Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Generative exploration of fragment-based chemical space via large language models enables the discovery of potent leads for targets lacking bioactive ligands

Bo Yang

ybo@jlu.edu.cn

Jilin University https://orcid.org/0000-0003-1927-8419

hao tuo

Jilin University

Geng Qin

First Hospital of Jilin University

Beilei Shen

Changchun Veterinary Research Institute

Haishi Zhao

College of Earth Sciences, Jilin University https://orcid.org/0000-0001-5696-8887

Yan Li

Jilin University

Xuanning Hu

Jilin University

Feng Wei

The First Hospital of Jilin University https://orcid.org/0000-0001-8451-4516

Xueyan Liu

Jilin University

Article

Keywords:

Posted Date: July 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10212415/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Generative exploration of fragment-based chemical space via large language models enables the discovery of potent leads for targets lacking bioactive ligands

Hao Tuo[1,2], Geng Qin[3], Beilei Shen[4,5], Haishi Zhao[6], Yan Li[1,2], Xuanning Hu[1,2], Feng Wei[7,8*], Xueyan Liu[1,2*], Bo Yang[1,2*]

[1]College of Computer Science and Technology, Jilin University, Changchun, China.
[2]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, China.
[3]Cancer Institute,First Hospital of Jilin University,Changchun,China.
[4]Changchun Veterinary Research Institute,Chinese Academy of Agricultural Sciences,Changchun,China.
[5]State Key Laboratory of Pathogen and Biosecurity,Key Laboratory of Jilin Province for Zoonosis Prevention and Control,Changchun,China.
[6]College of Earth Sciences, Jilin University, Changchun, China.
[7]Department of Hepatobiliary and Pancreatic Surgery, The First Hospital of Jilin University, Changchun, China.
[8]General Surgery Center, The First Hospital of Jilin University, Changchun, China.

*Corresponding author(s). E-mail(s): wei_feng@jlu.edu.cn; xueyanliu@jlu.edu.cn; ybo@jlu.edu.cn;

Abstract

Lead discovery is a critical step in drug development, yet current computational methods often face a trade-off between exploring novel chemical space and leveraging target-specific knowledge priors. Score-oriented algorithms can support broad exploration but lack medicinal-chemistry guidance, whereas knowledge-driven design benefits from target-specific knowledge while often staying close to existing chemotypes, limiting their utility in the pivotal real-world scenario of designing leads for targets that lack bioactive ligands. Here, we present a lead-discovery framework, AutoLeadDesign, that couples large language model (LLMs) reasoning with fragment-based chemical exploration without requiring target-specific ligand templates. Through a mutually reinforcing feedback loop between the LLM and chemical fragments, AutoLeadDesign enables target-aware and interpretable exploration of chemically diverse scaffolds while preserving critical binding interactions. Computational analysis reveals that AutoLeadDesign mitigates the tendency towards local optimization often observed in LLM-based molecular design. Analysis of the design trajectories indicates that the design process shares similar mechanisms to medicinal-chemistry strategies reported in expert-led studies (specifically fragment linking, merging, and growing), supporting the interpretability of the binding modes. Furthermore, we synthesized and experimentally validated leads targeting two therapeutically relevant targets: SARS-CoV-2 PLpro and the oncogenic KRAS G12D mutant. One designed inhibitor, PLP011, showed potent antiviral activity by suppressing viral replication and restoring the viability of infected cells. Another candidate, KP032, achieved a half-maximal inhibitory concentration (IC_{50}) of $50.24\,nM$ against the challenging target KRAS G12D. Finally, we released the fragment libraries targeting the therapeutic proteins investigated in this study. These libraries are intended to serve as fragment-based priors, providing medicinal chemists and future machine learning frameworks with essential structural insights for targeted drug design. By combining LLM-based biochemical reasoning with fragment-grounded exploration, AutoLeadDesign provides a practical strategy for more adaptive and interpretable lead discovery.

1 Introduction

Lead discovery represents a cornerstone in addressing global health challenges. It aims to systematically explore chemical space to identify bioactive compounds with high affinity for target proteins, as well as novel interaction modalities with promising therapeutic effects [1–14]. Traditional expert-driven approaches [15] rely heavily on human decision-making, and this dependence on potentially biased intuition often limits exploration of chemical space, ultimately constraining the efficiency of lead discovery. Meanwhile, data-driven generative models [16, 17] are trained to learn the distribution of known active compounds, which may bias generation toward existing chemotypes, thus limiting the novelty of molecules and leading to unsatisfactory optimization performance.

Fundamentally, lead design can be conceptualized as a combinatorial optimization problem within chemical space. A more straightforward solution involves algorithms that navigate discrete chemical space through an iterative design-and-evaluate paradigm oriented by affinity scores [18–23]. Starting from randomly initialized compound sets, these methods employ heuristic search algorithms to evolve molecular structures, using scoring functions to guide the exploration process. Their independence from known bioactive inhibitors makes them particularly suitable for drug discovery against targets that lack bioactive ligands. However, a fundamental limitation lies in their inability to effectively integrate domain-specific knowledge into the design process. Without expert-level biochemical guidance, these heuristic searches often remain weakly constrained, leading to the stochastic exploration of vast chemical landscapes. This frequently yields structures that lack biological relevance, thereby reducing the hit rate for therapeutic targets.

Large language models (LLMs), which have shown in understanding, summarizing and utilizing knowledge, have demonstrated remarkable capabilities in handling domain-specific tasks [24–28]. Recent studies [29–38] also suggest that LLMs now exhibit chemical reasoning capabilities comparable to human experts. These findings indicate that LLMs encode useful domain knowledge that can be leveraged for downstream scientific tasks. Recently, several knowledge-driven design frameworks [39, 40] have integrated LLMs to iteratively refine high-potential molecular candidates. Our analysis, however, shows that compounds designed by LLMs exhibit a pronounced tendency toward local optimization, indicating limited capacity to explore uncharted chemical space (Fig. 2a). This presents substantial challenges to de novo drug design against targets lacking known bioactive ligands, particularly for rare diseases or newly identified pathogenic proteins. This limitation may result from the vastness of chemical space and the difficulty of imposing effective structural constraints on LLM-based generation. Fragment-based design [41], which utilizes molecular fragments rather than individual atoms as fundamental building blocks, is a promising alternative for molecular construction. By reducing the search space while maintaining structural flexibility, it shows strong potential for improving the performance of LLM-based lead discovery.

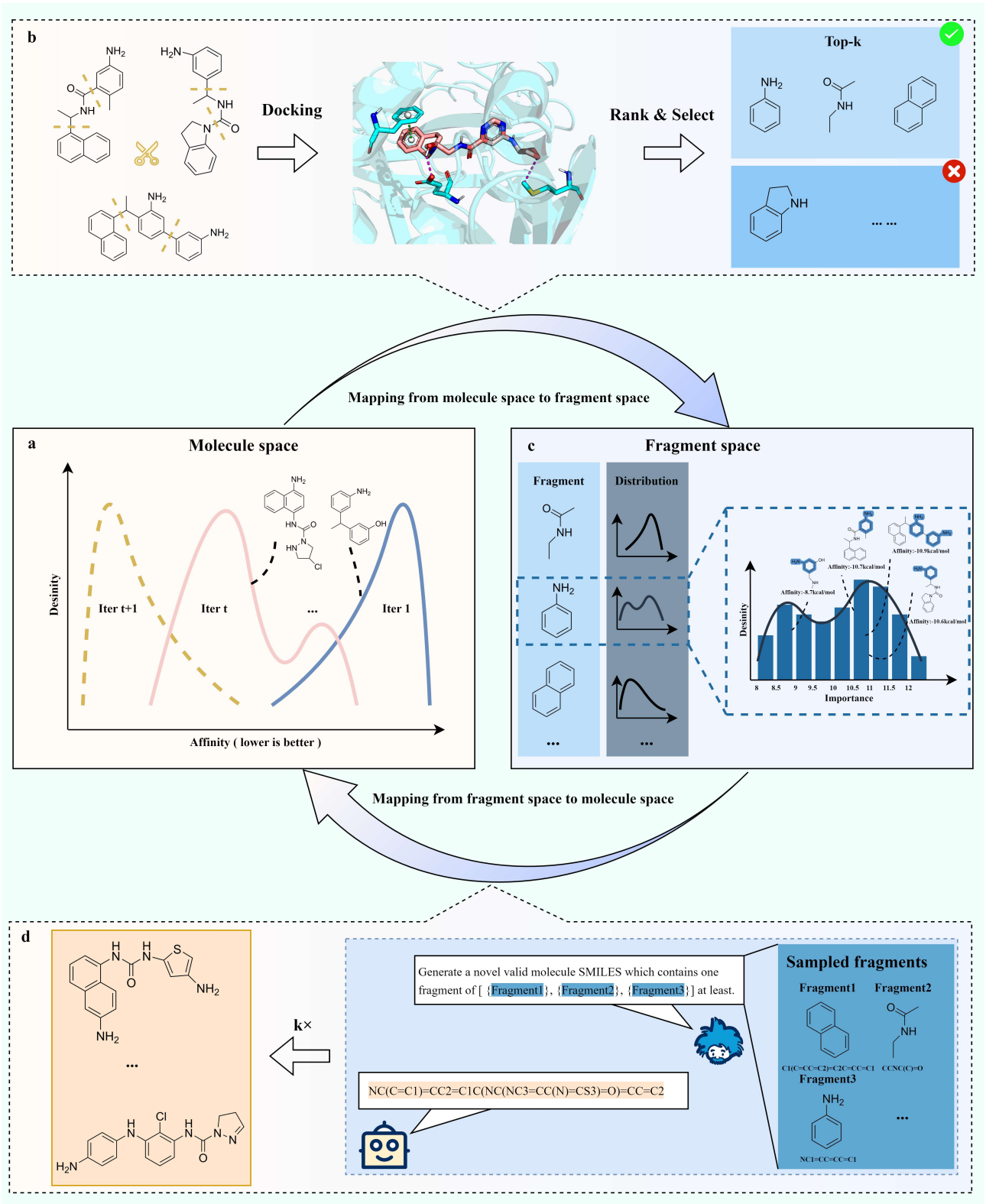


Fig. 1 The workflow of AutoLeadDesign. a, The molecule space. The compound library contains all the compounds designed by AutoLeadDesign with docking scores as their labels. b, The process of mapping from molecule space to fragment space. The compounds in the molecule library are decomposed into fragments, and the Top-k highest affinity fragments are selected to update the fragment library. c, Update the fragment library. Fragment importance is quantitatively determined by calculating the average binding affinity of all library compounds containing each specific fragment moiety. d, The process of mapping from fragment space to molecule space. The fragments are sampled based on the importance to bind with the target protein. Then, these key fragments are fed to Large Language Models to design novel compounds.

In this study, we propose AutoLeadDesign, a framework that integrates LLMs with chemical fragments to efficiently explore the chemical space for lead discovery. The closed-loop workflow, as illustrated in Fig. 1, consists of two key mappings. The mapping from molecule space to fragment space (Fig. 1b) decomposes explored molecules into fragments, thereby increasing design flexibility. By quantitatively analyzing the fragment occurrence frequencies and their contributions to binding energy, the framework continuously updates its statistical characterization of target-relevant chemical space. The mapping from fragment space to molecule space (Fig. 1d) then leverages rich domain knowledge encoded in the LLM to further explore chemical space based on the updated fragment-level characterization. Through the iterative execution of the two mappings, AutoLeadDesign enables more effective use of LLM knowledge and more comprehensive exploration of the chemical space, which is essential to discovering novel and high-affinity leads for the target protein.

Notably, AutoLeadDesign demonstrated superior performance compared to competing baseline models in de novo drug design. Through two case studies targeting PRMT5 and PLpro, we demonstrate its capability to de novo generate leads exhibiting high binding affinity and distinct binding modes without relying on known bioactive ligands. The designed compounds demonstrate superior docking scores, and the predicted binding modes are consistent with related studies, supporting their potential as bioactive candidates. The multidimensional analysis of the design process further reveals the underlying mechanisms of fragment growing, merging, and linking, which align with conventional fragment-based drug design (FBDD) strategies, thereby elucidating the rationale for its effectiveness. Furthermore, we synthesized and experimentally evaluated candidates targeting two distinct, clinically significant proteins: the SARS-CoV-2 papain-like protease (PLpro) and the oncogenic KRAS G12D mutant. The experimental results show that the designed inhibitor, PLP011, exhibits potent antiviral activity by effectively arresting viral replication and restoring the viability of infected cells. For the notoriously undruggable KRAS G12D, the candidate KP032 achieved a remarkable half-maximal inhibitory concentration (IC_{50}) of $50.24\,nM$. Additionally, we have released the target-specific fragment libraries generated during the design process, providing essential structural priors for medicinal chemists and future machine learning frameworks. These findings highlight the broad applicability and considerable potential of AutoLeadDesign in lead discovery.

2 Results

2.1 AutoLeadDesign Outperforms Baseline Methods in De Novo Molecular Design

To benchmark the overall performance of AutoLeadDesign, we compare our method against three combinatorial optimization methods: REINVENT citereinvent, ChemGE [23], RGA [21], and an LLM-based method, LMLF[39]. Commencing with randomly selected compound libraries, these methodologies progressively optimize molecular designs through iterative evaluation and expansion of uncharted chemical space. We randomly select 10 protein binding pockets in the CrossDocked2020 dataset [42] as a test set to evaluate these methods and AutoLeadGen. To ensure the comprehensiveness of validation, these proteins correspond to distinct classes: hydrolase, transferase, oxidoreductase, member protein, and immune system protein. In fact, the initial compound set may significantly impact the direction of exploration and the final result. To enable comprehensive evaluation, we have established two distinct initial compound sets under different experimental scenarios. One is

random initialization: we randomly select 100 molecules from ZINC as the 0-th generation. Another is prior initialization: on the basis of a random set, we add the co-crystal ligand of target proteins in the CrossDocked2020 dataset. We conducted 20 generations of optimization cycles, with each generation involving the design of 100 compounds per target protein in the test set. In practice, pharmacologists concern about the high-affinity compounds rather than the overall distribution of designed compounds. Hence, we select the top-100 compounds with the best docking scores for evaluation and consider the TOP-1/10/100 metrics (average docking score of top-1/10/100). For conciseness, Table 1 presents the averaged metrics for designed molecules across ten target proteins, with detailed results available in Supplementary Tables 1-10.

Table 1 Result of the de novo design task in the scene of random and prior initialization

Methods	Random			Prior		
	Top-1	Top-10	Top-100	Top-1	Top-10	Top-100
ChemGE	-10.76	-10.41	-10.02	-10.72	-10.38	-9.86
RGA	-7.66	-7.11	-6.21	-9.34	-8.94	-7.19
Reinvent	-10.79	-10.31	-9.59	-10.71	-10.24	-9.59
LMLF	-9.47	-8.97	-7.81	-11.13	-10.42	-8.97
AutoLeadDesign	**-11.76**	**-10.93**	**-10.07**	**-11.92**	**-11.34**	**-10.50**

As shown in Table 1, AutoLeadDesign demonstrates statistically superior performance in both random and prior initialization scenarios, designing leads with the highest binding affinity compared to baseline methods. In prior initialization scenarios, AutoLeadDesign and another LLM-based method, LMLF, achieved optimal and sub-optimal performance, respectively, illustrating the effectiveness of LLM in drug design. In the random initialization scenario, AutoLeadDesign maintains the SOTA performance, demonstrating the capability in designing high-affinity compounds without reliance on bioactive compounds, which are difficult to acquire. In contrast, the binding free energy of top compounds designed by LMLF exhibited a 15% reduction (from -11.13 kcal/mol to -9.47 kcal/mol), indicating its sensitivity to initialization. This limitation may originate from the inherent characteristic of local optimization. In-depth analyses will be conduct in the following case study of PRMT5.

2.2 In Silico Analysis of AutoLeadDesign in Navigating Uncharted Chemical Space

In this section, we provide a detailed in silico analysis of AutoLeadDesign's performance in lead discovery. Focusing on the therapeutic targets, the Protein Arginine Methyltransferase-5 (PRMT5) and the SARS-CoV-2 papain-like protease, our experiments are designed to verify the framework's proficiency in exploring uncharted chemical space and its robustness to initial conditions. By characterizing the interaction mechanisms and drug-likeness of the generated candidates, we demonstrate the reliability of our computational pipeline.

To demonstrate the capability of AutoLeadGen in designing novel leads with valid interactions for target proteins, we performed a lead generation task for PRMT5, a protein of significant therapeutic interest. PRMT5

catalyzes the transfer of methyl groups from S-adenosylmethionine (SAM) to arginine residues, forming symmetrical ω-NG-dimethyl arginine on protein substrates, which plays a key role in crucial cellular pathways, including cell growth and proliferation. The overexpression of PRMT5 involved in important cancer functions, indicating PRMT5 as a potential target for anticancer drug development [43–46]. LLY is a high-affinity inhibitor targeting PRMT5, forming hydrogen bonds to ASP419 and MET420, which play a crucial role in competing the binding pocket with SAM [47]. In our lead design study, we initialized the process with LLY and 100 randomly selected ZINC compounds.

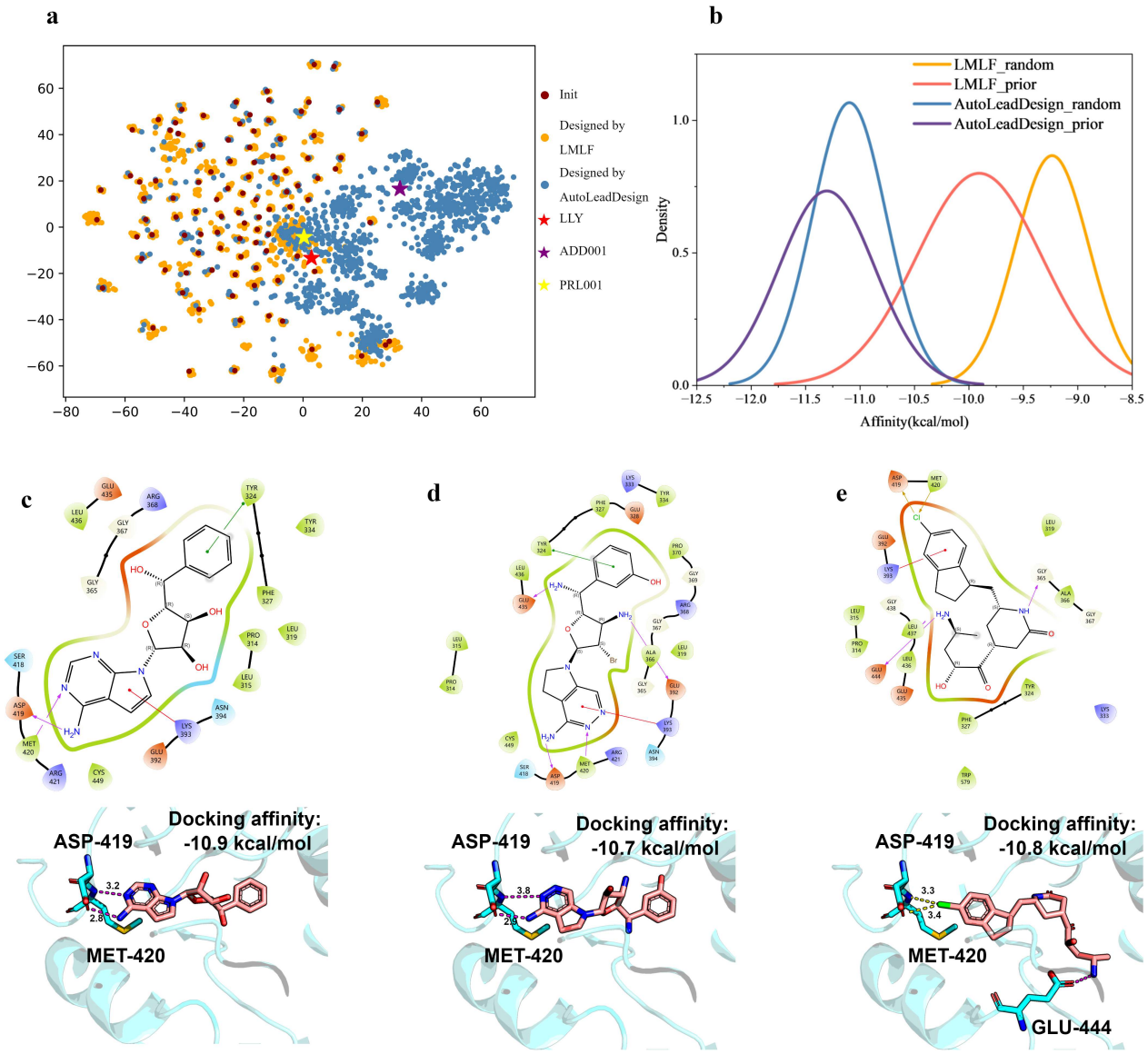


Fig. 2 Case study of PRMT5 in the scenario of prior initialization. In the visualization of binding interactions, purple lines stand for hydrogen bonds, yellow lines stand for halogen bonds, green lines stand for Pi-Pi Stacking, and red lines stand for Pi-cation interactions. a, The clustering result of designed compounds. b, The affinity distribution of designed compounds. c, The binding mode of LLY. d, The binding mode of PRL001. e, The binding mode of ADD001.

As shown in Fig. 2a, compounds designed by LMLF tend to cluster around the reference structure, demonstrating local optimization behavior. In contrast, AutoLeadDesign not only utilizes the known chemical space

around LLY but also explores uncharted regions. In addition, Fig. 2c-e depict the binding modes of compound LLY, along with PRL001 (designed by LMLF) and ADD001 (designed by AutoLeadDesign), in their interactions with the target protein. It can be seen that PRL001 exhibits a similar backbone and binding pattern to LLY. In contrast, ADD001 has a novel structure. In the precept of binding interactions, ADD001 not only maintains interactions with key residues ASP419 and MET420 (similar to LLY), but also establishes an additional hydrogen bond with the key residue GLU444 in the substrate protein binding site [45, 46, 48]. This dual binding mode represents a novel inhibition mechanism — simultaneously competing with both SAM cofactor and protein substrate binding. The results illustrate the ability of AutoLeadDesign in designing high-affinity compounds with novel interaction mechanisms, rather than merely optimizing known binding motifs.

To further evaluate the robustness of our approach, we construct a scenario with random initialization where we only randomly selected 100 compounds as start points. The visualization of results (Supplementary Fig. 1a) shows that compounds designed by LMLF still cluster around initial structures, indicating similar low bioactivity. This characteristic of local optimization results in its sensitivity to initialization (Fig. 2b), thereby limiting the capability of LLMs to design leads with novel binding modes to target proteins. By contrast, AutoLeadDesign exhibits exceptional proficiency in navigating uncharted chemical space, consistently yielding reliable results while demonstrating a key strength of low sensitivity to initial conditions. Interaction analysis of the top-ranked PRMT5 inhibitors (PRR001 and ADR001) yielded results similar to the previous observations, highlighting crucial interactions with key residues like GLU444. Further details and structural visualizations are available in Supplementary Fig 1. These results underscore the importance of comprehensive chemical space exploration in lead discovery. The superior performance of AutoLeadDesign, which does not depend on difficult-to-obtain bioactive ligands, represents a significant advancement over LMLF, offering greater potential for novel drug design.

The SARS-CoV-2 papain-like protease (PLpro), which mediates polyprotein processing, is indispensable for producing functional viral replicase complexes. Compared to existing antiviral therapeutics, PLpro-targeting agents may operate through divergent mechanisms of action, potentially providing improved therapeutic efficacy and resilience against emerging viral resistance [49–58]. We also exemplify the application of AutoLeadDesign in designing Plpro inhibitors in the setting of random initialization. The drug-like leads demonstrate superior binding affinities and favorable physicochemical profiles (e.g., QED, SAScore, and Lipinski compliance; see Supplementary Fig 2 for detailed distributions).

2.3 The Chemical Fragment Space Inspires LLM to Make Reasonable Decisions

The superior performance of AutoLeadDesign primarily stems from its fragment-inspired and LLM-guided design paradigm, where chemical fragments inspire extensive domain knowledge embedded in LLM to guide the drug design process. Here, we conduct an in-depth analysis of the LLM decision-making process from the perspective of fragment-based drug design (FBDD) strategies (particularly fragment linking and merging strategies), demonstrating the scientific rationality of LLM-driven decisions within this workflow.

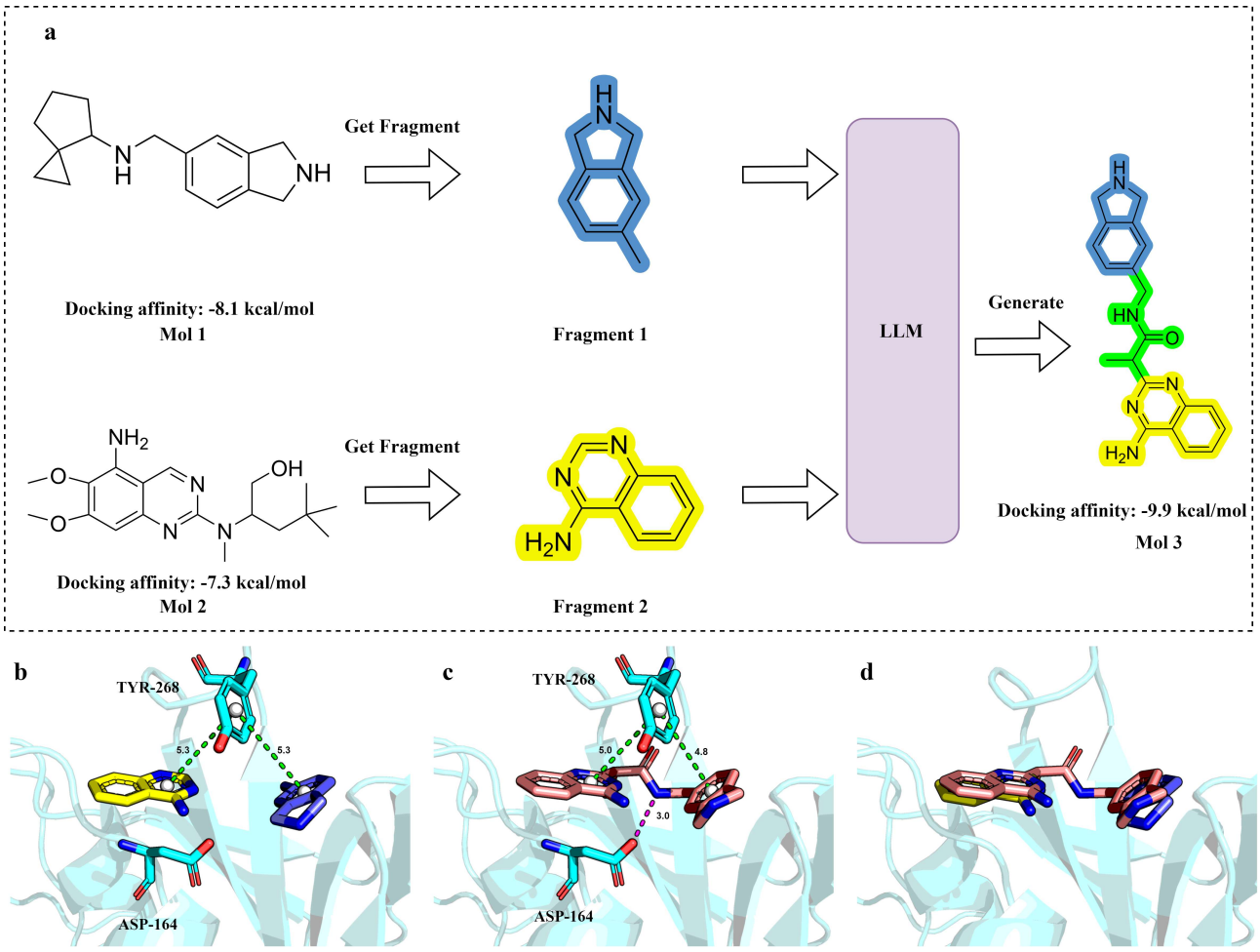


Fig. 3 The performance of LLM obeys the fragment linking strategy. In the visualization of binding interactions, purple lines stand for hydrogen bonds and green lines stand for Pi-Pi Stacking. a, A design process targeting PLpro. Fragment 1 and Fragment 2 are sampled, which are decomposed from Mol 1 and Mol 2, respectively. Then, two fragments are fed to LLM to design Mol 3. b, The binding mode of Fragment 1 (blue) and Fragment 2 (yellow). c, The binding mode of Mol 3. d, The different parts of Mol 3 overlap with two sampled fragments.

Fig. 3 illustrates the key decision-making process targeting PLpro. Fragment 1 and Fragment 2, derived from reference molecules Mol 1 and Mol 2, respectively (both exhibiting high binding potential), were selected as seed fragments and subsequently fed into LLM to generate Mol 3. Structural analysis revealed that the LLM designed a classic drug-like amide linker to bridge the two fragments. Detailed binding mode characterization demonstrated that both fragments occupy adjacent regions of the binding pocket and form $\pi-\pi$ stacking interactions with TYR-268 (Fig. 3b), exhibiting spatially proximal binding configurations. Notably, the amide linker not only preserves the native binding modes of the original fragments (Fig. 3d), but also forms an additional hydrogen bond with ASP-164 (Fig. 3c), significantly enhancing binding affinity ($\Delta G_{Mol1\to3} = -1.8\ kcal/mol$ and $\Delta G_{Mol2\to3} = -2.6\ kcal/mol$). Moreover, this amide-containing fragment will enrich the fragment library to guide subsequent molecular design iterations. This workflow rigorously adheres to fundamental fragment linking principles, providing experimental validation for the scientific rationality of LLM-driven decision-making in fragment-based drug design (FBDD). A similar decision-making case for PRMT5 leads demonstrates that a

strategic fragment merging process was employed to yield molecules that strictly adhere to fundamental FBDD principles (see Supplementary Fig 4 for detailed analysis).

In summary, detailed tracing of the decision-making process indicates that under the inspiration of sampled fragments, LLM intrinsically operationalizes the core tenets of fragment-based drug design (FBDD) with expert-like proficiency—mastering both fragment merging and linking strategies as canonical approaches.

2.4 The Exploration of Fragment Space

AutoLeadDesign demonstrates comprehensive capabilities in molecular design by not only implementing fragment linking and fragment merging strategies, but also exhibiting fragment growing-based progressive exploration of uncharted chemical space through evolutionary trajectory analysis of fragment library development.

The evolutionary trajectory of the fragment library targeting PLpro (Fig. 4a, c-e) demonstrates successful incorporation of a critical pyridine ring structure through multiple rounds of fragment recombination and decomposition (with Fragment 1 and Fragment 2 first integrated into the library at the 5th and 10th iterations, respectively). Detailed binding site analysis confirms this structural evolution: 1) Fragment 2 completely maintains the $\pi - \pi$ stacking interaction pattern with TYR-268 (Fig. 4e), showing high conformational similarity to the original fragment (RMSD = 1.1 Å); 2) the newly introduced pyridine ring forms strong $\pi - \pi$ stacking interactions with TYR-268 (centroid distance = 5.1 Å, dihedral angle = 86.1°), resulting in a significant reduction of binding free energy ($\Delta G = -1.1\ kcal/mol$). These findings conclusively validate the affinity enhancement achieved by AutoLeadDesign. Study targeting the PRMT5 similarly reveals characteristic fragment evolutionary patterns (Figs. 4b, f-h). This fragment evolution process strictly adheres to the theoretical framework of fragment growing—a localized exploration paradigm that optimizes known fragments to achieve high-affinity derivatives. Furthermore, the affinity distributions of fragment libraries targeting PLpro (Fig. 4i) and PRMT5 (Fig. 4j) provide further evidence that the dual exploration of both uncharted and local fragment spaces systematically drives the fragment library updating toward desired pharmacological profiles.

2.5 From In Silico Prediction to In Vitro Evaluation

Following the robust performance of AutoLeadDesign in identifying novel scaffolds in silico, we sought to empirically validate its efficacy through laboratory synthesis and biological testing. We targeted two proteins of immense clinical significance: PLpro, essential for viral replication, and KRAS G12D, a notoriously undruggable oncogenic driver.

To validate the antiviral potency of the molecules generated by AutoLeadDesign targeting PLpro, we selected the most promising candidate for chemical synthesis. From the generated molecule set, we applied dual criteria of binding affinity ($\leq -9kcal/mol$) and synthetic accessibility ($SAscore \leq 4$), yielding 96 candidate structures that met both requirements. Next, we examined the interaction patterns between the docked conformations of these molecules and the PLpro binding pocket (Supplementary Fig. 1). Molecules exhibiting key interactions with residues such as TYR268 and ASP164, as well as adjacent residues, were prioritized. Finally, after evaluating the synthetic feasibility of their proposed routes, the most promising candidate, named PLP011, was selected for chemical synthesis. The detailed synthetic routes can be found in Appendix C. As shown in Fig. 5a, PLP011

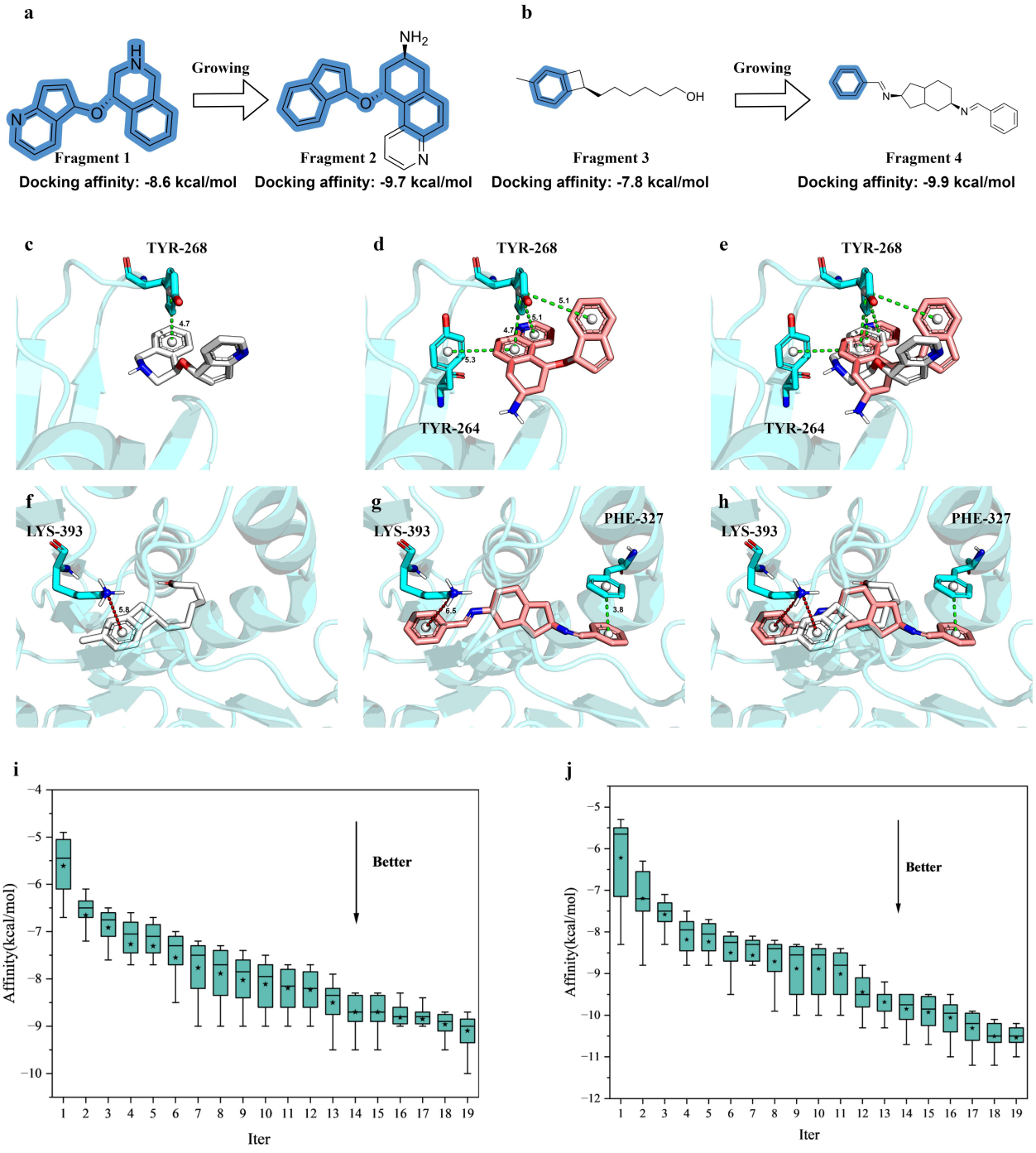


Fig. 4 The performance of LLM obeys the fragment growing strategy. In the visualization of binding interactions, green lines stand for Pi-Pi Stacking and red lines stand for Pi-cation interactions. a, A trace of fragment library targeting PLpro. Fragment 1 first appears in iteration 5. After many iterations of updating, Fragment 2 appeared in iteration 10. b, A trace of fragment library targeting PRMT5. Fragment 3 first appears in iteration 7. After many iterations of updating, Fragment 4 appeared in iteration 11. c, The binding mode of Fragment 1. d, The binding mode of Fragment 2. e, Fragment 2 overlaps with Fragment 1. f, The binding mode of Fragment 3. g, The binding mode of Fragment 4. h, Fragment 3 overlaps with Fragment 4. i, The updating of the fragment library targeting PLpro. j, The updating of the fragment library targeting PRMT5.

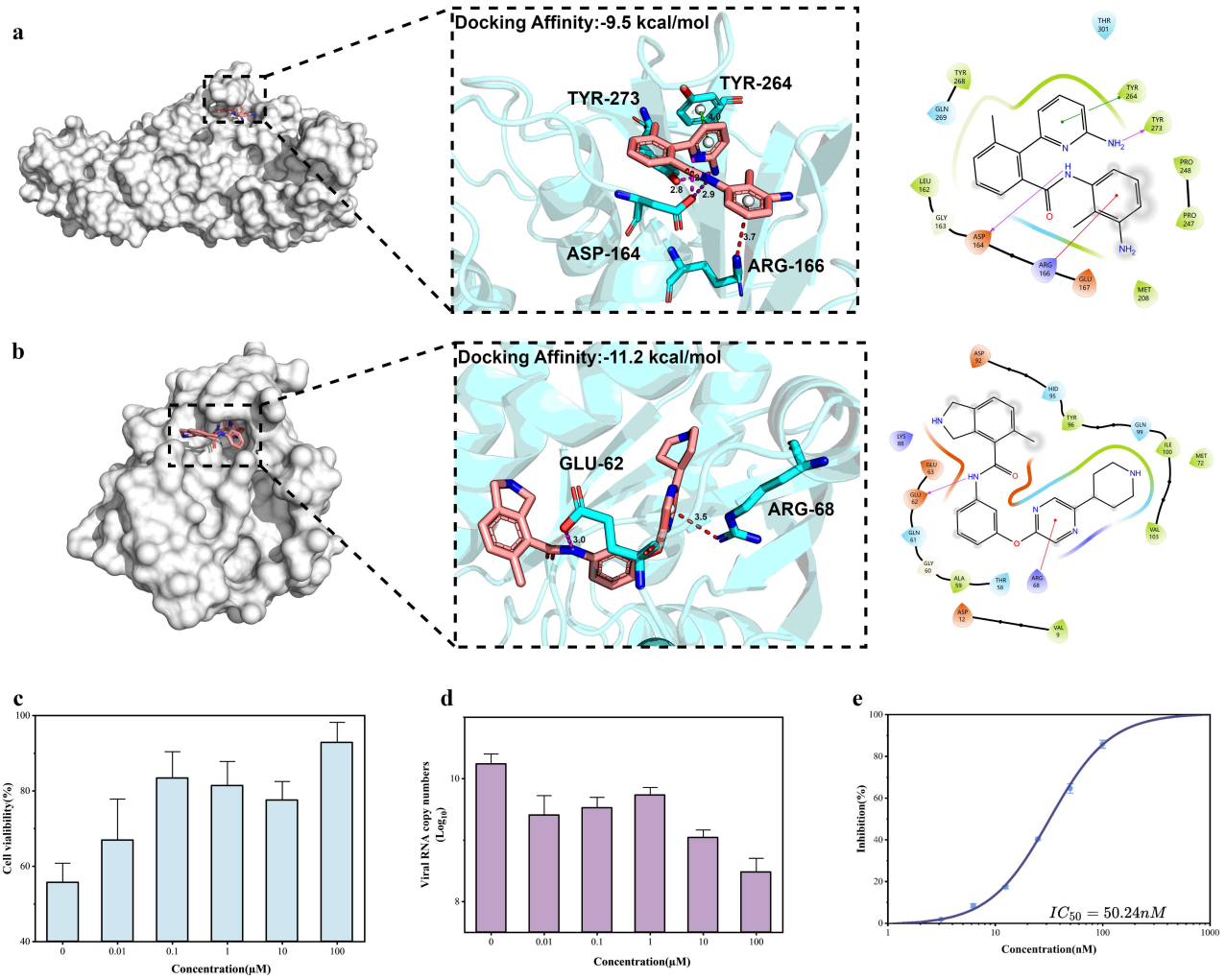


Fig. 5 The binding pose and the cell experiment result of the designed compounds. a, the interaction mode of PLP001 targeting PLpro. b, the interaction mode of KP032 targeting KRAS G12D. c, The viability of cells infected by SARS-CoV-2 at varying compound concentrations. d, The virtual RNA copy numbers of SARS-CoV-2 at varying compound concentrations. e, The Dose-dependent inhibition of KRAS G12D by the KP011.

effectively occupies the target binding pocket and interacts with ASP-164, ARG-166, TYR-264, and TYR-273 through key hydrogen bonds, $\pi-\pi$ and $\pi-$stacking interactions, indicating strong binding affinity to the target. Molecular Dynamics simulations (Supplement Fig. 5) showed that PLP011 remained stable (root mean square deviation $\leq 0.16nm$), with no exceptional movements (low root mean square fluctuation) in the simulation time.

The antiviral efficacy of PLP011 was systematically evaluated through cellular assays conducted in an animal biosafety level 3 (ABSL3) in Changchun Veterinary Research Institute, Chinese Academy of Agricultural Sciences. As illustrated in Fig. 5c, PLP011 demonstrated a significant, dose-dependent restorative effect on cells infected with SARS-CoV-2. At a low concentration of 0.01 μM, cell viability began to recover compared to the untreated control. Notably, the average cell viability reached 81.4% at a concentration of 1 μM, and further increased to nearly 93% at the maximum tested concentration of 100 μM, underscoring the compound's low cytotoxicity and strong protective capability. Complementing the cell viability data, Fig. 5d quantifies the inhibition of viral proliferation by measuring Viral RNA copy numbers (Log_{10}). The results reveal a substantial reduction in viral load as the concentration of PLP011 increases. Specifically, compared to the initial viral titer

in the control group (0 μM), PLP011 treatment induced a progressive decline in RNA copies, achieving a potent suppression of viral replication at the 100 μM dosage level. Collectively, the concentration-dependent trends observed in both the cell survival and viral inhibition assays confirm that PLP011 is a highly effective antiviral candidate with a significant therapeutic window.

Kirsten rat sarcoma virus gene (KRAS), particularly the G12D mutation, is a key driver in approximately one-third of human tumors. This mutation inhibits GTP hydrolysis, causing the protein to remain in a constitutively active state that fuels downstream oncogenic signaling [59, 60]. However, KRAS G12D has long been perceived as a formidable therapeutic target. Its characterization as 'undruggable' stems from its relatively featureless protein surface, which lacks the deep clefts necessary for stable ligand interaction [61]. Current strategies for developing KRAS G12D inhibitors frequently draw inspiration from the structural optimization of established G12C inhibitors. Building on this paradigm, we curated a set of small-molecule G12C inhibitors from existing literature to serve as initial compounds for our optimization. The generated molecules were rigorously filtered using a binding affinity threshold (≤ -9 kcal/mol) and a synthetic accessibility constraint (SA score ≤ 4). Among the 22 candidates that met these criteria, we performed detailed molecular interaction simulations and sought expert evaluation of their synthetic routes. Ultimately, compound KP032 was prioritized and selected for chemical synthesis based on its superior binding mode and synthetic feasibility. The detailed synthetic routes can be found in Appendix C. As shown in Fig. 5b, KP032 effectively occupies the Switch II pocket and interacts with GLU-62, ARG-68 through key hydrogen bond, $\pi-$stacking interactions. These robust molecular interactions underscore the compound's strong binding affinity for the KRAS G12D target. As shown in Fig. 5e, the compound exhibited a classic sigmoidal dose-response profile, indicating a clear, concentration-dependent suppression of the target's activity. Non-linear regression analysis of the inhibition curve yielded a half-maximal inhibitory concentration (IC_{50}) of 50.24 nM.

3 Discussion

In this work, we propose AutoLeadDesign, a novel property-conditioned molecular design algorithm that inspires large language models (LLMs) through molecular fragments while leveraging extensive domain knowledge embedded in LLMs to systematically explore fragment-based chemical space. This iterative closed-loop framework empowers the LLM to effectively synchronize with fragment-based heuristics, fostering a knowledge-driven expansion into the target-centric chemical space. As fragments prime the LLM's generative logic, the system autonomously identifies and explores potential chemical subspaces, systematically optimizing binding interactions while maintaining structural reasonability without the need for known ligands. Finally, the fragment libraries developed for our study's therapeutic targets have been released as a resource for the broader research community. These libraries function as critical structural priors, empowering medicinal chemists and facilitating the refinement of deep learning models by providing a high-quality starting point for exploring the chemical space surrounding these clinically relevant proteins.

Although AutoLeadDesign has shown its power in the model comparison, in silico analysis, and in vitro evaluation, demonstrating a successful application of knowledge embedded within the LLMs to drug discovery, it still faces some limitations. One notable challenge is LLM's indirect involvement in the deep reasoning process. Future improvement will focus on constructing a multi-agent framework to facilitate step-wise reasoning

and iterative summarization. Additionally, AutoLeadDesign primarily optimizes for docking scores, which may hinder the integration of medicinal chemists' prior knowledge regarding specific target sites during the design process. Future development will focus on incorporating explicit interaction constraints, such as Non-covalent bonds with key residues identified by medicinal chemists.

4 Methods

4.1 Overview of AutoLeadDesign

Fig. 1 illustrates the workflow of AutoLeadDesign, a property-conditioned lead discovery algorithm that operates as a closed-loop iterative framework to bridge the generative capacity of LLMs with the principles of fragment-based drug design. The workflow of this framework is structured as a continuous evolutionary cycle comprising four distinct stages: molecular decomposition, statistical characterization update of the fragment library, heuristic fragment sampling, and LLM-driven molecule recombination. This cyclical process can be fundamentally conceptualized as two bidirectional mappings between molecule space and fragment space that systematically navigate uncharted chemical landscapes.

The mapping from molecule space to fragment space, as shown in Fig. 1b, serves to deconstruct explored chemical entities into their constituent pharmacophoric elements to increase design flexibility. This transformation involves two primary steps: first, molecules within the evolving compound library, as shown in Fig. 1a, are decomposed into structural fragments based on predefined rules; second, the framework quantitatively evaluates fragment occurrence frequencies and their individual contributions to binding energy. By calculating these metrics, AutoLeadDesign continuously updates a statistical characterization of the target-relevant fragment space, effectively distilling global ligand data into a library of high-priority building blocks.

Conversely, the mapping from fragment space to molecule space, as shown in Fig. 1d, leverages the domain-specific knowledge encoded within the LLM to transition from discrete fragments back to integrated, high-affinity leads. This mapping begins with the systematic sampling of fragments from the filtered library, as shown in Fig. 1c, based on their probabilistic weights, which are derived from their historical binding performance. These selected fragments are then incorporated into structured prompts that guide the LLM to perform biochemical reasoning.

By iteratively navigating both fragment and molecule spaces, the framework creates a reciprocal feedback loop where fragments inform LLMs, while newly designed molecules, in turn, refine the framework's understanding of the fragment space. In this process, the prioritized fragment library provides heuristic guidance, offering the LLM a targeted directional anchor within high-potential yet uncharted chemical regions. By integrating these high-contribution fragments with its intrinsic domain knowledge, the LLM proposes novel molecular scaffolds that satisfy specific structural constraints. Conversely, as these newly designed molecules are deconstructed, they serve as a critical feedback mechanism that updates the framework's comprehension of the fragment space. This dual-mapping paradigm ensures that LLM-driven biochemical reasoning and fragment-level statistical insights mutually reinforce each other, facilitating a more comprehensive and interpretable exploration for de novo drug discovery against targets lacking established bioactive ligands.

4.2 Molecules decomposition

Molecular decomposition aims to construct a fragment library from current molecules, thereby establishing and iteratively updating the framework's understanding of target-specific preferences within the chemical space. The strategy in AutoLeadDesign is grounded in the BRICS (Breaking of Retrosynthetically Interesting Chemical Substructures) algorithm [62], a fragmentation methodology that advances the principles of RECAP (Retrosynthetic Combinatorial Analysis Procedure) by partitioning molecules according to 16 predefined cleavage rules. These rules target specific bond types that correspond to high-yielding synthetic reactions. By selectively cleaving these bonds, ensuring that the resulting fragment space is composed of synthetically accessible building blocks rather than arbitrary subgraphs.

While the BRICS algorithm provides a robust framework for retrosynthetic cleavage, its systematic application often generates chemically trivial fragments. Such fragments, including isolated methyl groups or single halide atoms, appear with high frequency regardless of target specificity. Their decomposition of these fragments from high-affinity molecules can lead the framework to erroneously identify them as key structural motifs. Without intervention, this phenomenon introduces a significant bias in the characterization of the fragment space, potentially misleading the generative process toward non-informative structural fragments. To preserve the integrity of the fragment library's statistical characterization and ensure the extraction of meaningful chemical information, we implemented a structural constraint requiring each fragment to contain at least two heavy atoms. By enforcing this size threshold, the framework effectively filters out structural noise, ensuring that the fragment library remains focused on pharmacophoric units with sufficient complexity to provide valid guidance for navigating the chemical space.

4.3 Fragment library update and sample

The fragment library at the t-th iteration, constructed through the systematic decomposition of molecules from all preceding $t-1$ cycles, serves as a repository that reflects the framework's cumulative comprehension of the target-specific fragment space. However, this accumulation leads to an excessively large and redundant collection of structural fragments. Inclusion of the entire global fragment pool as valid input for molecular generation inevitably triggers a multitude of unproductive explorations in low-potential chemical regions. Such exploration not only dilutes the influence of high-value structural insights but also diminishes the efficiency of the LLM in identifying truly novel leads. To focus on potential fragments, we adopt the assumption that fragments derived from compounds with lower binding affinity generally possess a reduced probability of recombining into high-affinity leads. We established a filtered library containing the top-k highest-affinity fragments selected from the global fragment pool.

Furthermore, to quantitatively evaluate fragment significance, we assessed their individual binding contributions to the target protein by analyzing the energy distribution of the molecular ensembles in which they reside. The core objective of this heuristic evaluation is to distill the intrinsic binding potential of discrete fragments from the collective performance of their parent compounds. We employ a Monte Carlo sampling approach to estimate its significance based on a representative subset of the chemical space. Specifically, we leverage the docking scores of all explored compounds containing the specific fragment f to reflect the probability distribution of that fragment's importance to target binding. By calculating the average binding energy across this

diverse molecular set, we effectively marginalize the influence of varied chemical environments, thereby isolating the fragment's consistent contribution to protein-ligand interactions. This scoring metric serves as a statistical proxy for the fragment's pharmacophoric value, providing a quantitative basis for the subsequent priority-based sampling within the AutoLeadDesign framework. The score is calculated as:

$$Score(f) = \frac{\sum_{s \in C, f \in Decompose(s)} Docking(s, p)}{|C|}, \tag{1}$$

where C stands for the compounds library containing all previously designed compounds until this iteration, and $Decompose(s)$ stands for the fragment set decomposed from compound s, and $Docking(s, p)$ stands for the binding energy of the compound s targeting the protein p.

To translate fragment evaluations into actionable design guidance while enhancing the robustness of the framework, we conceptualize the fragment selection process as a stochastic sampling task within the fragment space. Rather than adopting a deterministic greedy approach, which exclusively selects the highest-scoring fragments and risks converging on local optima, we model the fragment distribution as a categorical distribution. In this probabilistic framework, each fragment in the filtered library L is assigned a sampling weight derived from its importance score, ensuring that high-contribution fragments are prioritized while preserving the potential for structural innovation. The sampling weight is calculated based on the scores:

$$Weight(f) = \frac{Score(f)}{\sum_{f_i \in L} Score(f_i)}, \tag{2}$$

Through this systematic sampling process, AutoLeadDesign effectively distills a refined understanding of target-specific preferences from the currently explored chemical space. By utilizing the probability distribution as a robust heuristic prior, the framework provides the LLM with a targeted search guidance that prioritizes regions of high binding potential. This guided approach ensures that the generative process remains focused on promising chemical scaffolds, allowing the LLM to transition from blind searching to an informed exploration of uncharted chemical space.

4.4 Molecules recombination

To enhance the structural reasonability and drug-likeness of the generated candidates, we use knowledge embedded within LLMs to guide the design process. Since LLMs inherently lack a 3D structural understanding of target proteins, we utilize structural fragments as heuristic priors to anchor the generation in regions with high binding potential. By encoding these fragments into the SMILES-based prompt, we enforce a structural bias that steers the LLM's sampling process. This mechanism ensures that the model explores a chemically diverse yet target-relevant chemical space, utilizing its knowledge to maintain synthetic feasibility and pharmacophore integrity while adhering to the provided structural constraints.

We have developed a standardized prompt template that, when integrated with sampled molecular fragments, guides the LLM to effectively explore the chemical space. Typically, the template is presented as follows:

> Generate a novel and valid molecule SMILES that contains one fragment of [$fragment[0]$, $fragment[1]$, $fragment[2]$] at least, and do not generate any English text.

where $fragment[0]$, $fragment[1]$ and $fragment[2]$ will be filled with the SMILES of sampled fragments.

The large language model (LLM) employed in our study is DeepSeek-v3, utilizing the "deepseek-chat" conversational model with a temperature parameter set to 1.5.

4.5 Dataset setup

Random initial compounds library. To initialize the AutoLeadDesign pipeline, we randomly sampled 100 compounds from the ZINC100K database, and determined their protein binding affinities through docking simulations, providing the foundational start points for subsequent design cycles.

Target protein. We randomly select 10 target proteins in the CrossDocked2022 database [42]. These proteins belong to distinct classes: hydrolase, transferase, oxidoreductase, member protein, and immune system protein. The PDB IDs are 4AAW, 14GS, 1FMC, 2AZY, 1PHK, 3U5Y, 2HCJ, 3KC1, 3O96, and 3HY9. In the case study, we chose three therapeutically significant targets: PRMT5 (PDB ID: 8UOB), PLpro (PDB ID: 7L1G), and KRAS G12D (PDB ID: 7RPZ), all representing important clinical targets in their respective disease areas.

Docking simulation. We employed smina (a fork of Autodock Vina focused on improving scoring functions and energy minimisation) to approximate the Gibbs free energy changes during ligand-receptor binding[63]. Prior to docking, all ligands were converted from SMILES to 3D conformations using RDKit (v2023.9.6) with the ETKDG algorithm to ensure proper stereochemistry. These prepared conformations were then subjected to docking simulations using smina, with the binding site defined by the co-crystallized ligand coordinates with a 1 Å grid box expansion to ensure complete binding pocket coverage. Finally, the highest-affinity protein-ligand complexes were subjected to detailed binding interaction analysis in Maestro v14.2.

4.6 The settings of compared methods

REINVENT [22]. Building on policy-based reinforcement learning, REINVENT adapts recurrent neural networks for the de novo molecular design task. The framework's innovative use of augmented episodic likelihood enables the directed exploration of chemical space, designing compounds that satisfy predefined property requirements.

ChemGE[23]. ChemGE implements a population-based grammatical evolution strategy, where molecular structures undergo iterative optimization through evolutionary operations on their grammatical representations, facilitating directed exploration of chemical space.

Above baselines can be run using the software ($https://github.com/wenhao-gao/mol_opt$) in practical molecular optimization benchmark [64].

RGA[21]. RGA attempts to reformulate an evolutionary process as a Markov decision process and uses an E(3)-equivariant neural network to choose parents and mutation types based on the 3D structure of the ligands and proteins. The model weight was obtained from their public repository ($https://github.com/futianfan/reinforced-genetic-algorithm$).

LMLF[39]. LMLF separates the prompt into domain-constraints that can be written in a standard logical form, and a simple text-based query. The constraints are progressively refined using a logical notion of generalization to progressively design compounds. The settings of LMLF can be loaded from their repository ($https://github.com/Shreyas-Bhat/LMLF$).

Notably, to standardize the optimization objectives across all the frameworks, we unified the molecular property calculation module to focus exclusively on binding affinity with the target protein, which is calculated by smina.

4.7 Visualization

Clustering analysis. To visualize and compare the chemical space of the initial and designed compounds, we constructed t-SNE plots. The 2048-bit Morgan fingerprints were used as molecular descriptors. Employing the Barnes-Hut implementation of the t-SNE algorithm, we obtained two-dimensional representations for clustering analysis.

2D ligand-residue interaction. For detailed analysis of compound-target protein interactions, the highest-affinity molecular conformation selected by SMINA docking was further analyzed using Maestro v14.2. Typically, hydrogen bonds were identified with a maximum bond length of 2.8 Å, requiring minimum angles of 120° for donors (D-H→A) and 90° for acceptors (H→A-X); halogen bonds were detected with a 3.5 Å distance cutoff and similar angular constraints of ≥120° for donors (R-X→A) and ≥ 90° for acceptors (X→A-Y); $\pi-\pi$ interactions were classified as parallel stacking when within 4.4 Å with ≤ 30° angular deviation or as T-shaped stacking when within 5.5 Å with ≥60° angles; while $\pi-cation$ interactions were considered significant within 6.6 Å and maximum 30° angular deviation. These carefully optimized parameters ensured comprehensive and accurate characterization of all potential molecular interactions at the binding interface.

3D ligand-pocket binding pose. Structural visualization of the ligand-protein complex was performed using PyMOL v3.1 (Schrödinger, LLC), while all intermolecular interactions were computationally characterized using Maestro.

4.8 Chemical synthesis

We evaluated the synthetic feasibility of the designed compounds using the Synthetic Accessibility (SA) score and further validated the results with manual inspection by medicinal chemists. All the small molecules were synthesized by WuXi AppTec Company, with detailed synthetic routes described in the Supplementary Methods.

4.9 Cellular assays in vitro

Antiviral Activity of PLP011. The in vitro antiviral studies were carried out in an animal biosafety level 3 (ABSL3) in Changchun Veterinary Research Institute, Chinese Academy of Agricultural Sciences. Viral infection was established in live mammalian cells, followed by the measurement of cell viability and viral RNA copy numbers under a gradient of drug concentrations. These assays were designed to determine the compound's ability to inhibit viral proliferation and restore host cell health.

Biochemical Inhibition of KRAS G12D by KP032. The inhibitory potency of KP032 against the oncogenic KRAS G12D mutation was evaluated by measuring its ability to disrupt the interaction between KRAS G12D and its effector, cRAF. This was performed using a Time-Resolved Fluorescence Resonance Energy Transfer (TR-FRET) assay (Human KRAS G12D & cRAF Binding Kit, Absin). Briefly, KP032 was serially diluted

and incubated with Tag2-KRAS G12D (pre-loaded with GTP) and Tag1-cRAF protein, followed by the addition of Eu-labeled anti-Tag1 and Ac-labeled anti-Tag2 antibodies. TR-FRET signals were recorded at 620 nm and 665 nm to quantify the competitive inhibition.

Full details regarding cell culture conditions, viral titers, antibody concentrations, and instrumentation for both assays are provided in the Supplementary Methods and Appendix C.

4.10 Released fragment libraries

To facilitate future drug design efforts, we have compiled and released target-specific fragment libraries derived from the generative trajectories of AutoLeadDesign. These libraries were constructed through an incremental expansion process during execution, heuristically exploring the target-relevant fragment space. Each library entry integrates a structural fragment with the collection of candidates containing that motif, supplemented by their calculated binding affinities. This resource is intended to function as a structural prior, offering medicinal chemists and future machine learning frameworks high-quality, domain-specific heuristics to navigate the complex chemical landscapes of high-profile therapeutic targets.

5 Data availability

The code and all the data, including initial compounds, target proteins, and results, are available at https://github.com/TH1052638958/AutoLeadDesign.

6 Author contributions

Hao Tuo, Haishi Zhao and Bo Yang contributed to the idea. Hao Tuo contributed to the code. Hao Tuo, Haishi Zhao, Yan Li, Xuanning Hu, Xueyan Liu and Bo Yang provide advice on in silico experiment. Geng Qing and Feng Wei were responsible for the chemical synthesis of the generated leads. Geng Qin and Feng Wei performed the in vitro biochemical assay targeting KRAS G12D. Beilei Shen performed the cell-based viral infection assays targeting PLpro. Bo Yang and Xueyan Liu contributed to the manuscript revision.

References

[1] Mak, K.-K. & Pichika, M. R. Artificial intelligence in drug development: present status and future prospects. *Drug discovery today* **24**, 773–780 (2019).

[2] Keserü, G. M. & Makara, G. M. The influence of lead discovery strategies on the properties of drug candidates. *nature reviews Drug Discovery* **8**, 203–212 (2009).

[3] Corsello, S. M., Bittker, J. A., Liu, Z., Gould, J., McCarren, P., Hirschman, J. E., Johnston, S. E., Vrcic, A., Wong, B., Khan, M. *et al.* The drug repurposing hub: a next-generation drug library and information resource. *Nature medicine* **23**, 405–408 (2017).

[4] Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B. *et al.* Pubchem 2023 update. *Nucleic acids research* **51**, D1373–D1380 (2023).

[5] Irwin, J. J. & Shoichet, B. K. Zinc- a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling* **45**, 177–182 (2005).

[6] Hajduk, P. J. & Greer, J. A decade of fragment-based drug design: strategic advances and lessons learned. *Nature reviews Drug discovery* **6**, 211–219 (2007).

[7] Flemming, A. Resistance-guided discovery of new antibiotics. *Nature Reviews Drug Discovery* **12**, 826–826 (2013).

[8] Beckett, R. P., Davidson, A. H., Drummond, A. H., Huxley, P. & Whittaker, M. Recent advances in matrix metalloproteinase inhibitor research. *Drug Discovery Today* **1**, 16–26 (1996).

[9] Popov, K. I., Wellnitz, J., Maxfield, T. & Tropsha, A. Hit discovery using docking enriched by generative modeling (hidden gem): A novel computational workflow for accelerated virtual screening of ultra-large chemical libraries. *Molecular informatics* **43**, e202300207 (2024).

[10] Sadybekov, A. A., Sadybekov, A. V., Liu, Y., Iliopoulos-Tsoutsouvas, C., Huang, X.-P., Pickett, J., Houser, B., Patel, N., Tran, N. K., Tong, F. *et al.* Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. *Nature* **601**, 452–459 (2022).

[11] Bender, B. J., Gahbauer, S., Luttens, A., Lyu, J., Webb, C. M., Stein, R. M., Fink, E. A., Balius, T. E., Carlsson, J., Irwin, J. J. *et al.* A practical guide to large-scale docking. *Nature protocols* **16**, 4799–4832 (2021).

[12] Yoon, S., Smellie, A., Hartsough, D. & Filikov, A. Surrogate docking: structure-based virtual screening at high throughput speed. *Journal of computer-aided molecular design* **19**, 483–497 (2005).

[13] Lyu, J., Wang, S., Balius, T. E., Singh, I., Levit, A., Moroz, Y. S., O'Meara, M. J., Che, T., Algaa, E., Tolmachova, K. *et al.* Ultra-large library docking for discovering new chemotypes. *Nature* **566**, 224–229 (2019).

[14] Popov, K. I., Wellnitz, J., Maxfield, T. & Tropsha, A. Hit discovery using docking enriched by generative modeling (hidden gem): A novel computational workflow for accelerated virtual screening of ultra-large chemical libraries. *Molecular informatics* **43**, e202300207 (2024).

[15] Schönherr, H., Ayaz, P., Taylor, A. M., Casaletto, J. B., Touré, B. B., Moustakas, D. T., Hudson, B. M., Valverde, R., Zhao, S., O'Hearn, P. J. *et al.* Discovery of lirafugratinib (rly-4008), a highly selective irreversible small-molecule inhibitor of fgfr2. *Proceedings of the National Academy of Sciences* **121**, e2317756121 (2024).

[16] Wu, K., Xia, Y., Deng, P., Liu, R., Zhang, Y., Guo, H., Cui, Y., Pei, Q., Wu, L., Xie, S. *et al.* Tamgen: drug design with target-aware molecule generation through a chemical language model. *Nature Communications* **15**, 9360 (2024).

[17] Huang, L., Xu, T., Yu, Y., Zhao, P., Chen, X., Han, J., Xie, Z., Li, H., Zhong, W., Wong, K.-C. *et al.* A dual diffusion model enables 3d molecule generation and lead optimization based on target pockets. *Nature Communications* **15**, 2657 (2024).

[18] Du, Y., Jamasb, A. R., Guo, J., Fu, T., Harris, C., Wang, Y., Duan, C., Liò, P., Schwaller, P. & Blundell, T. L. Machine learning-aided generative molecular design. *Nature Machine Intelligence* **6**, 589–604 (2024).

[19] Jin, W., Barzilay, R. & Jaakkola, T. Junction tree variational autoencoder for molecular graph generation 2323–2332 (2018).

[20] Spiegel, J. & Durrant, J. Autogrow4: an open-source genetic algorithm for de novo drug design and lead optimization. j cheminform 12 (1): 25 (2020).

[21] Fu, T., Gao, W., Coley, C. & Sun, J. Reinforced genetic algorithm for structure-based drug design. *Advances in Neural Information Processing Systems* **35**, 12325–12338 (2022).

[22] Olivecrona, M., Blaschke, T., Engkvist, O. & Chen, H. Molecular de-novo design through deep reinforcement learning. *Journal of cheminformatics* **9**, 1–14 (2017).

[23] Yoshikawa, N., Terayama, K., Sumita, M., Homma, T., Oono, K. & Tsuda, K. Population-based de novo molecule generation, using grammatical evolution. *Chemistry Letters* **47**, 1431–1434 (2018).

[24] Williams, C. Y., Miao, B. Y., Kornblith, A. E. & Butte, A. J. Evaluating the use of large language models to provide clinical recommendations in the emergency department. *Nature Communications* **15**, 8236 (2024).

[25] Zheng, Z., Rampal, N., Inizan, T. J., Borgs, C., Chayes, J. T. & Yaghi, O. M. Large language models for reticular chemistry. *Nature Reviews Materials* 1–13 (2025).

[26] Wu, Z., Zhang, O., Wang, X., Fu, L., Zhao, H., Wang, J., Du, H., Jiang, D., Deng, Y., Cao, D. *et al.* Leveraging language model for advanced multiproperty molecular optimization via prompt engineering. *Nature Machine Intelligence* 1–11 (2024).

[27] Zheng, Y., Koh, H. Y., Ju, J., Yang, M., May, L. T., Webb, G. I., Li, L., Pan, S. & Church, G. Large language models for drug discovery and development. *Patterns* **6** (2025).

[28] Lee, N., De Brouwer, E., Hajiramezanali, E., Biancalani, T., Park, C. & Scalia, G. *Rag-enhanced collaborative llm agents for drug discovery*, Vol. 40, 561–569 (2026).

[29] Castro Nascimento, C. M. & Pimentel, A. S. Do large language models understand chemistry? a conversation with chatgpt. *Journal of Chemical Information and Modeling* **63**, 1649–1655 (2023).

[30] Ramos, M. C., Collison, C. J. & White, A. D. A review of large language models and autonomous agents in chemistry. *Chemical Science* (2025).

[31] M. Bran, A., Cox, S., Schilter, O., Baldassari, C., White, A. D. & Schwaller, P. Augmenting large language models with chemistry tools. *Nature Machine Intelligence* **6**, 525–535 (2024).

[32] Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A. S., Ceder, G., Persson, K. A. & Jain, A. Structured information extraction from scientific text with large language models. *Nature Communications* **15**, 1418 (2024).

[33] Sun, L., Huang, Y., Wang, H., Wu, S., Zhang, Q., Gao, C., Huang, Y., Lyu, W., Zhang, Y., Li, X. *et al.* Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561* **3** (2024).

[34] Guo, T., Nan, B., Liang, Z., Guo, Z., Chawla, N., Wiest, O., Zhang, X. *et al.* What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems* **36**, 59662–59688 (2023).

[35] Zhang, X., Wang, L., Helwig, J., Luo, Y., Fu, C., Xie, Y., Liu, M., Lin, Y., Xu, Z., Yan, K. *et al.* Artificial intelligence for science in quantum, atomistic, and continuum systems. *arXiv preprint arXiv:2307.08423* (2023).

[36] Wu, Z., Zhang, O., Wang, X., Fu, L., Zhao, H., Wang, J., Du, H., Jiang, D., Deng, Y., Cao, D. *et al.* Leveraging language model for advanced multiproperty molecular optimization via prompt engineering. *Nature Machine Intelligence* 1–11 (2024).

[37] Bai, X., Xie, Y., Zhang, X., Han, H. & Li, J.-R. Evaluation of open-source large language models for metal–organic frameworks research. *Journal of Chemical Information and Modeling* **64**, 4958–4965 (2024).

[38] Yang, F., Zhao, P., Wang, Z., Wang, L., Zhang, J., Garg, M., Lin, Q., Rajmohan, S. & Zhang, D. Empower large language model to perform better on industrial domain-specific question answering. *arXiv preprint arXiv:2305.11541* (2023).

[39] Brahmavar, S. B., Srinivasan, A., Dash, T., Krishnan, S. R., Vig, L., Roy, A. & Aduri, R. Generating novel leads for drug discovery using llms with logical feedback. *Proceedings of the AAAI Conference on Artificial Intelligence* **38**, 21–29 (2024).

[40] Gao, B., Huang, Y., Liu, Y., Xie, W., He, B., Tan, H., Ma, W.-Y., Zhang, Y.-Q. & Lan, Y. *Cidd: Collaborative intelligence for structure-based drug design empowered by llms* (2025).

[41] Rees, D. C., Congreve, M., Murray, C. W. & Carr, R. Fragment-based lead discovery. *Nature Reviews Drug Discovery* **3**, 660–672 (2004).

[42] Francoeur, P. G., Masuda, T., Sunseri, J., Jia, A., Iovanisci, R. B., Snyder, I. & Koes, D. R. Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. *Journal of chemical information and modeling* **60**, 4200–4215 (2020).

[43] Wang, Y., Hu, W. & Yuan, Y. Protein arginine methyltransferase 5 (prmt5) as an anticancer target and its inhibitor discovery. *Journal of medicinal chemistry* **61**, 9429–9441 (2018).

[44] Zheng, J., Li, B., Wu, Y., Wu, X. & Wang, Y. Targeting arginine methyltransferase prmt5 for cancer therapy: updated progress and novel strategies. *Journal of Medicinal Chemistry* **66**, 8407–8427 (2023).

[45] Wu, Q., Schapira, M., Arrowsmith, C. H. & Barsyte-Lovejoy, D. Protein arginine methylation: from enigmatic functions to therapeutic targeting. *Nature reviews Drug discovery* **20**, 509–530 (2021).

[46] Feustel, K. & Falchook, G. S. Protein arginine methyltransferase 5 (prmt5) inhibitors in oncology clinical trials: a review. *Journal of Immunotherapy and Precision Oncology* **5**, 58–67 (2022).

[47] Bonday, Z. Q., Cortez, G. S., Grogan, M. J., Antonysamy, S., Weichert, K., Bocchinfuso, W. P., Li, F., Kennedy, S., Li, B., Mader, M. M. *et al.* Lly-283, a potent and selective inhibitor of arginine methyltransferase 5, prmt5, with antitumor activity. *ACS medicinal chemistry letters* **9**, 612–617 (2018).

[48] Brehmer, D., Beke, L., Wu, T., Millar, H. J., Moy, C., Sun, W., Mannens, G., Pande, V., Boeckx, A., van Heerde, E. *et al.* Discovery and pharmacological characterization of jnj-64619178, a novel small-molecule inhibitor of prmt5 with potent antitumor activity. *Molecular cancer therapeutics* **20**, 2317–2328 (2021).

[49] Klemm, T., Ebert, G., Calleja, D. J., Allison, C. C., Richardson, L. W., Bernardini, J. P., Lu, B. G., Kuchel, N. W., Grohmann, C., Shibata, Y. *et al.* Mechanism and inhibition of the papain-like protease, plpro, of sars-cov-2. *The EMBO journal* **39**, e106275 (2020).

[50] Amin, S. A., Banerjee, S., Ghosh, K., Gayen, S. & Jha, T. Protease targeted covid-19 drug discovery and its challenges: Insight into viral main protease (mpro) and papain-like protease (plpro) inhibitors. *Bioorganic & medicinal chemistry* **29**, 115860 (2021).

[51] Rut, W., Lv, Z., Zmudzinski, M., Patchett, S., Nayak, D., Snipas, S. J., El Oualid, F., Huang, T. T., Bekes, M., Drag, M. *et al.* Activity profiling and structures of inhibitor-bound sars-cov-2-plpro protease provides a framework for anti-covid-19 drug design. *biorxiv* 2020–04 (2020).

[52] Shen, Z., Ratia, K., Cooper, L., Kong, D., Lee, H., Kwon, Y., Li, Y., Alqarni, S., Huang, F., Dubrovskyi, O. *et al.* Design of sars-cov-2 plpro inhibitors for covid-19 antiviral therapy leveraging binding cooperativity. *Journal of medicinal chemistry* **65**, 2940–2955 (2021).

[53] Yan, S. & Wu, G. Spatial and temporal roles of sars-cov plpro—a snapshot. *The FASEB Journal* **35**, e21197 (2021).

[54] McClain, C. B. & Vabret, N. Sars-cov-2: the many pros of targeting plpro. *Signal Transduction and Targeted Therapy* **5**, 223 (2020).

[55] Puhl, A. C., Godoy, A. S., Noske, G. D., Nakamura, A. M., Gawriljuk, V. O., Fernandes, R. S., Oliva, G. & Ekins, S. Discovery of plpro and mpro inhibitors for sars-cov-2. *ACS omega* **8**, 22603–22612 (2023).

[56] Kouznetsova, V. L., Zhang, A., Tatineni, M., Miller, M. A. & Tsigelny, I. F. Potential covid-19 papain-like protease plpro inhibitors: repurposing fda-approved drugs. *PeerJ* **8**, e9965 (2020).

[57] Ershov, P., Yablokov, E., Mezentsev, Y. V., Chuev, G., Fedotova, M., Kruchinin, S. & Ivanov, A. Sars-cov-2 coronavirus papain-like protease plpro as an antiviral target for inhibitors of active site and protein–protein interactions. *Biophysics* **67**, 902–912 (2022).

[58] Shin, D., Mukherjee, R., Grewe, D., Bojkova, D., Baek, K., Bhattacharya, A., Schulz, L., Widera, M., Mehdipour, A. R., Tascher, G. *et al.* Papain-like protease regulates sars-cov-2 viral spread and innate immunity. *Nature* **587**, 657–662 (2020).

[59] Zhou, C., Li, C., Luo, L., Li, X., Jia, K., He, N., Mao, S., Wang, W., Shao, C., Liu, X. *et al.* Anti-tumor efficacy of hrs-4642 and its potential combination with proteasome inhibition in kras g12d-mutant cancer. *Cancer Cell* **42**, 1286–1300 (2024).

[60] Isermann, T., Sers, C., Der, C. J. & Papke, B. Kras inhibitors: resistance drivers and combinatorial strategies. *Trends in cancer* **11**, 91–116 (2025).

[61] Ghazi Vakili, M., Gorgulla, C., Snider, J., Nigam, A., Bezrukov, D., Varoli, D., Aliper, A., Polykovsky, D., Padmanabha Das, K. M., Cox Iii, H. *et al.* Quantum-computing-enhanced algorithm unveils potential kras inhibitors. *Nature biotechnology* **43**, 1954–1959 (2025).

[62] Degen, J., Wegscheid-Gerlach, C., Zaliani, A. & Rarey, M. On the art of compiling and using'drug-like'chemical fragment spaces. *ChemMedChem* **3**, 1503 (2008).

[63] Koes, D. R., Baumgartner, M. P. & Camacho, C. J. Lessons learned in empirical scoring with smina from the csar 2011 benchmarking exercise. *Journal of chemical information and modeling* **53**, 1893–1904 (2013).

[64] Gao, W., Fu, T., Sun, J. & Coley, C. Sample efficiency matters: a benchmark for practical molecular optimization. *Advances in neural information processing systems* **35**, 21342–21357 (2022).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementGenerativeexplorationoffragmentbasedchemicalspacevialargelanguagemodelsenablesthediscoveryofpotentleadsfortargetslackingbioactivel