LARGE DRUG DISCOVERY MODEL

Ilia Igashov[1*], **Arne Schneuing**[1*], **Adrian W. Dobbelstein**[1*], **Irina Morozova**[1], **Rebecca M. Neeser**[1], **Evgenia Elizarova**[1], **Philippe Schwaller**[1], **Michael M. Bronstein**[2,3] **& Bruno Correia**[1]
[1]École Polytechnique Fédérale de Lausanne, [2]University of Oxford, [3]Aithyra
{ilia.igashov,arne.schneuing,adrian.dobbelstein,bruno.correia}@epfl.ch

ABSTRACT

We introduce the Large Drug Discovery Model (LDDM), a generative framework for target-specific 3D molecule design, leveraging fragment-based masked modeling and large-scale training on the new synthetic dataset of protein-ligand complexes SynthDock. In addition to *de novo* drug design, LDDM is also able to solve more constrained drug discovery tasks, which allows users to interact with the model during the design process. We furthermore leverage this feature to introduce a new controlled generation strategy for multi-objective optimization. We benchmarked LDDM across various tasks, including *de novo* generation, fragment-based drug design, and molecular docking. Finally, we experimentally validated LDDM-designed molecules that bind to the oncogenic target KRAS.

1 INTRODUCTION

Computational drug discovery has long relied on virtual screening (Walters & Wang, 2020), but its effectiveness is limited by available compound libraries, making it less suitable for challenging and highly constrained targets like protein-protein interactions. Besides, even the largest screening libraries (Walters & Wang, 2020) cover only a tiny fraction of the entire drug-like chemical space (Bohacek et al., 1996).

In this work, we follow another computational drug discovery strategy: generative modeling. Contrary to virtual screening, generative models are not limited by available compound libraries and can perform custom molecular design taking into account geometric and chemo-physical properties of the target pocket. We introduce the Large Drug Discovery Model (LDDM), a generative framework that simultaneously operates on atom types, coordinates, and covalent bonds of small molecules, and samples all these modalities in the context of protein pockets. Unlike previous structure-based drug design methods, LDDM is not only capable of generating new ligands from scratch but also supports various substructure design and docking tasks. To achieve this, we introduce a fragment-based masked modeling approach, in which portions of the input molecule are masked, and the model is trained to recover these missing parts. This technique, inspired by language modeling methods (Devlin et al., 2018), enables efficient self-supervised learning by modeling the probabilistic structure of the chemical space. To define meaningful chemical units, we fragment the molecule using BRICS (Degen et al., 2008) (Figure 1A). During training, we randomly mask different subsets of fragments and task the model with predicting the missing parts. We perform masking not only across molecular fragments, but also across data modalities. We randomly mask either the entire fragment (design regime) or only its coordinates (docking regime), as shown in Figure 1B. To recover the masked information, we follow Schneuing et al. (2025) and train a multi-domain generative model that combines equivariant flow matching (Lipman et al., 2023) for sampling three-dimensional (3D) atom coordinates and Markov bridge models (Igashov et al., 2024a) for discrete atom and bond type generation, as shown in Figure 1C.

Furthermore, fragment-based masked modeling enables an efficient optimization strategy for decomposable objectives (Figure 5D). Starting with an empty protein pocket, we generate an initial set of molecules, split them into fragments, and retain only those meeting all specified criteria. These fragments are used as inputs for the next generation cycle, where the model fills in the gaps with new chemical matter. This iterative process refines molecules over time, optimizing local properties

[*]These authors contributed equally

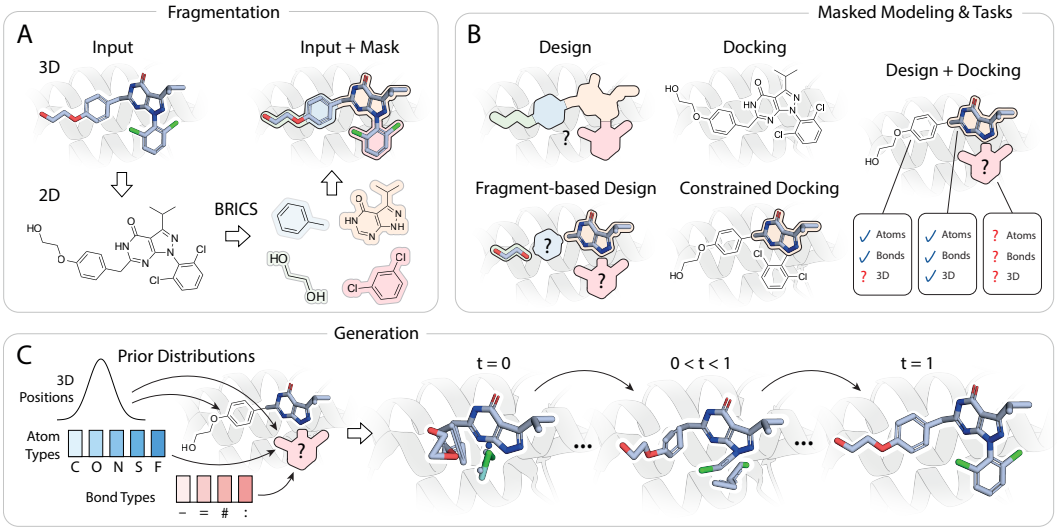

Figure 1: **Method overview.** LDDM operates on molecules that are fragmented using BRICS (Degen et al., 2008) (A), tackles diverse tasks depending on the level of the available information (B), and is trained to reconstruct missing data through generative denoising (C).

such as protein-ligand interactions. The same approach also enables us to sample new molecules while respecting synthesis constraints (see Section 2.4).

We benchmark LDDM across various design and docking tasks and showcase the design of a novel KRAS binder, for which we experimentally validated binding to the targeted pocket.

2 Results

2.1 Molecular Design

In silico evaluation of computational drug discovery methods remains an open and extremely challenging problem. Without a definitive ground truth and with the known limitations of existing computational metrics (Plewczynski et al., 2011; Tian et al., 2015; Ertl & Schuffenhauer, 2009), it is unclear how to reliably validate drug design methods on the computer. Here, we assess *de novo* design capabilities of our model from two perspectives: its ability to recover geometric and pharmacophoric patterns of known binders and to model the underlying chemical space.

First, using experimental data from the PoseBusters test set (Buttenschoen et al., 2024), we explore how well our model recovers known binding signals. To this end, we compare LDDM designs to known binders using a shape and color similarity metric SC_{RDKit} (Putta et al., 2005; Landrum et al., 2006). The color similarity function scores two 3D conformers against each other based on the overlap of their pharmacophoric features, while the shape similarity measure is a simple volumetric comparison between the two conformers (Imrie et al., 2020). We benchmark LDDM against alternative drug discovery approaches, including docking-based virtual screening and the reinforcement learning-based framework REINVENT (Blaschke et al., 2020). As positive and negative controls, we report shape-color similarities of re-docked reference ligands and randomly selected molecules, respectively. All docking experiments are conducted using variants of AutoDock Vina (Ding et al., 2023). Further details on the test set, baseline configurations, and evaluation metrics are provided in Appendix A.3-A.4. As shown in Figures 2A and 6A, LDDM demonstrates a considerably higher ability to recover the reference binding signals. Figures 2C and 6B provide examples of generated molecules that achieve high shape-color similarities with the reference molecules. In all cases, the designed and ground-truth molecules have similar coarse-grained topologies and high volumetric overlap. Besides, heteroatoms with similar properties are often placed in the same pocket regions.

Following Schneuing et al. (2025), we endow LDDM with an uncertainty estimation mechanism that allows to detect potentially flawed samples. As shown in Figure 2B, the model tends to have

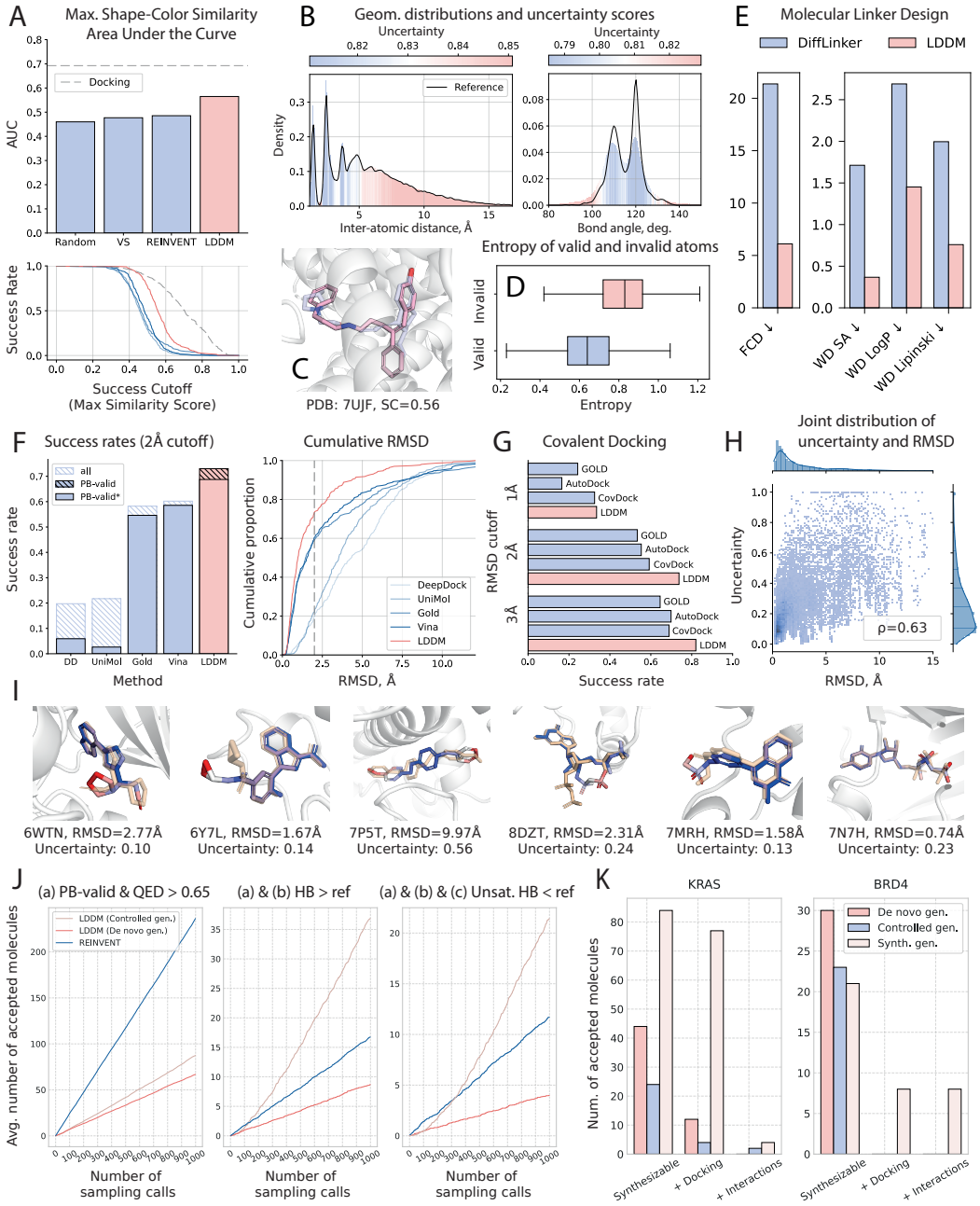

Figure 2: **Benchmarks.** Recovery of binding patterns in *de novo* design (A, C), uncertainty estimation and geometric distributions (B), entropy and chemical validity (D), linker design results (E), local and covalent docking (F-I), controlled generation efficiency (J), and synthesizable design (K).

higher uncertainty scores when it generates less likely geometries, highlighting the tails of interatomic distance and bond angle distributions. We also observed that LDDM assigns on average lower entropy values to chemically valid molecular substructures, compared to the invalid ones, as we demonstrate in Figure 2D.

More results on the ability of LDDM to model the underlying chemical space, as well as characteristics of the space itself are provided in Appendix B.1 and Figure 6.

2.2 FRAGMENT-BASED DESIGN

By design, LDDM can be conditioned on known molecular parts, making it well-suited for various fragment-based design scenarios. Here, we focus on molecular linker design and compare our approach to the specialized method DiffLinker (Igashov et al., 2024b). As shown in Figure 2E, LDDM exhibits superior distribution learning capabilities in terms of Fréchet ChemNet Distance (FCD) and Wasserstein distance (WD) between the sampled and reference molecules across various molecular properties. Here, both models were trained on the same Pockets dataset (Igashov et al., 2024b). More details are provided in Appendix A.2.

2.3 MOLECULAR DOCKING

By constraining atom and bond type modalities during generation, LDDM effectively addresses local and constrained docking tasks, as illustrated in Figure 1B. Here, we evaluate LDDM's performance in both settings and compare it against various state-of-the-art baselines (see Appendix A.3 for details).

First, we explore local docking capabilities of our model. As shown in Figure 2F, LDDM outperforms other docking methods on the PoseBusters test set (Buttenschoen et al., 2024), achieving docking root mean square deviation (RMSD) below 2Å in over 70% of test cases. For each target, we sampled 50 conformations and selected only one based on our model's uncertainty score. LDDM's uncertainty score correlates well with docking RMSD (Spearman's $\rho = 0.63$), enabling reliable scoring of docking poses (Figure 2H). Notably, LDDM can apply uncertainty estimation at the atomic level, identifying specific molecular regions that may have been docked imprecisely. Figure 2I showcases selected examples where atomic uncertainties strongly align with RMSD values, highlighting incorrectly docked molecular fragments.

Next, we evaluate LDDM's ability to perform covalent docking by additionally constraining the position of the ligand atom covalently attached to the target protein. We test our model on the benchmark set introduced by Scarpino et al. (2018), and compare its performance to other state-of-the-art covalent docking methods. As shown in Figure 2G, LDDM outperforms competing methods substantially in top-1 success rate for 2Å and 3Å RMSD cutoffs. As in the previous benchmark, we use LDDM's uncertainty estimates to score the docking conformations.

2.4 OPTIMISATION VIA CONTROLLED GENERATION

Controlled generation allows to design molecules with improved local properties that are rarely observed in *de novo*-generated compounds. This is achieved through an iterative process, illustrated in Figure 5D and discussed in detail in Appendix A.1.

We conduct three experiments, gradually increasing the number of local and global validation filters applied. Successful designs are defined as those that are (a) valid according to the PoseBusters criteria (Buttenschoen et al., 2024) and have quantitative drug-likeness (QED) (Bickerton et al., 2012) above 0.65, (b) have more hydrogen bonds than the reference molecule, and (c) have less unsatisfied hydrogen bond donors and acceptors than the reference molecule. We compare our method to REINVENT (Blaschke et al., 2020), a reinforcement learning framework that is tasked to optimize QED and the number of hydrogen bonds between the designed molecule and the target protein, as explained in Appendix A.3.

Figure 2J compares the average cumulative number of unique valid molecules per target for REINVENT, *de novo* design with LDDM, and controlled generation on 50 randomly chosen PoseBusters targets. While REINVENT efficiently optimizes global properties such as QED, our controlled generation procedure outperforms it in local molecular optimization and produces more compounds with higher numbers of hydrogen bonds and lower numbers of unsatisfied hydrogen bond donors and acceptors. This highlights the unique advantages of our approach over structure-implicit optimization methods in enhancing target-dependent and decomposable properties. Additionally, controlled generation dramatically increases the yield compared to simple *de novo* generation.

Using the same iterative design strategy with a different fragment tree expansion policy, we can also improve the synthetic accessibility of generated molecules. To this end, we expand our fragment tree only if a designed molecule contains substructures that are reachable via a single step in our

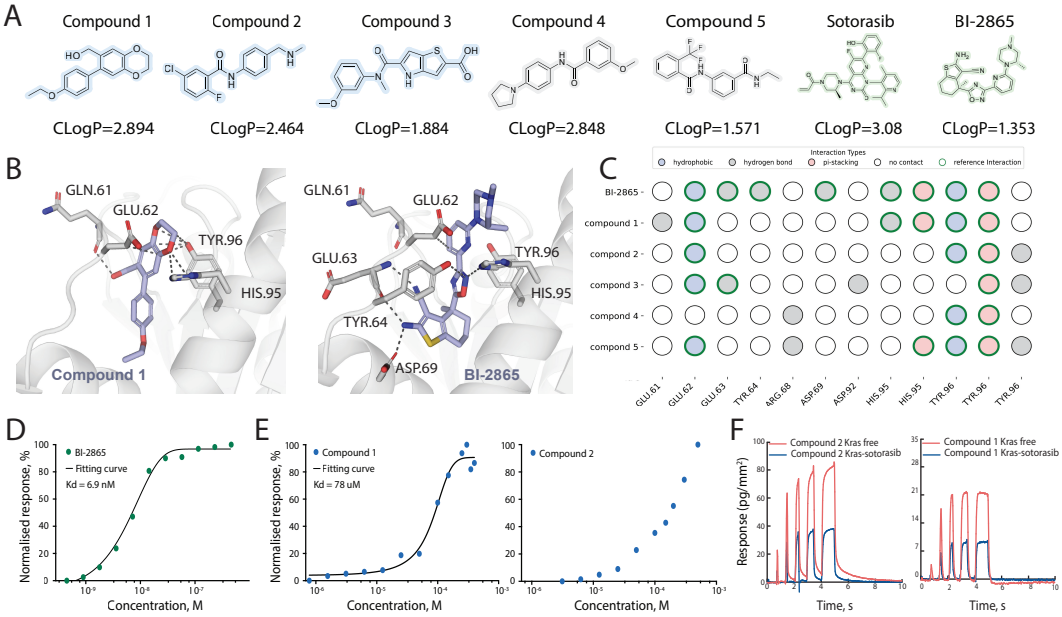

Figure 3: **KRAS binder design.** Chemical structures of 5 compounds tested experimentally, sotorasib, and BI-2865 (A), binding poses of the designed compound 1 and reference BI-2865 (B), interaction profiles of LDDM designs compared to the reference BI-2865 (PDB ID: 8AZR) (C), binding kinetics for positive control BI-2865 (D), binding kinetics for hit compounds 1 and 2 (E), and rapid kinetic sensograms for compounds 1 and 2 in the blocking experiment with sotorasib (F).

synthesis framework, which includes around $140\,000$ building blocks and 13 reaction rules following SyntheMol (Swanson et al., 2024). For each fragment in the tree, in turn, we obtain the set of all reachable molecules by enumerating possible derivatives from the current fragment using our synthesis framework. Therefore, every generated molecule is already associated with a robust synthesis route. In Figure 2K, we sampled 1000 molecules each for two target proteins, KRAS and BRD4. We then applied increasingly strict filters based on (a) AiZynthFinder synthetic accessibility (Genheden et al., 2020), (b) Vina docking scores (Alhossary et al., 2015), and (c) hydrogen-bond interactions (satisfied vs. unsatisfied). Synthesizable generation yields molecules that meet all these criteria, while generation without synthesis constraints produces substantially fewer accepted molecules.

2.4.1 Design of KRAS Binders

In this study, we explore the capability of LDDM to design small molecules targeting KRAS (Kessler et al., 2019). The Switch II region was selected as the target pocket, the same site known to be bound by the pan-KRAS inhibitor BI-2865 (Kim et al., 2023). After sampling and computational scoring (see Appendix A.5), 2263 designed molecules passed all filters and were searched against the Enamine REAL database (Shivanyuk et al., 2007) for exact matches. The top five molecules were selected on the basis of their interaction profiles and ordered for experimental testing. The selected candidates and their interaction profiles are presented in Figures 3A and 3C, respectively. Figure 3B shows the generated 3D structure of compound 1 together with its key interaction in comparison to positive control BI-2865. All five compounds were further tested for binding using grating-coupled interferometry (GCI). The screening revealed 3 successful binders according to our hit calling criteria (Appendix A.5) albeit with substantially lower affinity (high micromolar to millimolar, Figure 3E) than the positive control ($6.9\,\mathrm{nM}$). To confirm whether the designed ligands bind at the targeted site, we then performed a blocking experiment, in which $\mathrm{KRAS}^{\mathrm{G12C}}$ was preincubated with the covalent inhibitor sotorasib. This resulted in reduced signal (Figure 3F) suggesting that the hits bind in the target pocket as predicted.

3 CONCLUSION

We introduced Large Drug Discovery Model (LDDM), a 3D generative model inspired by masked pre-training in large language models. By selectively masking molecular fragments and modulating input modalities, LDDM seamlessly adapts to constrained and unconstrained docking and design tasks, achieving competitive performance across various benchmarks. Leveraging controlled and synthesizable generation, our approach efficiently optimizes decomposable, structure-dependent properties while ensuring synthetic feasibility. Experimental validation confirmed binding to the oncogenic target KRAS for three out of five computationally designed compounds. While affinity remains an open challenge, integrating stricter filtering and human-in-the-loop selection could enhance specificity and potency of the designed compounds. Overall, LDDM serves as a versatile tool to accelerate hit identification and lead optimization, streamlining the drug discovery process.

REFERENCES

Amr Alhossary, Stephanus Daniel Handoko, Yuguang Mu, and Chee-Keong Kwoh. Fast, accurate, and reliable molecular docking with quickvina 2. *Bioinformatics*, 31(13):2214–2216, 2015.

Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision*, pp. 348–367. Springer, 2022.

A Patrícia Bento, Anne Hersey, Eloy Félix, Greg Landrum, Anna Gaulton, Francis Atkinson, Louisa J Bellis, Marleen De Veij, and Andrew R Leach. An open source chemical structure curation pipeline using rdkit. *Journal of Cheminformatics*, 12:1–16, 2020.

G Richard Bickerton, Gaia V Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L Hopkins. Quantifying the chemical beauty of drugs. *Nature chemistry*, 4(2):90–98, 2012.

Thomas Blaschke, Josep Arús-Pous, Hongming Chen, Christian Margreitter, Christian Tyrchan, Ola Engkvist, Kostas Papadopoulos, and Atanas Patronov. Reinvent 2.0: an ai tool for de novo drug design. *Journal of chemical information and modeling*, 60(12):5918–5922, 2020.

Regine S Bohacek, Colin McMartin, and Wayne C Guida. The art and practice of structure-based drug design: a molecular modeling perspective. *Medicinal research reviews*, 16(1):3–50, 1996.

Cédric Bouysset and Sébastien Fiorucci. Prolif: a library to encode molecular interactions as fingerprints. *Journal of cheminformatics*, 13(1):72, 2021.

Martin Buttenschoen, Garrett M Morris, and Charlotte M Deane. Posebusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15(9):3130–3139, 2024.

Jorg Degen, Christof Wegscheid-Gerlach, Andrea Zaliani, and Matthias Rarey. On the art of compiling and using'drug-like'chemical fragment spaces. *ChemMedChem*, 3(10):1503, 2008.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Ji Ding, Shidi Tang, Zheming Mei, Lingyue Wang, Qinqin Huang, Haifeng Hu, Ming Ling, and Jiansheng Wu. Vina-gpu 2.0: further accelerating autodock vina and its derivatives with graphics processing units. *Journal of chemical information and modeling*, 63(7):1982–1998, 2023.

Janani Durairaj, Yusuf Adeshina, Zhonglin Cao, Xuejin Zhang, Vladas Oleinikovas, Thomas Duignan, Zachary McClure, Xavier Robin, Daniel Kovtun, Emanuele Rossi, et al. Plinder: The protein-ligand interactions dataset and evaluation resource. *bioRxiv*, pp. 2024–07, 2024.

Peter Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of cheminformatics*, 1:1–11, 2009.

Wei P Feinstein and Michal Brylinski. Calculating an optimal box size for ligand docking and virtual screening against experimental and predicted binding pockets. *Journal of cheminformatics*, 7:1–10, 2015.

Paul G Francoeur, Tomohide Masuda, Jocelyn Sunseri, Andrew Jia, Richard B Iovanisci, Ian Snyder, and David R Koes. Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. *Journal of chemical information and modeling*, 60(9):4200–4215, 2020.

Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107, 2012.

Samuel Genheden, Amol Thakkar, Veronika Chadimová, Jean-Louis Reymond, Ola Engkvist, and Esben Bjerrum. Aizynthfinder: a fast, robust and flexible open-source software for retrosynthetic planning. *Journal of cheminformatics*, 12(1):70, 2020.

Golnaz Ghiasi, Barret Zoph, Ekin D Cubuk, Quoc V Le, and Tsung-Yi Lin. Multi-task self-training for learning general representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8856–8865, 2021.

Ilia Igashov, Arne Schneuing, Marwin Segler, Michael M. Bronstein, and Bruno Correia. Retrobridge: Modeling retrosynthesis with markov bridges. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=770DetV8He.

Ilia Igashov, Hannes Stärk, Clément Vignac, Arne Schneuing, Victor Garcia Satorras, Pascal Frossard, Max Welling, Michael Bronstein, and Bruno Correia. Equivariant 3d-conditional diffusion model for molecular linker design. *Nature Machine Intelligence*, pp. 1–11, 2024b.

Fergus Imrie, Anthony R Bradley, Mihaela van der Schaar, and Charlotte M Deane. Deep generative models for 3d linker design. *Journal of chemical information and modeling*, 60(4):1983–1995, 2020.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Dirk Kessler, Michael Gmachl, Andreas Mantoulidis, Laetitia J Martin, Andreas Zoephel, Moriz Mayer, Andreas Gollner, David Covini, Silke Fischer, Thomas Gerstberger, et al. Drugging an undruggable pocket on kras. *Proceedings of the National Academy of Sciences*, 116(32):15823–15829, 2019.

Dongsung Kim, Lorenz Herdeis, Dorothea Rudolph, Yulei Zhao, Jark Böttcher, Alberto Vides, Carlos I Ayala-Santos, Yasin Pourfarjam, Antonio Cuevas-Navarro, Jenny Y Xue, et al. Pan-kras inhibitor disables oncogenic signalling and tumour growth. *Nature*, 619(7968):160–166, 2023.

Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *European conference on machine learning*, pp. 282–293. Springer, 2006.

Radoslav Krivák and David Hoksza. P2rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *Journal of cheminformatics*, 10:1–12, 2018.

Gregory A Landrum and Sereina Riniker. Combining ic50 or k i values from different sources is a source of significant noise. *Journal of Chemical Information and Modeling*, 64(5):1560–1567, 2024.

Gregory A Landrum, Julie E Penzotti, and Santosh Putta. Feature-map vectors: a new class of informative descriptors for computational drug discovery. *Journal of computer-aided molecular design*, 20:751–762, 2006.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Matthias Rarey Louis Bellmann, Patrick Penner. Topological similarity search in large combinatorial fragment spaces. *Journal of Chemical Information and Modeling*, 2020.

Shitong Luo, Jiaqi Guan, Jianzhu Ma, and Jian Peng. A 3d generative model for structure-based drug design. *Advances in Neural Information Processing Systems*, 34:6229–6239, 2021.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Andrew T McNutt, Paul Francoeur, Rishal Aggarwal, Tomohide Masuda, Rocco Meli, Matthew Ragoza, Jocelyn Sunseri, and David Ryan Koes. Gnina 1.0: molecular docking with deep learning. *Journal of cheminformatics*, 13(1):43, 2021.

David Mizrahi, Roman Bachmann, Oguzhan Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4m: Massively multimodal masked modeling. *Advances in Neural Information Processing Systems*, 36:58363–58408, 2023.

Dariusz Plewczynski, Michał Łaźniewski, Rafał Augustyniak, and Krzysztof Ginalski. Can we trust docking results? evaluation of seven commonly used programs on pdbbind database. *Journal of computational chemistry*, 32(4):742–755, 2011.

Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Günter Klambauer. Fréchet chemnet distance: A metric for generative models for molecules in drug discovery. *arXiv preprint arXiv:1803.09518*, 2018.

Santosh Putta, Gregory A Landrum, and Julie E Penzotti. Conformation mining: an algorithm for finding biologically relevant conformations. *Journal of medicinal chemistry*, 48(9):3313–3318, 2005.

Andrea Scarpino, György G Ferenczy, and György M Keseru. Comparative evaluation of covalent docking tools. *Journal of Chemical Information and Modeling*, 58(7):1441–1458, 2018.

Arne Schneuing, Charles Harris, Yuanqi Du, Kieran Didi, Arian Jamasb, Ilia Igashov, Weitao Du, Carla Gomes, Tom L Blundell, Pietro Lio, et al. Structure-based drug design with equivariant diffusion models. *Nature Computational Science*, 4(12):899–909, 2024.

Arne Schneuing, Ilia Igashov, Adrian W. Dobbelstein, Thomas Castiglione, Michael M. Bronstein, and Bruno Correia. Multi-domain distribution learning for de novo drug design. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=g3VCIM94ke.

Alexander N Shivanyuk, Sergey V Ryabukhin, A Tolmachev, AV Bogolyubsky, DM Mykytenko, AA Chupryna, W Heilman, and AN Kostyuk. Enamine real database: Making chemical diversity real. *Chemistry today*, 25(6):58–59, 2007.

Baris E Suzek, Hongzhan Huang, Peter McGarvey, Raja Mazumder, and Cathy H Wu. Uniref: comprehensive and non-redundant uniprot reference clusters. *Bioinformatics*, 23(10):1282–1288, 2007.

Kyle Swanson, Gary Liu, Denise B Catacutan, Autumn Arnold, James Zou, and Jonathan M Stokes. Generative ai for designing and validating easily synthesizable and structurally novel antibiotics. *Nature Machine Intelligence*, 6(3):338–353, 2024.

Sheng Tian, Junmei Wang, Youyong Li, Dan Li, Lei Xu, and Tingjun Hou. The application of in silico drug-likeness predictions in pharmaceutical research. *Advanced drug delivery reviews*, 86: 2–10, 2015.

Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, 50(D1):D439–D444, 2022.

Pat Walters. Useful rdkit utils, 2021. URL https://github.com/PatWalters/useful_rdkit_utils.

Pat Walters. Mining ring systems in molecules for fun and profit, 2022. URL https://practicalcheminformatics.blogspot.com/2022/12/identifying-ring-systems-in-molecules.html.

W Patrick Walters and Renxiao Wang. New trends in virtual screening, 2020.

W Patrick Walters, Matthew T Stahl, and Mark A Murcko. Virtual screening—an overview. *Drug discovery today*, 3(4):160–178, 1998.

Renxiao Wang, Xueliang Fang, Yipin Lu, Chao-Yie Yang, and Shaomeng Wang. The pdbbind database: methodologies and updates. *Journal of medicinal chemistry*, 48(12):4111–4119, 2005.

Scott A Wildman and Gordon M Crippen. Prediction of physicochemical parameters by atomic contributions. *Journal of chemical information and computer sciences*, 39(5):868–873, 1999.

A METHODS

A.1 LDDM

Architecture We use the geometric heterogeneous graph neural network architecture from DrugFlow (Schneuing et al., 2025), including self-conditioning, confidence head, and virtual nodes. Furthermore, we modify the input layer to accept three additional binary flags indicating whether a bond type, atom type or atom coordinate is known. Small molecule atoms are featurized as one-hot vectors representing 13 different types, {C, N, O, S, B, Br, Cl, P, I, F, NH, N+, O-} where charges and explicit hydrogens are included in selected cases following (Schneuing et al., 2025). Protein residue nodes carry a feature encoding their type (one of the 20 standard amino acids) as well as vector-valued features indicating the positions of all atoms belonging to that residue relative to its C_α position. Bond types are either single, double, triple, aromatic, or "None". We remove all edges between pocket residues or between residues and ligand atoms that are more than $10\,\text{Å}$ apart for computational efficiency.

Masked modeling LDDM is trained as a generative model with varying degrees of conditioning. Randomly sampled parts of the training molecules are masked out and the model is tasked to recover the missing bits similar to masked language modeling. Before training, all molecules are fragmented into retrosynthetically interesting substructures using the BRICS (Degen et al., 2008) method. At each training iteration, we then select a role for each of the resulting fragments before noise is added. These roles are sampled with equal probability. Fragments can be designed, docked, or provided purely as context. If selected for design, both the molecular graph and the coordinates are noised and subsequently predicted by the model. For docking, coordinates are noised but the model has access to the clean molecular graph of the fragment. Context fragments are not noised at all, and the model can use their full geometric and chemical information to infer the structures of other parts of the ligand. To predict missing parts of the molecular system, we employ a generative modelling framework that integrates continuous flow matching (Lipman et al., 2023) for atom coordinates with Markov bridge models (Igashov et al., 2024a) for categorical atom and bond types.

Controlled generation Building on LDDM's substructure-constrained design ability, the controlled generation algorithm iteratively proposes candidate molecules, analyses their properties, and selects substructures that fulfill predefined criteria to condition new samples on. At each iteration, two filtering functions are applied:

- *Global Filtering* (`FilterMol`) discards invalid or undesired molecules. For demonstrating sampling efficiency in Section 2.4, we exclude molecules with QED below 0.6.
- *Local Filtering* (`FilterFrag`) evaluates each fragment individually. Here, we discard fragments that are not PB-valid or do not form a hydrogen bond with the target (as assessed with ProLIF (Bouysset & Fiorucci, 2021)). For the benchmark shown in Figure 2J, we additionally require that fragments have no unsatisfied hydrogen bonds.

Each valid fragment is stored as a node in a *fragment tree* $\mathcal{T}$, where each parent node is a substructure of its children. We manage the sampling budget B at each iteration using a scheme inspired by the Upper Confidence Bound (UCB) for Monte Carlo tree search (Kocsis & Szepesvári, 2006):

$$\text{UCB}(v) = \frac{\text{successes}(v)}{\text{calls}(v)+1} \;+\; c\sqrt{\frac{\ln\big(\text{pCalls}(v)+1\big)}{\text{calls}(v)+1}},$$

where $\text{successes}(v)$ and $\text{calls}(v)$ track the number of accepted molecules derived from node v and the total sampling attempts at v, respectively. $\text{pCalls}(v)$ stands for the total number of samples generated from the parent of the considered node. The hyperparameter c controls the exploration-exploitation trade-off and was set to 1 in our experiments. The number of samples allocated to each node is proportional to its UCB score, balancing exploration of under-sampled fragments with exploitation of promising ones. Full algorithmic details are provided in Algorithm 1.

Synthesizable generation To generate molecules that are synthetically feasible, we replace BRICS fragmentation with a *building block* approach. Each fragment node v in the tree is linked

Algorithm 1 Controlled Generation with LDDM

Require: $\mathcal{T}$ (fragment tree, empty at start), `FilterMol` (global filter), `FilterFrag` (local filter), `Alloc` (allocation function), $N_{\max}$ (max iterations)

for $i = 1$ to $N_{\max}$ **do**
  $B \leftarrow \texttt{Alloc}(\mathcal{T})$ ▷ Allocate sampling budget to tree nodes
  $M \leftarrow \texttt{SampleMolecules}(B)$ ▷ Generate candidate molecules
  $M_{\text{valid}} \leftarrow \{\, m \in M : \texttt{FilterMol}(m) = 1 \}$ ▷ Global filter
  for all $m \in M_{\text{valid}}$ **do**
    $F \leftarrow \texttt{FragmentBRICS}(m)$
    $F_{\text{sel}} \leftarrow \{\, x \in F : \texttt{FilterFrag}(x) = 1 \}$ ▷ Local filter
    $\mathcal{T} \leftarrow \texttt{AddFragments}(\mathcal{T}, F_{\text{sel}})$
  end for
end for
return $\texttt{ExtractMolecules}(\mathcal{T})$

to a library of reachable reaction products, precomputed from reaction templates $\mathcal{R}$ and building blocks $\mathcal{B}$ that are compatible with the current molecule. Newly generated molecules must contain a substructure match with one of these precomputed products to create a new child node that contains the matching substructure. This ensures that expansion pathways align with tentative synthetic routes.

To enable sampling tree expansion also for similar reaction products (i.e. partial substructure matches), we additionally employ a hybrid design/docking strategy. The maximum common substructure is kept fixed, while unmatched atoms of the reaction product are docked flexibly.

The same `FilterMol` and `FilterFrag` checks apply to all newly formed fragments. By combining fragment-based masked modeling with reaction-driven expansions, synthesizable generation systematically explores combinatorial fragment spaces while optimizing for local properties (see Algorithm 2).

We perform 1024 sampling calls over 20 iterations (64 molecules per batch). We use the ENAMINE building block database and combine them with suitable reaction templates from Synthemol (Swanson et al., 2024). Generated molecules are subjected to three sequential filters:

1. *Synthetic Feasibility.* Assessed with AiZynthFinder (Genheden et al., 2020), requiring each molecule to be synthesizable in no more than two steps.
2. *Docking Success.* Molecules must achieve a Vina score below -8 kcal/mol *or* surpass the reference ligand's score, and must have a minimized RMSD ≤ 2 Å.
3. *Interaction Success.* Hydrogen-bond counts are compared to those of the reference ligand; valid molecules must have at least as many hydrogen bonds as the reference count and not more unsatisfied sites then the reference.

A.2 Datasets

SynthDock training set In this project, we aim to achieve two key goals: to explore a broad chemical space of drug-like small molecules, and to learn physics of protein-ligand interactions. While training on millions of experimentally determined complexes of proteins and drug-like molecules would be the preferred option, such data is limited. Therefore, we set out to use protein structure prediction (Jumper et al., 2021) and molecular docking algorithms (McNutt et al., 2021), and generated a large, high-quality synthetic protein-ligand dataset, SynthDock. Our dataset includes 3491 proteins grouped in 2704 UniRef50 clusters (Suzek et al., 2007), and $264\,449$ unique drug-like compounds from ChEMBL Gaulton et al. (2012). The main advantage of this synthetic data approach lies in the scale and diversity of the dataset, which allows us to capture a large chemical space, as illustrated in Figure 6C. Additionally, similar to the effects reported in the computer vision domain (Ghiasi et al., 2021; Bachmann et al., 2022; Mizrahi et al., 2023), we anticipate that large-scale probabilistic training on high-quality pseudo labeled data enables the transfer of inductive biases from the *teacher* docking and structure prediction methods to LDDM, ensuring that our model learns the underlying physics of protein-ligand interactions.

Algorithm 2 Synthesizable Generation

Require: $\mathcal{T}$ (fragment tree, empty at start), `FilterMol` (global filter), `FilterFrag` (local filter), `Alloc` (allocation function), $N_{\max}$ (max iterations), $\mathcal{R}$ (reaction templates), $\mathcal{B}$ (building blocks)

for $i = 1$ to $N_{\max}$ **do**
  $B \leftarrow \texttt{Alloc}(\mathcal{T})$ ▷ Budget allocation
  $M \leftarrow \texttt{SampleMolecules}(B)$ ▷ Candidate molecules
  $M_{\text{valid}} \leftarrow \{\, m \in M : \texttt{FilterMol}(m) = 1 \}$ ▷ Global filter
  for all $m \in M_{\text{valid}}$ **do** ▷ Iterate valid molecules
    $v \leftarrow \texttt{SampledNode}(m)$ ▷ Node from which m was sampled
    $F \leftarrow \texttt{FragmentBuildingBlocks}(m, P_v)$ ▷ Check library P_v
    $F' \leftarrow \texttt{DockSimilarProducts}(m, P_v)$ ▷ Optional docking
    $F_{\text{merged}} \leftarrow F \cup F'$
    $F_{\text{sel}} \leftarrow \{\, x \in F_{\text{merged}} : \texttt{FilterFrag}(x) = 1 \}$ ▷ Local filter
    $\mathcal{T} \leftarrow \texttt{AddFragments}(\mathcal{T}, F_{\text{sel}})$
  end for
  $\texttt{ComputeReactions}(\mathcal{T}, \mathcal{R}, \mathcal{B})$ ▷ Reaction step
end for
return $\texttt{ExtractMolecules}(\mathcal{T})$

To create the SynthDock dataset we obtained all molecules with available binding assays from the ChEMBL database (release 33) and removed protein mutants as suggested by Landrum & Riniker (2024). After also removing common salts, metals and solvents, the structures were standardized using the ChEMBL Structure Pipeline (Bento et al., 2020) by taking the neutral (parent) compound of a salt or charged compound and extracting the non-isomeric counterpart of racemic mixtures. For each of the resulting target-compound pairs, we downloaded the corresponding protein structure model from the AlphaFold Protein Structure Database (Varadi et al., 2022). Next, we ran P2Rank (Kriv´ak & Hoksza, 2018) to identify binding pockets in these structures and performed docking using Gnina (McNutt et al., 2021) without CNN scoring option. The docking box was centered in the predicted pocket centers and its size set to be $2.857\times$ ligand radius of gyration (Feinstein & Brylinski, 2015). For each pocket, Gnina generated nine binding modes from which we kept only those with Vina docking scores ≤ 0. We also discarded poses with RMSD < 1.7 Å to other poses for the same pocket-ligand pair. As a final sanity check, we applied the PoseBusters evaluation suite (Buttenschoen et al., 2024) to the docked complexes and kept only the ligand-pocket pairs that passed all filters. The SynthDock training set includes $1\,714\,308$ pocket-ligand pairs in total.

CrossDocked training set Some models were trained on the CrossDocked dataset (Francoeur et al., 2020) which was created by docking ligands from the PDB into cognate and non-cognate receptors with structurally similar binding pockets. We use the same subset and cross-validation splits as Luo et al. (2021) with $100\,000$ structures for training and 100 for testing.

PoseBusters test set Our primary test set is the PoseBusters benchmark set (Buttenschoen et al., 2024) with 308 unique ligands and proteins from the Protein Data Bank released since 2021.

DiffLinker pockets dataset We follow the dataset curation procedure described by Igashov et al. (2024b) to benchmark LDDM against DiffLinker. Specifically, we build upon the protein–ligand dataset curated by Schneuing et al. (2024) and apply the same fragmentation scheme as introduced in DiffLinker. This results in a dataset of $185\,678$ training examples and 566 test examples. To ensure a fair comparison between LDDM and DiffLinker, we re-train LDDM on the same dataset. We then evaluate the distribution learning capabilities of both methods, using the molecules in the test set as a reference.

A.3 Baselines

REINVENT We compare our method to the reinforcement learning (RL) framework REINVENT (Blaschke et al., 2020). REINVENT employs a recurrent neural network to generate molecules as SMILES (symbolic representation) and implements goal-directed learning through RL. It thus can only be implicitly conditioned on the target protein through the RL reward. We design two scoring functions taking target structure into consideration outlined below. For both ex-

periments, we limit the number of oracle calls (successful docking and interaction profiling) to 1000 to ensure a fair and computationally feasible comparison with a batch size of 32. Additionally, we incorporate the QED score as a secondary objective, where molecules with a QED score above 0.65 receive a reward of 1, while those below this threshold receive a reward of 0. The combined reward is the product of the transformed QED and docking/interaction score. Finally, to promote structural diversity among generated molecules, we apply REINVENT's built-in diversity filter (bucket size of 25 based on identical Murcko scaffold and minimum similarity of 0.4).

To assess the capability of REINVENT to recover reference ligands we optimize for the Vina score by docking using Gnina (McNutt et al., 2021) with the binding site specified by the reference ligand (autobox parameter). We transform the raw docking scores using a reverse sigmoid function,

$$R_{\text{dock}} = \frac{1}{1 + 10^{k\left(S - \frac{S_{\text{high}} + S_{\text{low}}}{2}\right)\frac{10}{S_{\text{high}} - S_{\text{low}}}}}, \tag{1}$$

where S is the docking score, S_{low} and S_{high} define the range of scores set to -14 and -2, respectively, and $k = 0.25$ is a factor defining smoothness.

In the second experiment, we assess the ability of REINVENT to optimize molecular interactions with a target protein. Here, the reward function is designed to maximize the number of hydrogen bonds formed, as identified using ProLIF (Bouysset & Fiorucci, 2021). The reward is transformed as follows,

$$R_{\text{hbonds}} = \frac{H_{\text{sat}}}{H_{\text{tot}}}, \tag{2}$$

where H_{sat} is the number of satisfied hydrogen bonds and H_{tot} is the total number of unsatisfied and satisfied hydrogen bonds in the molecule. Unsatisfied hydrogen bonds are hydrogen bond acceptors and donors that do not fulfill their potential while satisfied ones do form an interaction with the protein. This transformation ensures that molecules where all potential hydrogen bond sites are involved in a hydrogen bond to the target receive a maximum reward of 1.

Virtual screening (VS) To simulate virtual screening and showcase the ability of generative models to propose solutions beyond those in its training data, we dock molecules from LDDM's training set into the test set proteins using a GPU-optimized version of QuickVina (Ding et al., 2023; Alhossary et al., 2015), and select the top 100 hits according to docking score. For computational tractability, we randomly select a subset of $10\,000$ SynthDock molecules as our screening library. For docking, we use target-specific bounding boxes that fully cover the associated reference ligand plus 5Å margin added to all six sides.

Random As a trivial baseline, we include a random selection of docked molecules from the screening library (rather than the top performing ones).

Local docking baselines For local docking, all baselines and the corresponding results were taken from Buttenschoen et al. (2024). We note that PB-validity* reported for all baselines was measured on top-1 selected poses only. For LDDM poses, we first filtered out invalid poses and then selected the best candidates using the uncertainty scores. That is why Figure 2F has two legends for PB-validity. We label the PB-validity evaluation suggested by Buttenschoen et al. (2024) with asterisk. In these experiments, we used LDDM trained on CrossDocked (Francoeur et al., 2020).

Covalent docking baselines For covalent docking, all baselines and the corresponding results were taken from Scarpino et al. (2018).

A.4 METRICS

We use a symmetrized version of the *shape-color similarity* score proposed in (Imrie et al., 2020). Shape similarity measures the volumetric overlap of two molecules (Putta et al., 2005) and color similarity compares pharmacophores in 3D (Landrum et al., 2006). Shape-color similarity is the average of both scores, which output values between 0 and 1. We obtain a symmetric score by using both molecules as reference and averaging the outcomes.

Fréchet ChemNet Distance (FCD) (Preuer et al., 2018) measures how closely a set of generated molecules resembles the training (reference) distribution. FCD uses the activations of the penultimate layer of a pretrained bioactivity prediction model (ChemNet) to parameterize two Gaussian distributions and finally compute the Fréchet distance between them. For the visualization of the chemical space covered by LDDM samples, we used the same ChemNet embeddings, mapped them on the first 100 principal components using PCA and computed 2D projections with UMAP (McInnes et al., 2018).

We also report *validity* as the fraction of molecules passing RDKit's sanitization filters, *connectivity* as the fraction of molecules without disconnected fragments, a *quantitative estimate of drug-likeness (QED)* (Bickerton et al., 2012), a *synthetic accessibility proxy (SA)* (Ertl & Schuffenhauer, 2009) and *lipophilicity (logP)* (Wildman & Crippen, 1999). For some of these quantities we compute Wasserstein-1 distances (WD) or Jensen-Shannon divergences (JSD) to the distributions observed in the training set rather than absolute values to evaluate if the generative model successfully recreates patters from the reference set (Schneuing et al., 2025). *PoseBusters (PB) validity* checks what fraction of molecules passes all filters in the PoseBusters test suite (Buttenschoen et al., 2024).

A.5 Design of KRAS Binders

A.5.1 Computational Design

To design KRAS binders, we used LDDM trained on the CrossDocked dataset (Francoeur et al., 2020) in *de novo* design regime. During generation, we discarded molecules that did not pass PoseBusters validity filters (Buttenschoen et al., 2024), REOS filters (Walters et al., 1998), or had at least one ring system that was not found in ChEMBL (Walters, 2022; 2021). Eventually, we obtained $10\,000$ unique samples.

A.5.2 Filtering Criteria

A total of $10\,000$ molecules were designed using LDDM. To enhance the likelihood of experimental validation, a series of filtering steps were applied. To ensure the solubility of the test compounds, the logP values of the selected molecules were restricted to a range of 1 to 3. Additionally, to reduce molecular flexibility and limit binding degrees of freedom, the number of rotatable bonds was capped at a maximum of 5. After applying these two filters, 2268 molecules remained. These molecules were then subjected to an exact-match search against the Enamine REAL database, which contains 48 billion compounds. The search was performed using the SpaceLight (Louis Bellmann, 2020) mode of the InfiniSee tool, based on Tanimoto similarity on fCSFP4 fingerprints. This process identified 207 molecules within the database. To validate the binding mode, molecular docking was conducted using Gnina (McNutt et al., 2021) with an exhaustiveness parameter of 32. A total of 133 molecules were docked with a normalized RMSD of < 0.3Å per atom. For all 133 molecules, an interaction profile was generated based on the predicted binding pose from LDDM, with a focus on recovering the known interaction fingerprint of small-molecule binders targeting the same pocket. The top five molecules, ranked according to the number and type of recovered interactions, were selected for experimental testing.

A.5.3 Experimental Methods

Grating-coupled interferometry (GCI, Creoptix) was used for the hit screening. The KRAS(GDP)G12C was immobilised by amine coupling on a carboxymethyl-$5'$-dextran (4PCH) sensor chip. The surface of a chip was activated with 1:1 mixture of $(100\,\mathrm{mM})$ N-hydroxysuccinimide and $(400\,\mathrm{mM})$ 1-ethyl-3-(3-dimethylaminopropyl)-carbodiimide. The immobilisation level of KRAS(GDP)G12C protein $(60\,\mu\mathrm{g/mL})$ ($10\,\mathrm{mM}$ sodium acetate, pH 4.5) reached the immobilisation level of $15\,000$ response units.

The hit identification was based on the kinetic parameters from the rapid kinetic, where we applied two hit calling criteria: 1) dissociation and association errors are equal or less than 100, 2) maximum response value R_{max} is greater than 15. The hit identification was done using the rapid kinetic, where all the compounds are injected at a single concentration $(1\,\mathrm{mM})$ for increasing times, the flow rate was $400\,\mu\mathrm{L/min}$, association time of 5 s and dissociation time of 20 s. Each compound was addressed through two channels: reference channel and the channel with immobilised KRAS.

First, KRAS(GDP)G12C protein was immobilized on a 4PCH chip, the activity of a protein after immobilization was checked by binding of the positive control BI-2865. The protein on a chip was active and the K_d value was 6.9 nM for BI-2865 positive control, which is close to the literature range of 4.5-12 nM (Kim et al., 2023).

The binding of the second positive control—covalent inhibitor sotorasib—was done in an independent experiment, and it was shown that cysteine 12 is active in KRAS(GDP)G12C protein after immobilization, and sotorasib binding performed well. The hit screening was done in the running buffer 20 mM Hepes, 150 mM NaCl, 1 mM DTT, $5\,\mathrm{mM}$ $MgCl_2$, 1% DMSO, 0,02% P020 using rapid kinetic method. We applied the following parameters for hit identification in the program software of Creoptix method: the threshold value for association and dissociation errors – is less than or equal to 100, and the threshold value for R_{max} is greater than 15 (R_{max} for BI-2865 was 18). Applying this filtering metrics we detected three hit molecules of five designed.

Next, we performed a blocking experiment with the sotorasib covalent inhibitor to show that our hit molecules indeed bind in the right pocket. We incubated KRAS(GDP)G12C free in the running buffer with 1% DMSO and KRAS(GDP)G12C with $500\,\mu\mathrm{M}$ of covalent binder sotorasib as well in the running buffer with 1% DMSO for 10 min, the KRAS(GDP)G12C and KRAS(GDP)G12C-sotorasib were immobilised on 2 independent channels. Hit molecules were screened again against these 2 channels (with the free KRAS and blocked KRAS) using rapid kinetic. It was found that the R_{max} value is much lower and the dissociation/association errors are much higher for blocked KRAS(GDP)G12C. Thus we concluded that our hit molecules bind indeed in the correct pocket of KRAS(GDP)G12C.

Finally, the multicycle kinetic was done for our top hits to determine the Kd values, the Kd for our top hit—compound 1— was $79\,\mu\mathrm{M}$.

B Extended Results

B.1 Chemical Space Modeling

Traditionally, the evaluation of generative design methods has focused on the absolute values of various quality metrics, such as docking scores, drug-likeness (Bickerton et al., 2012), or synthetic accessibility (Ertl & Schuffenhauer, 2009). However, as discussed in Schneuing et al. (2025), this strategy often fails to meet the primary objective of generative modeling: learning the underlying distribution of the training data.

To train LDDM, we curated SynthDock, a new dataset of docked protein-ligand complexes. It includes 2704 distinct protein clusters and $264\,449$ unique molecules from ChEMBL (Gaulton et al., 2012), which makes it orders of magnitude larger and more diverse than other existing datasets for 3D structure-based drug design (Francoeur et al., 2020; Wang et al., 2005; Durairaj et al., 2024). We visualize the entire SynthDock chemical space in Figure 6C using the pre-trained neural network ChemNet (Preuer et al., 2018) and dimensionality reduction techniques, as explained in Methods. The vast chemical space of SynthDock includes different classes of molecular compounds including amines, amides, sulphones, sugar and phosphonic acid derivatives and N-heterocyclic molecules.

Notably, LDDM samples, obtained on the diverse held-out test set, uniformly cover the entire chemical space while remaining structurally novel: over 90% of LDDM samples have less than 0.5 Tanimoto similarity to the SynthDock training set. Examples of LDDM samples belonging to the different clusters of the chemical space are shown in the corresponding panels of Figure 6C.

Finally, to quantify the ability of LDDM to learn the underlying data distribution, we perform the same evaluation as in Schneuing et al. (2025). To be able to compare our method with other baselines, we train a separate model using the CrossDocked dataset Francoeur et al. (2020). As shown in Tables 1, 2, and 3, LDDM achieves competitive performance in various distribution learning metrics.

Table 1: Fréchet ChemNet Distance and Jensen-Shannon divergence between distributions of discrete molecular data. The best result is highlighted in bold, the second best is underlined. All baselines including LDDM were trained on the same CrossDocked dataset.

Method	FCD	Atoms	Bonds	Rings
POCKET2MOL	12.703	0.081	**0.044**	<u>0.446</u>
DIFFSBDD	11.637	<u>0.050</u>	0.227	0.588
TARGETDIFF	13.766	0.076	0.240	0.632
DRUGFLOW	**4.278**	**0.043**	<u>0.060</u>	**0.391**
LDDM	<u>7.192</u>	0.084	0.143	0.477

Table 2: Wasserstein distance between marginal distributions of continuous molecular data (bond distances and angles), drug-likeness (QED), synthetic accessibility (SA), lipophilicity (logP) and numbers of rotatable bonds (RB). The last column reports the Jensen-Shannon divergence between the joint distributions of four molecular properties (QED, SA, logP and Vina efficiency score). The best result is highlighted in bold, the second best is underlined. All baselines including LDDM were trained on the same CrossDocked dataset.

	Top-3 bond distances			Top-3 bond angles			Molecular properties				
Method	C–C	C–N	C=C	C–C=C	C–C–C	C–C–O	QED	SA	logP	RB	JSD_{all}
POCKET2MOL	0.050	0.024	0.045	2.173	2.936	3.938	0.072	<u>0.576</u>	1.209	2.861	0.223
DIFFSBDD	0.041	0.039	0.042	3.632	8.166	7.756	0.065	1.570	0.774	0.928	0.274
TARGETDIFF	<u>0.017</u>	<u>0.019</u>	0.028	4.281	3.422	4.125	<u>0.050</u>	1.518	**0.489**	<u>0.354</u>	0.242
DRUGFLOW	**0.017**	**0.016**	**0.016**	**0.952**	**2.269**	<u>1.941</u>	**0.014**	**0.317**	0.665	**0.144**	**0.099**
LDDM	0.041	0.024	<u>0.016</u>	<u>1.651</u>	<u>2.290</u>	**1.938**	0.096	0.944	<u>0.529</u>	1.358	<u>0.196</u>

Table 3: Wasserstein distance between distributions of binding efficiency scores and normalized numbers of different protein-ligand interactions. The best result is highlighted in bold, the second best is underlined.

	Binding efficiency		Protein-ligand interactions				
Method	Vina	Gnina	H-bond	H-bond (acc.)	H-bond (don.)	π-stacking	Hydrophobic
POCKET2MOL	0.064	0.044	0.040	0.026	0.014	<u>0.007</u>	**0.027**
DIFFSBDD	0.086	0.043	0.047	0.030	0.017	0.011	0.044
TARGETDIFF	<u>0.034</u>	<u>0.030</u>	0.031	0.021	0.010	0.012	0.039
DRUGFLOW	**0.028**	**0.013**	<u>0.019</u>	<u>0.012</u>	<u>0.007</u>	**0.006**	<u>0.036</u>
LDDM	0.066	0.032	**0.015**	**0.008**	**0.007**	0.009	0.054

B.2 DESIGN OF KRAS BINDERS

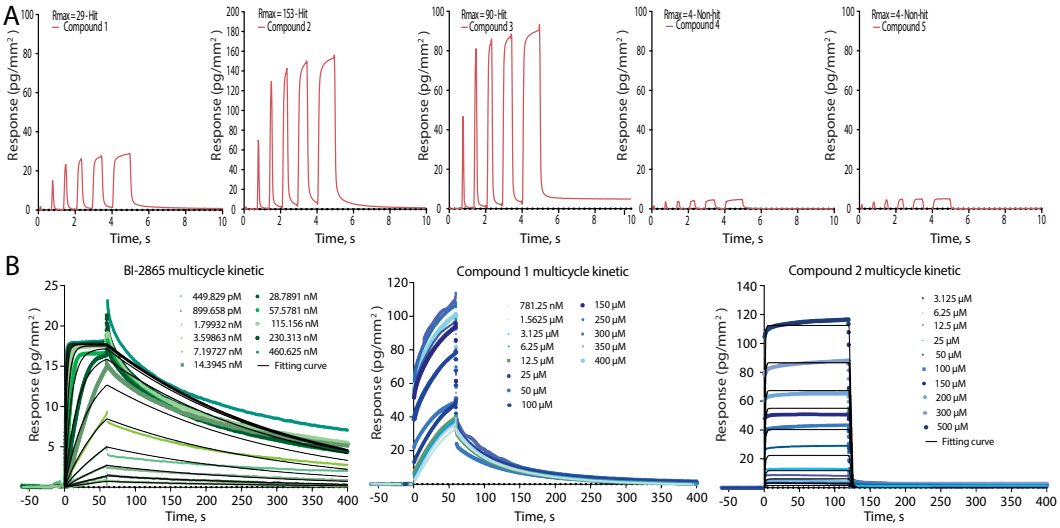

Figure 4: **Experimental Data.** (A) Rapid kinetic sensograms for all five tested compounds. (B) Binding kinetic sensograms for varying compound concentrations of the positive control BI-2865, compound 1 and compound 2.

Table 4: **Hit identification criteria.** We defined experimental hits based on two criteria: (1) Observed $\mathrm{R_{max}} > 15\,\mathrm{pg/mm^2}$ and (2) dissociation and association errors ≤ 100. The theoretical $\mathrm{R_{max}}$ value was computed as $\mathrm{R_{max}} = R_L \cdot S_m \cdot \frac{\text{analyte MW}}{\text{ligand MW}}$ where R_L is the amount of immobilized ligand in $\mathrm{pg/mm^2}$ and S_m is the stoichiometry (number of binding sites for the analyte on the ligand).

Compound	Theoretical $\mathrm{R_{max}}$	Observed $\mathrm{R_{max}}$	Diss. Error	Ass. Error	Hit?
Compound 1	205	29	9	50	yes
Compound 2	209	153	9.2	32	yes
Compound 3	236	90	53	37	yes
Compound 4	212	4	120	120	no
Compound 5	240	4	4.8×10^8	170	no

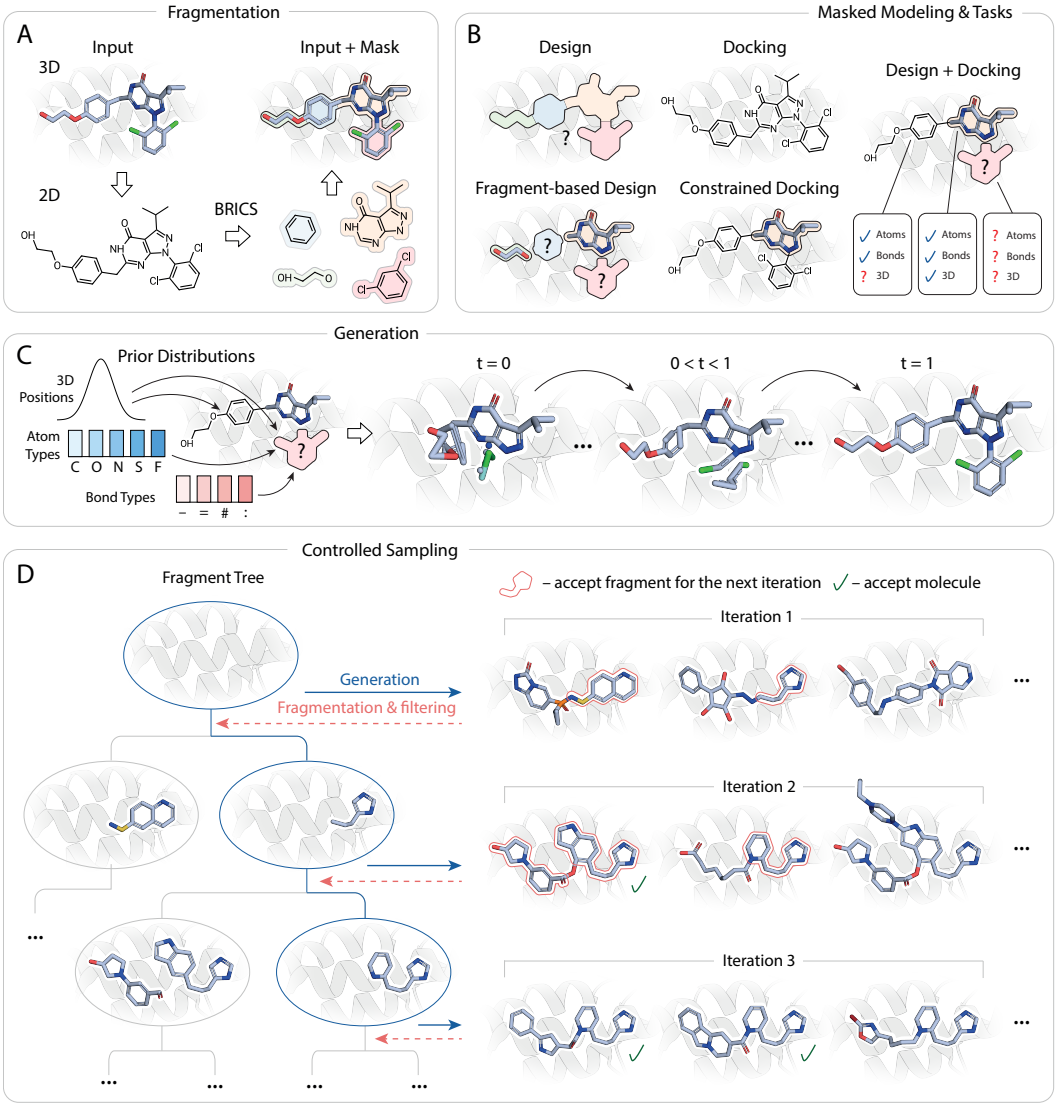

Figure 5: **Method overview.** (A) To prepare the data for training, we identify retrosynthetically interesting substructures using the BRICS (Degen et al., 2008) method, and assign fragment labels to different parts of each molecule. (B) During training, each fragment can either be masked out completely, partially given as a molecular graph without coordinates or fully provided as additional context for the denoising network. The trained model can then be applied naturally to diverse tasks by specifying the appropriate amount of given information for each fragment. (C) The generative model denoises continuous atom positions and simultaneously categorical atom and bond types for de novo designed fragments. (D) Controlled generation scheme. The process starts with the empty protein pocket (on the left) for which LDDM samples the first generation of molecules (on the right). These samples are then fragmented, and each fragment is accepted or rejected based on one or several validation criteria. The next iteration starts with these accepted fragments, which are input to LDDM to sample the next generation of molecules. At each step, we also evaluate the entire molecules, and the ones that pass all the validation filters are saved in the final pool. The process continues for a specified number of iterations or until a certain number of molecules is saved to the final pool.

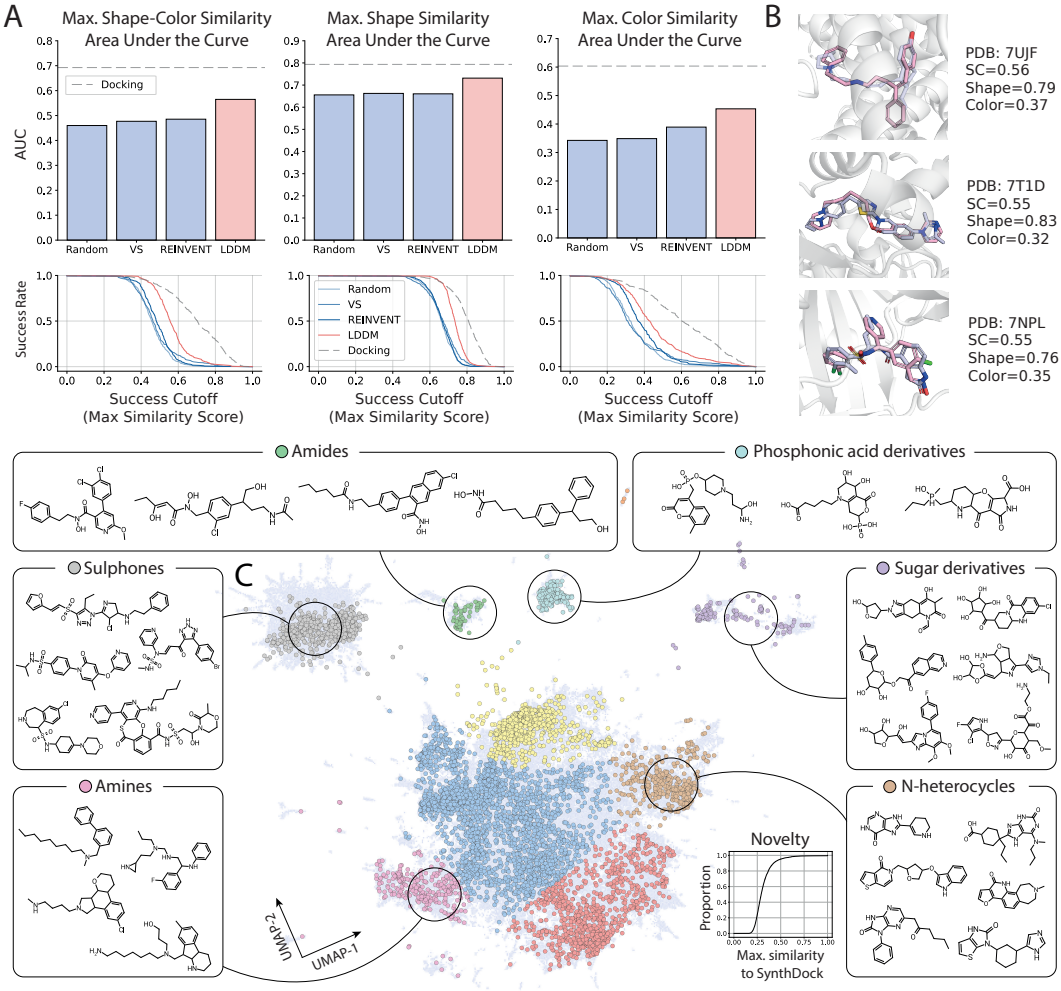

Figure 6: **Molecular design.** **(A)** Ground-truth recovery. We use SC_{RDKit} to efficiently compute shape and color (i.e. pharmacophore-based) similarity between generated 3D molecules and reference compounds from PoseBusters test set (304 targets). For each test target, we consider 100 generated (or proposed) candidates and choose the most similar one to the reference molecule. In the absence of a universal similarity cutoff, we report the area under the curve (AUC) of the cumulative success rate for the whole range of similarity values (i.e. between 0 and 1). As a positive control, we evaluate docking conformations of the reference molecules (100 conformations per target) obtained using AutoDock Vina (dashed gray lines). **(B)** Examples of LDDM samples (red) with high similarity to the reference molecules (gray). In all pairs, both molecules have high volumetric overlap and similar pharmacophore patterns. **(C)** Chemical space of LDDM samples ($n = 6856$). Each generated molecule is represented as a point corresponding to the two-dimensional UMAP vector computed on the PCA-projection (with 100 components) of the molecule's ChemNet embedding (with 512 components). Different colors of LDDM samples correspond to the clusters computed on the UMAP vectors using spectral clustering algorithm with the predefined number of clusters ($n = 10$). Six panels around the map provide examples of LDDM samples from different clusters. For each cluster, we show several compounds manually selected from the 10 closest to the cluster centroid molecules. While LDDM entirely covers the chemical space of the SynthDock training set (blue background, $n = 254140$), it produces novel chemical matter as shown in the cumulative histogram of novelty. To quantify novelty of a generated molecule, we compute its maximum Tanimoto similarity (using Morgan fingerprints) to the SynthDock training set. As shown in the plot, approximately 90% of samples have less than 0.5 similarity to the training set.