{-1} puromycin to obtain Ba/F3-CLIP1-LTK cell line.

Ba/F3-CLIP1-LTK activity assay

1×10^4 Ba/F3-CLIP1-LTK cells are seeded in 96-well plates with RPMI-1640 and treated with gradient concentrations of interest compounds for 48 h. Afterward, 10 μL of 5 mg ml^{-1} MTT solution is added into each well and the cells are further incubated for another 4 h. Then, 100 μL of triplex 10% SDS-0.1% HCl-PBS solution is added to dissolve the formazan deposited on the bottom of the plates, and the plates are then further retained in an incubator overnight. The absorbance at 570 nm is measured with the reference wavelength at 650 nm using a Synergy H1 microplate reader (BioTek).

Data availability

The data and source code of this study is freely available at GitHub (**https://github.com/HaotianZhangAI4Science/FragGen**) to allow replication of the results.

Author contributions

O. Z. and Y. H. contributed to the main idea and code; S. C. and P. C. contributed to the bioassays; M. Y. and S. C. contributed to the chemical synthesis; X. Z. and H. T. contributed to the ablation study; Y. Z. and M. W. contributed to the data presentation and data collection; Z. W., H. F., Z. Z., and H. C. contributed to the baseline models application; Y. K. and C.-Y. H. contributed to the manuscript envision and experimental design. T. H. contributed to the essential financial support, the conceptualization, and was responsible for the overall quality.

Conflicts of interest

The authors declare that there is no conflict of interest.

Acknowledgements

This work was supported by This work was financially supported by National Key Research and Development Program of China (2022YFF1203003), National Natural Science Foundation of China (22220102001), and Natural Science Foundation of Zhejiang Province (LD22H300001).

References

1 A. S. Rifaioglu, H. Atas, M. J. Martin, R. Cetin-Atalay, V. Atalay and T. Doğan, *Brief. Bioinform.*, 2019, **20**, 1878–1912.
2 A. N. Jain, *Curr. Opin. Drug Discov. Dev.*, 2004, **7**, 396–403.
3 D. Xue, Y. Gong, Z. Yang, G. Chuai, S. Qu, A. Shen, J. Yu and Q. Liu, *Wiley: Comput. Mol. Sci.*, 2019, **9**, e1395.
4 D. Jiang, Z. Ye, C.-Y. Hsieh, Z. Yang, X. Zhang, Y. Kang, H. Du, Z. Wu, J. Wang and Y. Zeng, *Chem. Sci.*, 2023, **14**, 2054–2069.
5 J. Wang, C.-Y. Hsieh, M. Wang, X. Wang, Z. Wu, D. Jiang, B. Liao, X. Zhang, B. Yang and Q. He, *Nat. Mach. Intell.*, 2021, **3**, 914–922.
6 P. Bongini, M. Bianchini and F. Scarselli, *Neurocomputing*, 2021, **450**, 242–252.
7 N. Brown, J. Cambruzzi, P. J. Cox, M. Davies, J. Dunbar, D. Plumbley, M. A. Sellwood, A. Sim, B. I. Williams-Jones and M. Zwierzyna, *Prog. Med. Chem.*, 2018, **57**, 277–356.
8 J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek and A. Potapenko, *Nature*, 2021, **596**, 583–589.
9 R. J. Townshend, S. Eismann, A. M. Watkins, R. Rangan, M. Karelina, R. Das and R. O. Dror, *Science*, 2021, **373**, 1047–1051.
10 H. Zhang, J. Zhang, H. Zhao, D. Jiang and Y. Deng, *bioRxiv*, 2023, preprint, 2023.2003. 2008.531607.
11 Z. Gao, Y. Hu, C. Tan and S. Z. Li, *arXiv*, 2023, preprint, arXiv:2302.07120, DOI: **10.48550/arXiv.2302.07120**.
12 M. Ragoza, T. Masuda and D. R. Koes, *Chem. Sci.*, 2022, **13**, 2701–2713.
13 X. Peng, S. Luo, J. Guan, Q. Xie, J. Peng and J. Ma, presented in part at the, *International Conference on Machine Learning*, 2022.

14 H. Lin, Y. Huang, M. Liu, X. Li, S. Ji and S. Z. Li, *arXiv*, 2022, preprint, arXiv:2211.11214, DOI: 10.48550/arXiv.2211.11214.
15 O. Zhang, J. Zhang, J. Jin, X. Zhang, R. Hu, C. Shen, H. Cao, H. Du, Y. Kang, Y. Deng, F. Liu, G. Chen, C.-Y. Hsieh and T. Hou, *Nat. Mach. Intell.*, 2023, **5**, 1020–1030.
16 M. Liu, Y. Luo, K. Uchino, K. Maruhashi and S. Ji, presented in part at the, Proceedings of the 39th International Conference on Machine Learning, *Proceedings of Machine Learning Research*, 2022.
17 N. Ståhl, G. Falkman, A. Karlsson, G. Mathiason and J. Boström, *J. Chem. Inf. Model.*, 2019, **59**, 3166–3176.
18 Z. Zhang, Y. Min, S. Zheng and Q. Liu, presented in part at the, *The Eleventh International Conference on Learning Representations*, 2022.
19 G. N. Simm and J. M. Hernández-Lobato, *arXiv*, 2019, preprint, arXiv:1909.11459, DOI: 10.48550/arXiv.1909.11459.
20 H. Zhang, S. Li, J. Zhang, Z. Wang, J. Wang, D. Jiang, Z. Bian, Y. Zhang, Y. Deng and J. Song, *Chem. Sci.*, 2023, **14**, 1557–1568.
21 J. Zhu, Y. Xia, C. Liu, L. Wu, S. Xie, T. Wang, Y. Wang, W. Zhou, T. Qin and H. Li, *arXiv*, 2022, preprint, arXiv:2202.01356, DOI: 10.48550/arXiv.2202.01356.
22 W. Lu, Q. Wu, J. Zhang, J. Rao, C. Li and S. Zheng, *Adv. Neural Inf. Process. Syst.*, 2022, **35**, 7236–7249.
23 H. Stärk, O. Ganea, L. Pattanaik, R. Barzilay and T. Jaakkola, presented in part at the, *International Conference on Machine Learning*, 2022.
24 X. Zhang, O. Zhang, C. Shen, W. Qu, S. Chen, H. Cao, Y. Kang, Z. Wang, E. Wang and J. Zhang, *Nat. Comput. Sci.*, 2023, **3**, 789–804.
25 G. Corso, B. Jing, R. Barzilay and T. Jaakkola, presented in part at the, *International Conference on Learning Representations (ICLR 2023)*, 2023.
26 V. G. Satorras, E. Hoogeboom and M. Welling, presented in part at the, *International Conference on Machine Learning*, 2021.
27 K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko and K.-R. Müller, *J. Chem. Phys.*, 2018, **148**, 241722.
28 O. Zhang, T. Wang, G. Weng, D. Jiang, N. Wang, X. Wang, H. Zhao, J. Wu, E. Wang and G. Chen, *Nat. Comput. Sci.*, 2023, **3**, 849–859.
29 Y. Luo and S. Ji, presented in part at the, *International Conference on Learning Representations*, 2021.
30 A. Schneuing, Y. Du, C. Harris, A. Jamasb, I. Igashov, W. Du, T. Blundell, P. Lió, C. Gomes and M. Welling, *arXiv*, 2022, preprint, arXiv:2210.13695, DOI: 10.48550/arXiv.2210.13695.
31 C. Shi, S. Luo, M. Xu and J. Tang, presented in part at the, Proceedings of Machine Learning Research, *Proceedings of the 38th International Conference on Machine Learning*, 2021.
32 W. A. Bingel and W. Lüttke, *Angew. Chem., Int. Ed.*, 1981, **20**, 899–911.
33 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.
34 O. Trott and A. J. Olson, *J. Comput. Chem.*, 2010, **31**, 455–461.
35 D. E. Clark and S. D. Pickett, *Drug Discov. Today*, 2000, **5**, 49–58.
36 P. Ertl and A. Schuffenhauer, *J. Cheminf.*, 2009, **1**, 1–11.
37 A. Ganesan, *Curr. Opin. Chem. Biol.*, 2008, **12**, 306–317.
38 S. Axelrod and R. Gomez-Bombarelli, *Sci. Data*, 2022, **9**, 185.
39 F. M. Ferguson and N. S. Gray, *Nat. Rev. Drug Discovery*, 2018, **17**, 353–377.
40 Z. Zhao, H. Wu, L. Wang, Y. Liu, S. Knapp, Q. Liu and N. S. Gray, *ACS Chem. Biol.*, 2014, **9**, 1230–1241.
41 A. Abuhammad and M. O. Taha, *Expert Opin. Drug Discov.*, 2016, **11**, 197–214.
42 O. Daoui, H. Nour, O. Abchir, S. Elkhattabi, M. Bakhouch and S. Chtita, *J. Biomol. Struct. Dyn.*, 2023, **41**, 7768–7785.
43 H. Izumi, S. Matsumoto, J. Liu, K. Tanaka, S. Mori, K. Hayashi, S. Kumagai, Y. Shibata, T. Hayashida and K. Watanabe, *Nature*, 2021, **600**, 319–323.
44 E. Bethke, B. Pinchuk, C. Renn, L. Witt, J. Schlosser and C. Peifer, *ChemMedChem*, 2016, **11**, 2664–2674.
45 C. C. Lee, Y. Jia, N. Li, X. Sun, K. Ng, E. Ambing, M.-Y. Gao, S. Hua, C. Chen and S. Kim, *Biochem. J.*, 2010, **430**, 425–437.
46 C. Chen, P. Pan, Z. Deng, D. Wang, Q. Wu, L. Xu, T. Hou and S. Cui, *Bioorg. Med. Chem. Lett.*, 2019, **29**, 912–916.
47 S. Salentin, S. Schreiber, V. J. Haupt, M. F. Adasme and M. Schroeder, *Nucleic Acids Res.*, 2015, **43**, W443–W447.
48 C. Deng, O. Litany, Y. Duan, A. Poulenard, A. Tagliasacchi and L. J. Guibas, presented in part at the, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.
49 B. Jing, S. Eismann, P. Suriana, R. J. Townshend and R. Dror, *arXiv*, 2020, preprint, arXiv:2009.01411, DOI: 10.48550/arXiv.2009.01411.
50 P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio and Y. Bengio, *arXiv*, 2017, preprint arXiv:1710.10903, DOI: 10.48550/arXiv.1710.10903.