FLOWR – FLOW MATCHING FOR STRUCTURE- AND INTERACTION- AWARE *de novo* LIGAND GENERATION

Julian Cremer[1,*,†], **Ross Irwin**[2,3,*,†], **Alessandro Tibo**[2], **Jon Paul Janet**[2], **Simon Olsson**[3], **Djork-Arné Clevert**[1]

[1]Machine Learning & Computational Sciences, Pfizer Worldwide R&D, Berlin, Germany
[2]Molecular AI, Discovery Sciences, R&D, AstraZeneca, Gothenburg, Sweden
[3]Department of Computer Science and Engineering,
Chalmers University of Technology and University of Gothenburg,
Gothenburg, Sweden
[*]Corresponding authors. Email: julian.cremer@pfizer.com, rossir@chalmers.se
[†]These authors contributed equally to this work.

ABSTRACT

We present our progress on overcoming key challenges in applying generative models to 3D ligand design, including generating high-quality binders and reducing inference times. We introduce FLOWR, a flow matching framework for 3D ligand generation conditioned on a protein pocket and a set of desired interaction between the protein and the ligand. To thoroughly evaluate our model we also introduce SPIRE, a refined dataset of high-quality protein-ligand complexes derived from crystallographic data. Evaluations on this dataset show that FLOWR outperforms an existing state-of-the-art diffusion model, while achieving up to a 50-fold speed-up in inference time. We also propose an interaction-aware training and inference strategy that enables the generation of novel ligands tailored to predefined interaction profiles. Our findings suggest that FLOWR is an important step forward for efficient, AI-driven *de novo* ligand generation.

1 INTRODUCTION

Structure-based drug design (SBDD) aims to design ligands which can bind to a desired protein pocket in order to effectively modulate the protein's biological function (Anderson, 2003; 2012). Despite its successes, SBDD remains challenging due to the inherent complexity of molecular interactions, the vast chemical space to be explored, and the difficulty in accurately predicting ligand binding poses and affinities (Ferreira et al., 2015; Shoichet, 2004). Due to their ability to accurately capture the geometric properties of protein-ligand interactions, generative models based on diffusion approaches (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020) have emerged has a promising tool for 3D molecule generation (Guan et al., 2023; Schneuing et al., 2023; Le et al., 2023; Cremer et al., 2024). However, existing diffusion-based models for SBDD have been shown to generate molecules with strained conformations, uncommon substructures, and reduced drug-likeness (Cremer et al., 2024), while suffering from very long sampling times compared to other approaches (Irwin et al., 2024). Additionally, due to data leakage issues in commonly used datasets (Škrinjar et al., 2025; Durairaj et al., 2024), assessing the ability of these models to generalise to unseen data has been challenging.

In this work we present our progress on overcoming some of these key challenges. We introduce FLOWR, a novel flow matching model for *de novo* 3D ligand generation conditioned on structural constraints. Our approach allows ligands to be generated based on a protein pocket and, optionally, a desired interaction profile between the pocket and the ligand. To train our model and to assess its ability to generalise to unseen pockets we introduce SPIRE, a novel high-quality benchmark dataset for SBDD, based on the recently introduced PLINDER (Durairaj et al., 2024) set. To construct SPIRE we apply an extensive filtering and processing pipeline which cleans up many of the structural defects that are present in existing datasets (Wang et al., 2024), infers atomic resolution protein-ligand interaction profiles, and minimises data leakage between train and test sets. On this challenging

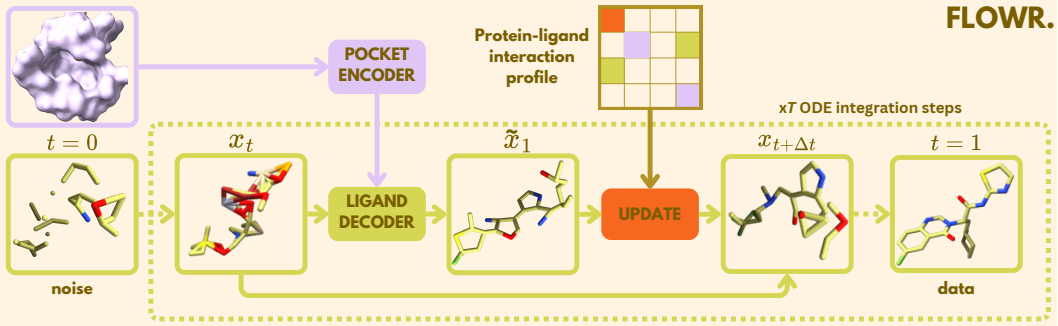


Figure 1: Overview of the FLOWR model for 3D ligand generation. A protein pocket is encoded and passed, along with the noisy ligand l_t, into the ligand decoder, which is trained to produce a denoised ligand $\tilde{l}_t$. Optionally, a set of desired interactions can be incorporated for conditional generation. A flow matching integration scheme is then used to push l_t towards the data distribution and generate a sample $\tilde{l}_1$.

dataset FLOWR demonstrates superior performance over an existing state-of-the-art diffusion-based method, while achieving up to 50-times faster inference times. Furthermore, we show that our method for generating ligands conditioned on a desired interaction profile leads to a significant increase in the number of generated ligands with interaction profiles that match that of the reference complex.

2 THE SPIRE DATASET

Modelling interactions between protein pockets and ligands has recently been gaining attention as a method for evaluating the quality of binding poses and designing better small molecule drug candidates (Errington et al., 2024; Harris et al., 2023). At the same time questions have been raised about the quality of existing benchmark datasets – PDBBind (Wang et al., 2005) has been found to contain covalently bound ligands, missing atoms in pockets, and steric clashes between the pocket and the ligand (Wang et al., 2024). CrossDocked2020 (Francoeur et al., 2020), another commonly used dataset for pocket-conditioned ligand generative models, is based on the PDBBind General set and is also likely to share some of these structural defects. Additionally, questions have also been raised as to how well temporal data splits, which are commonly used to create benchmark test sets, are able to assess models' abilities to generalise to unseen data since there are often close structural similarities between complexes in the training and test sets.

To address the issues of data quality and information leakage, and to provide rich, fine-grained information on the interactions between protein pockets and small molecule ligands, we present the SPIRE (Small molecule Protein Interaction Refined) dataset. Using the recently proposed PLINDER dataset (Durairaj et al., 2024) as a starting point we apply an extensive processing pipeline to produce a set refined set of high-quality structures. Our processing begins by filtering the complexes from PLINDER to retain only those containing single small molecule ligands and single proteins. We then apply structural refinement, including adding missing atoms, assigning protonation states, and performing energy minimisation. We also infer bond orders and atomic resolution protein-ligand interactions and perform a final quality filtering step. Our final dataset contains 35,666 protein-ligand complexes, making SPIRE the largest dataset of high-quality, refined structures derived directly from crystallographic data. We include the full details of the SPIRE dataset processing in Appendix A.1. We maintain the same data splits as PLINDER which were chosen to minimise data leakage between train and test sets which allows realistic assessment of models' generalisability to unseen data. Since existing datasets often contain significant structural redundancy, we also experiment with two data deduplication strategies, which we outline in Appendix A.2.

3 FLOWR – STRUCTURE-AWARE LIGAND GENERATION

We present FLOWR – a flow-based generative model for *de novo* ligand generation conditioned on a protein pocket and desired interactions between the pocket and the ligand. We assume access to a

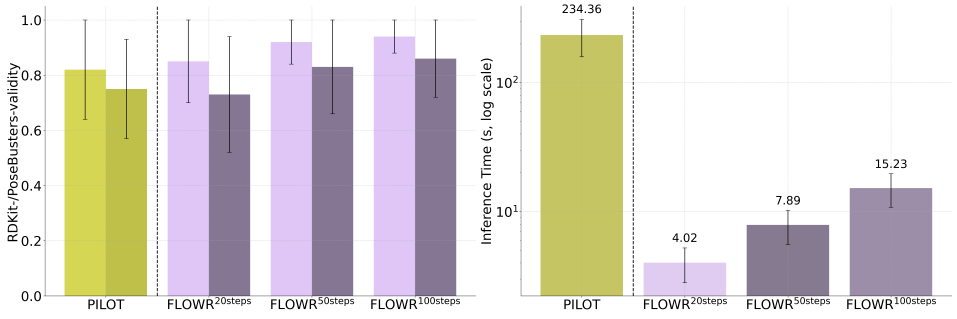


Figure 2: Comparison of PILOT and FLOWR in terms of RDKit- and PoseBusters-validity (left) and inference speed (right, log scale). Results for FLOWR are reported using three different inference step settings: 20, 50, and 100 steps. For each of the 225 targets in the SPIRE test set, we generate 100 ligands and compute the average validity scores and inference time per target.

dataset containing tuples of a ligand l, a protein pocket $\mathcal{P}$ to which the ligand binds, and a matrix $\mathcal{I} \in \mathbb{N}^{M \times N}$ of atomic protein-ligand interactions which the binding pose satisfies, where M and N refer to the number of atoms in the protein and ligand, respectively. In Fig 1 we show an overview of how our model generates novel ligands based on protein pocket and interaction conditioning.

Built upon SemlaFlow (Irwin et al., 2024), FLOWR extends its E(3)-equivariant architecture with a pocket encoder and a cross-attention module, enabling structural conditioning on $\mathcal{P}$ and $\mathcal{I}$. The pocket encoder processes $\mathcal{P}$ once per generation, ensuring efficiency, while improvements to self-attention and feed-forward modules further enhance performance. Full architectural details are in Appendix B.

FLOWR jointly models continuous (coordinates) and discrete (atom types, bond orders) molecular features. Training follows Irwin et al. (2024), using continuous flow matching (Tong et al., 2024) for coordinates and discrete flow models (Campbell et al., 2024) for categorical properties. Ligand formal charges are directly predicted. The model learns to recover l_1 from l_t via $p^{\theta}_{1|t}(l_1|l_t, \mathcal{P}, \mathcal{I})$, minimizing mean-squared error for coordinates and cross-entropy for categorical features (Appendix C).

Given $\mathcal{P}$ and optionally $\mathcal{I}$, novel ligands are generated by iteratively refining an initial noisy ligand $l_0 \sim p_{\text{noise}}$. The model follows a learned vector field v^{θ}_t for continuous features and a discrete integration scheme for categorical attributes (Campbell et al., 2024). Full sampling details are in Appendix B.

4 EXPERIMENTS AND RESULTS

We compare FLOWR against PILOT, a recently proposed diffusion-based model (Cremer et al., 2024). The authors of PILOT report significant improvements in distribution learning and ligand quality, demonstrating superior performance compared to previous models such as TargetDiff (Guan et al., 2023) and DiffSBDD (Schneuing et al., 2023). Given PILOT's strong performance and its claimed state-of-the-art results, we use it as the primary baseline for comparison, although we aim to work towards benchmarking other models on the SPIRE dataset. Such an approach allows for a thorough evaluation against PILOT, while ensuring that all results are fair and reproducible (more details in Appendix B).

Due to the extra efficiency and scalability achieved by FLOWR we also investigate the impact of generating hydrogen atoms in the ligand explicitly. Hydrogen bonds are a crucial element of protein-ligand binding, but explicit ligand hydrogens have larger been neglected in prior studies. Finally, we also evaluate a FLOWR model which has been trained to generate ligands conditioned on protein pockets and desired interaction profiles.

4.1 RESULTS

In Fig. 2, we compare PILOT and FLOWR in terms of RDKit-validity, PoseBusters-validity (PB-validity), and inference speed. Our results indicate that FLOWR generates ligands with significantly

Table 1: Benchmark comparison of the proposed FLOWR model against the PILOT model using the SPIRE test dataset, which consists of 225 targets. For FLOWR, results are reported for inference steps of 20, 50, and 100. For both models, 100 ligands were sampled per target. The evaluation includes strain energy and AutoDock-Vina scores. Additionally, we report the Wasserstein distance of the generated ligands' bond angle and bond length distributions relative to those in the SPIRE test set.

MODEL	STRAIN ENERGY	VINA SCORE	VINA SCORE (MINIMIZED)	BONDANGLESW1	BONDLENGTHSW1 [10^{-2}]
TEST SET	30.07 ± 36.96	-7.69 ± 2.00	-7.88 ± 2.00	-	-
PILOT	73.50 ± 64.30	-6.06 ± 0.95	-6.45 ± 0.95	1.71 ± 1.1	0.6 ± 0.01
FLOWR$^{20\ \text{STEPS}}$	98.80 ± 130.14	-6.12 ± 0.83	-6.55 ± 0.82	1.97 ± 1.2	0.7 ± 0.02
FLOWR$^{50\ \text{STEPS}}$	72.20 ± 72.78	-6.28 ± 0.82	-6.70 ± 0.82	1.49 ± 0.9	0.5 ± 0.01
FLOWR$^{100\ \text{STEPS}}$	64.61 ± 69.50	-6.36 ± 0.85	-6.80 ± 0.82	1.34 ± 0.9	0.4 ± 0.01

Table 2: Benchmark of the proposed FLOWR model against the PILOT model on the SPIRE test dataset with explicit hydrogens in training and inference.

MODEL	STRAIN ENERGY	VINA SCORE	VINA SCORE (MINIMIZED)	BONDANGLESW1	BONDLENGTHSW1 [10^{-2}]
TEST SET	30.07 ± 36.96	-7.69 ± 2.00	-7.88 ± 2.00	-	-
PILOT	53.07 ± 22.84	-5.00 ± 0.65	-5.50 ± 0.66	2.81 ± 1.3	0.2 ± 0.02
FLOWR$^{100\ \text{STEPS}}$	54.11 ± 33.36	-6.48 ± 0.87	-6.86 ± 0.87	0.82 ± 0.8	0.1 ± 0.01

higher validity on average. While RDKit-validity is a 2D ligand-centric measure, the PoseBusters suite (Buttenschoen et al., 2024) evaluates ligand conformations using well-established 3D ligand-pocket-based metrics, providing a more comprehensive assessment of pose accuracy. FLOWR achieves a substantial improvement over PILOT in both metrics, with an average RDKit-validity of 0.94 ± 0.24 vs. 0.82 ± 0.39 and an average PB-validity of 0.86 ± 0.21 vs. 0.75 ± 0.18, respectively. Notably, FLOWR significantly improves inference speed, outperforming PILOT by a factor of 15 when using 100 inference steps, as shown in Fig. 2 (right). This efficiency gain is primarily attributed to FLOWR's model architecture; the protein pocket encoder requires only a single forward when integrating the vector field. In contrast, prior models (Guan et al., 2023; Schneuing et al., 2023; Le et al., 2023; Cremer et al., 2024) often recompute protein pocket embeddings at every sampling step. Notably, the number of integration steps can be reduced as low as 20, achieving a 50-fold speed-up over PILOT without significantly impacting model performance. We provide full results for various numbers of integration steps in Appendix D.2

In Tab. 1 we compare PILOT and FLOWR in terms of strain energy, AutoDock-Vina score (used as an approximate measure of pose quality and binding affinity (Eberhardt et al., 2021)), and their ability to generalize to the test set distribution based on Wasserstein distance measures for bond angles and bond lengths, following Vignac et al. (2023); Le et al. (2023). Both models exhibit reasonable strain energy values, though on average, they are twice as high as the mean strain energy of the test set. However, FLOWR outperforms PILOT in docking assessments, suggesting a higher pose accuracy. Note, we use Vina's scoring function with no re-docking applied. We also report the minimized Vina score, where local energy minimisation is applied to the ligand. Additionally, in terms of bond angle and bond length distances, FLOWR demonstrates significantly better generalization compared to PILOT. A fully comprehensive overview of evaluation metrics can be found in Appendix D.1.

In Tab. 2, we repeat the same experiments while incorporating explicit hydrogens in the ligands for both training and inference. Under these conditions, PILOT exhibits a clear decrease in performance, while FLOWR maintains comparable results. For both models, validity drops significantly, with RDKit-validity decreasing to 0.64 ± 0.48 for FLOWR and 0.52 ± 0.50 for PILOT, while PB-validity declines to 0.60 ± 0.22 and 0.47 ± 0.14, respectively. However, since SPIRE provides limited coverage of both chemical and conformational space, we hypothesize that increasing data availability will alleviate this decline, particularly given the demonstrated learning efficiency of FLOWR.

Overall, FLOWR outperforms PILOT across all evaluated metrics, in some cases by large margins, indicating a superior ability to learn and generalize over the distribution of ligand-pocket complexes. On average, we observe a $\sim$10% increase in ligand and ligand-pocket validity, while Vina scores suggest that FLOWR generates significantly better poses. Importantly, FLOWR achieves these improvements while achieving a significant increase in inference speed.

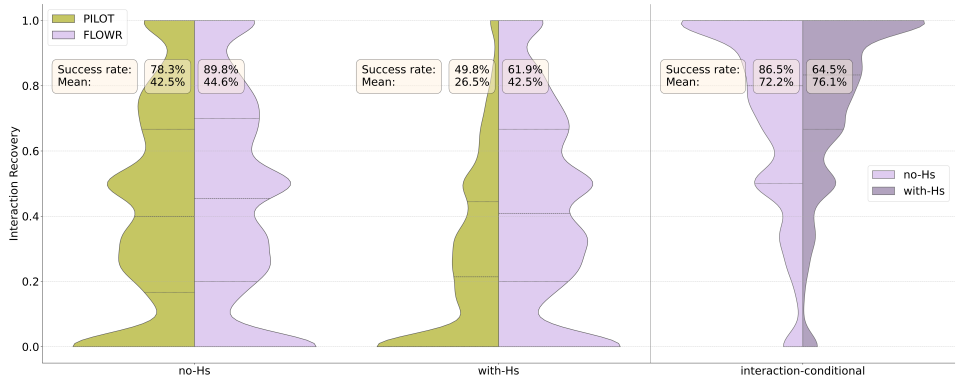


Figure 3: Comparison of PILOT and FLOWR on interaction recovery rates (left). Both models are either trained without explicit hydrogens (no-Hs) or with explicit hydrogens (with-Hs). We also show the results for interaction-conditional sampling with FLOWR (right). The success rate is the percentage of ligands for which interaction fingerprints could be retrieved for 100 sampled ligands for every test set target.

4.2 INTERACTION RECOVERY

In SBDD, understanding how a ligand engages with its target binding site at the atomic level is essential for optimising potency, selectivity, and pharmacological properties (Salentin et al., 2015; Jubb et al., 2016; Bouysset & Fiorucci, 2021). Ligand-pocket interactions, including hydrogen bonds, hydrophobic contacts, π–π and π–cation stacking, salt bridges, and electrostatic or van der Waals interactions, collectively determine binding affinity and specificity. To systematically identify such interactions, protein-ligand interaction fingerprints (PLIFs) are commonly used (Bouysset & Fiorucci, 2021; Errington et al., 2024). In Figure 3 we illustrate the distribution of interaction recovery performance of FLOWR across the SPIRE test set targets, following the same evaluation setting as above. We also report the success rate which refers to the proportion of RDKit- and PoseBusters-valid ligands for which interactions could be identified. As shown, FLOWR consistently outperforms PILOT, particularly when considering explicit hydrogen modelling, achieving an average interaction recovery rate of 44%/42% with a success rate of 89%/62%, whereas PILOT achieves 42%/26% with a success rate of 78%/50%.

However, to further improve interaction recovery, we propose an interaction-based masking for training and inference that is applied onto the learned vector field, which ensures that ligand atoms involved in pocket interactions are kept fixed. More details are given in Appendix C. Using this strategy, we achieve an average interaction recovery rate of 72.2%/76.1%, while maintaining a high success rate of 86%/61%. Notably, despite the guided generation process, the model retains its ability to explore chemical space, with an average molecular diversity of 0.83/0.84 (compared to 0.86/0.87 for the unconditional model). Additionally, this approach significantly improves binding affinity, as indicated by a decrease in Vina score to -6.93/-6.85 kcal/mol (vs. -6.36/-6.48 kcal/mol in the unconditional setting).

5 CONCLUSION

We introduced FLOWR, a flow matching-based generative framework for structure-conditioned 3D ligand design. Alongside FLOWR, we presented SPIRE, a refined dataset of high-quality protein-ligand complexes along with their interaction profiles. In comparison to the recent state-of-the-art diffusion-based PILOT model, FLOWR consistently demonstrates higher validity and improved ligand-pocket interaction recovery rates. Notably, FLOWR achieves a 10% increase in PoseBusters-validity, while offering up to a 30-fold improvement in inference speed. These results suggest that FLOWR provides more reliable pose quality and more consistent pocket-specific interactions than existing approaches. A key distinction of FLOWR is its ability to explicitly model hydrogens in ligand molecules, addressing a critical factor for achieving physically plausible ligand-pocket interactions. Unlike prior models, FLOWR is able to account for explicit hydrogen placement without a significant

loss in model expressivity. Furthermore, we proposed a novel interaction-based training and inference scheme, enabling the targeted generation of ligands that fulfill pre-specified interaction profiles. This approach enhances FLOWR's applicability in early-stage drug discovery, particularly in hit expansion and lead optimization campaigns.

REFERENCES

Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants, 2023. URL https://arxiv.org/abs/2209.15571.

Amy C. Anderson. The process of structure-based drug design. *Chemistry & Biology*, 10(9):787–797, 2003. ISSN 1074-5521. doi: https://doi.org/10.1016/j.chembiol.2003.09.002. URL https://www.sciencedirect.com/science/article/pii/S1074552103001947.

Amy C. Anderson. *Structure-Based Functional Design of Drugs: From Target to Lead Compound*, pp. 359–366. Humana Press, Totowa, NJ, 2012. ISBN 978-1-60327-216-2. doi: 10.1007/978-1-60327-216-2_23. URL https://doi.org/10.1007/978-1-60327-216-2_23.

Benoit Baillif, Jason Cole, Patrick McCabe, and Andreas Bender. Benchmarking structure-based three-dimensional molecular generative models using genbench3d: ligand conformation quality matters, 2024. URL https://arxiv.org/abs/2407.04424.

Cédric Bouysset and Sébastien Fiorucci. Prolif: a library to encode molecular interactions as fingerprints. *Journal of Cheminformatics*, 13(1):72, Sep 2021. ISSN 1758-2946. doi: 10.1186/s13321-021-00548-6. URL https://doi.org/10.1186/s13321-021-00548-6.

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, 2024. URL http://dx.doi.org/10.1039/D3SC04185A.

Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design, 2024. URL https://arxiv.org/abs/2402.04997.

Ting Chen, Ruixiang ZHANG, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=3itjR9QxFw.

Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking, 2023. URL https://arxiv.org/abs/2210.01776.

Julian Cremer, Tuan Le, Frank Noé, Djork-Arné Clevert, and Kristof T Schütt. Pilot: Equivariant diffusion for pocket conditioned de novo ligand generation with multi-objective guidance via importance sampling. *arXiv preprint arXiv:2405.14925*, 2024.

Janani Durairaj, Yusuf Adeshina, Zhonglin Cao, Xuejin Zhang, Vladas Oleinikovas, Thomas Duignan, Zachary McClure, Xavier Robin, Danny Kovtun, Emanuele Rossi, Guoqing Zhou, Srimukh Veccham, Clemens Isert, Yuxing Peng, Prabindh Sundareson, Mehmet Akdel, Gabriele Corso, Hannes Stärk, Zachary Carpenter, Michael Bronstein, Emine Kucukbenli, Torsten Schwede, and Luca Naef. Plinder: The protein-ligand interactions dataset and evaluation resource. *bioRxiv*, 2024. doi: 10.1101/2024.07.17.603955. URL https://www.biorxiv.org/content/early/2024/07/17/2024.07.17.603955.

Jerome Eberhardt, Diogo Santos-Martins, Andreas F. Tillack, and Stefano Forli. Autodock vina 1.2.0: New docking methods, expanded force field, and python bindings. *Journal of Chemical Information and Modeling*, 61(8):3891–3898, Aug 2021. ISSN 1549-9596. doi: 10.1021/acs.jcim.1c00203. URL https://doi.org/10.1021/acs.jcim.1c00203.

David Errington, Constantin Schneider, Cédric Bouysset, and Frédéric A. Dreyer. Assessing interaction recovery of predicted protein-ligand poses, 2024. URL https://arxiv.org/abs/2409.20227.

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):1–11, 2009.

Leonardo G Ferreira, Ricardo N Dos Santos, Glaucius Oliva, and Adriano D Andricopulo. Molecular docking and structure-based drug design strategies. *Molecules*, 20(7):13384–13421, July 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, Sep 2020. ISSN 1549-9596. doi: 10.1021/acs.jcim.0c00411. URL https://doi.org/10.1021/acs.jcim.0c00411.

Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky TQ Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. *arXiv preprint arXiv:2407.15595*, 2024.

Jiaqi Guan, Wesley Wei Qian, Xingang Peng, Yufeng Su, Jian Peng, and Jianzhu Ma. 3d equivariant diffusion for target-aware molecule generation and affinity prediction. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=kJqXEPXMsE0.

Charles Harris, Kieran Didi, Arian R Jamasb, Chaitanya K Joshi, Simon V Mathis, Pietro Lio, and Tom Blundell. Benchmarking generated poses: How rational is structure-based drug design with generative models? *arXiv preprint arXiv:2308.07413*, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Ross Irwin, Alessandro Tibo, Jon Paul Janet, and Simon Olsson. Efficient 3d molecular generation with flow matching and scale optimal transport, 2024. URL https://arxiv.org/abs/2406.07266.

Harry C Jubb, Alicia P Higueruelo, Bernardo Ochoa-Montaño, Will R Pitt, David B Ascher, and Tom L Blundell. Arpeggio: A web server for calculating and visualising interatomic interactions in protein structures. *J Mol Biol*, 429(3):365–371, December 2016.

Leon Klein, Andreas Krämer, and Frank Noé. Equivariant flow matching, 2023. URL https://arxiv.org/abs/2306.15030.

Tuan Le, Julian Cremer, Frank Noé, Djork-Arné Clevert, and Kristof Schütt. Navigating the design space of equivariant diffusion-based generative models for de novo 3d molecule generation, 2023.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Sebastian Salentin, Sven Schreiber, V Joachim Haupt, Melissa F Adasme, and Michael Schroeder. PLIP: fully automated protein-ligand interaction profiler. *Nucleic Acids Res*, 43(W1):W443–7, April 2015.

Arne Schneuing, Yuanqi Du, Charles Harris, Arian Jamasb, Ilia Igashov, Weitao Du, Tom Blundell, Pietro Lió, Carla Gomes, Max Welling, Michael Bronstein, and Bruno Correia. Structure-based drug design with equivariant diffusion models, 2023. URL https://arxiv.org/abs/2210.13695.

Brian K Shoichet. Virtual screening of chemical libraries. *Nature*, 432(7019):862–865, December 2004.

Peter Škrinjar, Jérôme Eberhardt, Janani Durairaj, and Torsten Schwede. Have protein-ligand co-folding methods moved beyond memorisation? *bioRxiv*, 2025.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Alexander Tong, Kilian FATRAS, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=CD9Snc73AW.

Clément Vignac, Nagham Osman, Laura Toni, and Pascal Frossard. Midi: Mixed graph and 3d denoising diffusion for molecule generation. In *Machine Learning and Knowledge Discovery in Databases: Research Track - European Conference, ECML PKDD 2023, Turin, Italy, September 18-22, 2023, Proceedings, Part II*, volume 14170 of *Lecture Notes in Computer Science*, pp. 560–576. Springer, 2023. doi: 10.1007/978-3-031-43415-0_33. URL https://doi.org/10.1007/978-3-031-43415-0_33.

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.

Yingze Wang, Kunyang Sun, Jie Li, Xingyi Guan, Oufan Zhang, Dorian Bagni, and Teresa Head-Gordon. Pdbbind optimization to create a high-quality protein-ligand binding dataset for binding affinity prediction. *arXiv preprint arXiv:2411.01223*, 2024.

A ADDITIONAL DETAILS ON THE SPIRE DATASET

A.1 PREPROCESSING PIPELINE

To create the SPIRE dataset, we began with the PLINDER dataset release 06/2024 (PLINDER version 2) and applied the following pre-processing pipeline:

1. **Initial filtering**. We remove all PLINDER systems which contain more than one ligand or more than one protein chain in the pocket. We then remove all systems where the ligand is marked as one or more of the following: 'oligo', 'ion', 'cofactor', 'artifact', 'fragment', 'covalent', or 'other'.
2. **Structure refinement**. We use Schrodinger tools to refine the structure of the remaining systems. These tools perform the following:
   (a) Add missing atoms to partially filled residues in the protein.
   (b) Convert some non-standard residue types to standard ones.
   (c) Assign protonation states to heavy atoms and add hydrogen atoms to both the protein and ligand.
   (d) Infer bonds and formal charges for both the protein and ligand.
   (e) Apply local energy minimisation to the protein-ligand complex.
3. **Infer protein-ligand interactions**. We use ProLIF to infer the interactions between the protein and ligand at an atomic resolution, creating a binary matrix of shape N_{prot}xN_{lig}x$|S|$, where N_{prot} is the number of atoms in the protein, N_{lig} is the number of atoms in the ligand, and S is the set of possible interaction types. We apply ProLIF with the default settings and infer all supported interaction types, $|S|$ = 13.
4. **Quality filtering**. Finally, we apply a final filtering step and accumulate the processed systems into train, validation and testing splits. Here we ensure that all systems contain RDKit valid ligands. We also filter out any system which contains atoms other than [H, C, N, O, F, P, S, Cl, Se, Br] and any system with fewer than 5 residues in the pocket. Additionally, we filter out all systems containing NAG ligands since we found these were highly overrepresented which would likely create an unwanted bias for generative models. We also filter out all systems derived from the PDB complex "1mvm" since it contains many small DNA fragments and was not originally filtered by PLINDER.

A.2 DATASET DEDUPLICATION

Like existing datasets of protein-ligand complexes, the SPIRE training set contains many redundant systems – systems which have significant structural similarity to another training system. Understanding the impact of this redundancy on model performance is a relatively unexplored topic but could have an important influence on the design of future datasets. We therefore apply two data deduplication strategies to SPIRE and report results on all three datasets. Deduplication is only applied to the training data and all models are evaluated identically.

Our first deduplication strategy works by creating groups of systems such that all systems within the group have identical ligands (based on their canonical SMILES after hydrogen atoms have been removed) and identical pockets atoms where the pocket coordinates are within an RMSD of 1.0 of some reference system for the group. We find that for system groups defined like this the distribution of RMSD values to the reference is very close to zero, so the choice of reference system and the RMSD threshold is not so important. In practice we iterate over all systems in the dataset, if a system cannot be added to an existing group a new group is created with this system as the group's reference system. Once all systems in the training dataset have been grouped a single system is randomly selected from each group to form the deduplicated training set. We refer to this dataset as SPIRE$^{\text{RMSD}}$. We also explore an extension of this deduplication strategy which allows systems to be in the same group if the sequence identity between a query pocket and the pocket of the reference system for a group is greater than 90%. In this case the RMSD between the query and reference pockets is taken by comparing the coordinates only on matching residues. Again, once groups have been constructed, a single system is randomly sampled from each group to form the deduplicated training set. We refer to this dataset as SPIRE$^{\text{RMSD-SEQID}}$. The sizes of the three versions of the dataset are shown in Table 3.

Table 3: Sizes of train, validation and test dataset splits for the three proposed versions of the SPIRE dataset.

Dataset	Train Systems	Val Systems	Test Systems
SPIRE	35,373	68	225
SPIRE$^{\text{RMSD}}$	24,885	68	225
SPIRE$^{\text{RMSD-SEQID}}$	20,349	68	225

B ADDITIONAL MODEL DETAILS

B.1 BACKGROUND ON FLOW MATCHING

Here we provide a short introduction to the flow matching methods used in this paper and introduce the notation we will use for the full training and inference details below.

Flow Matching for Continuous Data Flow matching (Lipman et al., 2023; Albergo & Vanden-Eijnden, 2023; Liu et al., 2023) is a generative model framework which aims to transport samples from an initial distribution p_0 to a target distribution p_1 by learning a vector field $v_t^\theta(x_t)$, which induces a time-dependent density p_t with p_0 and p_1 as endpoints. The key insight in flow matching is that such a vector field can be learned by firstly sampling data $x_1 \sim p_1(x_1)$, then sampling from a conditional probability path $x_t \sim p_{t|1}(x_t|x_1)$, which has an associated vector field $u_t(x_t|x_1)$, and finally regressing $v_t^\theta(x_t)$ against $u_t(x_t|x_1)$ (Tong et al., 2024).

Discrete Flow Models Frameworks for generating discrete sequences based on continuous-time markov chains (CTMC) have recently been proposed as an extension of flow matching to categorical data (Campbell et al., 2024; Gat et al., 2024). These methods work in a similar way to continuous flow matching by firstly defining a conditional probability path $p_{t|1}(.|x_1)$ and then learning a data denoiser $p_{1|t}^\theta(.|x_t)$ which is used during inference to push x_t towards the data distribution. The full details of the discrete flow model method we use in this paper can be found in (Campbell et al., 2024).

B.2 MODEL ARCHITECTURE

We base the neural network architecture for FLOWR off the recently proposed *SemlaFlow* model (Irwin et al., 2024), which achieves state-of-the-art results on unconditional 3D molecular generation tasks. SemlaFlow proposes Semla, an E(3)-equivariant architecture which includes a number of innovations making it significantly more efficient and scalable than previous models. We extend the Semla architecture to allow conditional generation by incorporating a separate pocket encoder and adding a cross attention module within the ligand decoder. This module follows a similar design to the attention module proposed by Semla, using a 2-layer MLP to produce attention scores. The module takes invariant and equivariant embeddings of $\mathcal{P}$ and l_t and, optionally, embeds $\mathcal{I}$, therefore allowing structural conditioning on the protein pocket and a set of desired protein-ligand interactions. We make use of the latent attention operation proposed in (Irwin et al., 2024) to significantly increase the efficiency of this operation. Notably, the pocket encoder module for FLOWR does not depend on t or l_t, meaning only one forward pass through the encoder is required when generating ligands, further ensuring the efficiency of our approach.

In addition to extending the architecture to allow conditional generation we also made improvements to various existing components within Semla, which we found we able to push the model's performance and efficiency even further. These improvements include:

- We replace the equivariant feed-forward module in Semla with a version based on a gating component. Specifically, if the invariant and equivariant input features for the component for atom i are denoted by $\mathbf{h}_i \in \mathbb{R}^{d_{\text{inv}}}$ and $\mathbf{x}_i \in \mathbb{R}^{3 \times d_{\text{equi}}}$, respectively, then the output is given by $\mathbf{x}_i^{\text{out}} = \mathbf{W}_\theta^2 \hat{\mathbf{x}}_i$ where $\hat{\mathbf{x}}_i = \sigma(\Phi_\theta(\mathbf{h}_i, ||\mathbf{x}_i||)) \odot \mathbf{W}_\theta^1 \mathbf{x}_i$. Here σ refers to an elementwise sigmoid function applied to invariant features, $\odot$ denotes elementwise multiplication, $\mathbf{W}_\theta^1 \in \mathbb{R}^{d_{\text{equi}} \times d_{\text{equi}}}$ and $\mathbf{W}_\theta^2 \in \mathbb{R}^{d_{\text{equi}} \times d_{\text{equi}}}$ are both weight matrices, and Φ_θ is

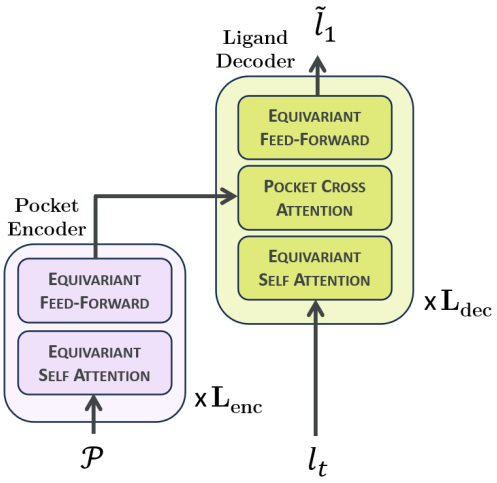


Figure 4: Overview of the equivariant architecture used by FLOWR. Our architecture extends the *Semla* architecture by adding a pocket encoder and a cross attention module within the ligand decoder. In addition to protein pockets, FLOWR also supports conditional generation based on protein-ligand interaction profiles.

a two-layer MLP. We find this module is significantly faster than the equivariant feed-forward block used by Semla.

- We pass bond embeddings into the self attention module on every layer, as opposed to only passing them to the first layer as proposed by Semla. We found this change led to improved validities of the generated molecules, while having only a very minor effect on inference time.

We parameterise FLOWR with a 4-layer pocket encoder with $d_{\text{inv}}^{\text{enc}} = 256$ and a 12-layer ligand decoder with $d_{\text{inv}}^{\text{dec}} = 384$. $d_{\text{equi}} = 64$ is the same for both encoder and decoder. For latent attention we use a latent size of 64 with 32 attention heads. A full overview of the FLOWR architecture for ligand generation conditioned on a pocket and interaction profile is shown in Figure 4.

B.3 TRAINING AND INFERENCE

We train FLOWR to generate novel ligands conditioned on a given structure. Since 3D molecular graphs contain a mixture of continuous and categorical data types, FLOWR jointly generates continuous and discrete distributions. Our approach follows a similar setup to Irwin et al. (2024). Specifically, we apply the continuous flow matching framework from Tong et al. (2024) to learn ligand coordinates, and the discrete flow models framework from Campbell et al. (2024) to learn atom types and bond orders. Ligand formal charges are not learned through a flow, but simply predicted by the model.

Model Training Training proceeds by sampling ligand noise $l_0 \sim p_{\text{noise}}$, a ligand, pocket and interaction tuple $(l_1, \mathcal{P}, \mathcal{I}) \sim p_{\text{data}}$, and a time $t \in [0, 1]$. We use Gaussian noise for coordinates and uniform distributions for atom and bond types to create p_{noise}. We then sample a noisy ligand from the same conditional probability path $l_t \sim p_{t|1}(l_t|l_1)$ used in Irwin et al. (2024) and is defined as follows:

$$t \sim \text{Beta}(\alpha, \beta) \qquad \mathbf{x}_t \sim \mathcal{N}(t\mathbf{x}_1 + (1-t)\mathbf{x}_0, \sigma^2) \tag{1}$$

$$\mathbf{a}_t \sim \text{Cat}(t\delta(\mathbf{a}_1) + (1-t)\frac{1}{|A|}) \qquad \mathbf{b}_t \sim \text{Cat}(t\delta(\mathbf{b}_1) + (1-t)\frac{1}{|B|}) \tag{2}$$

Where A and B are the sets of possible values for atom types and bond orders, respectively, and $\delta(.)$ is the one-hot encoding operation applied to each item in a sequence individually. We use values $\alpha = 2.0$, $\beta = 1.0$, and $\sigma = 0.2$ for all FLOWR models.

Following (Vignac et al., 2023; Le et al., 2023; Cremer et al., 2024) we train FLOWR to predict l_1 directly by learning the distribution $p^{\theta}_{1|t}(l_1|l_t, \mathcal{P}, \mathcal{I})$. This leads to the same loss function as

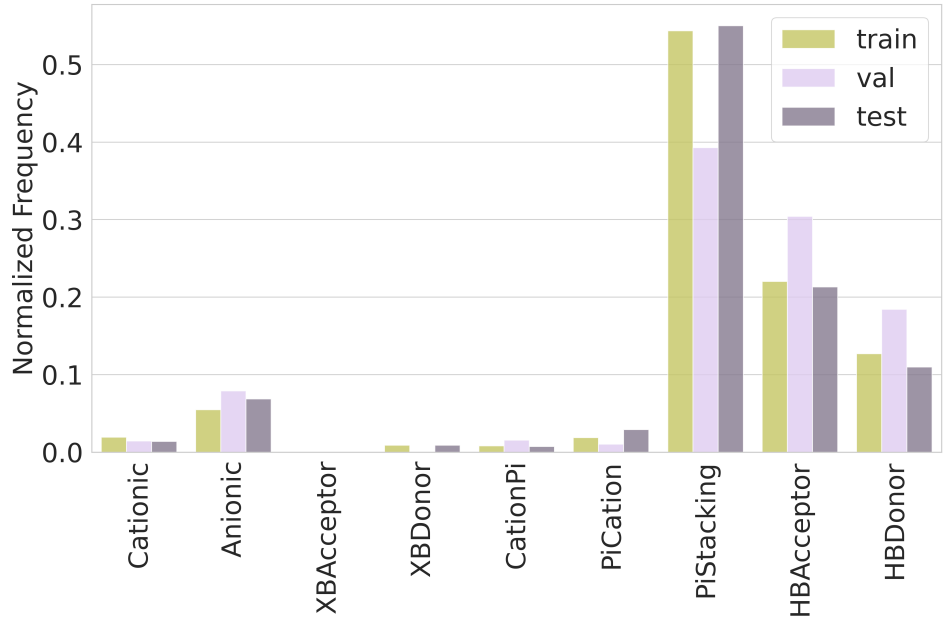


Figure 5: Distribution of interaction types on the train, validation and test sets of the SPIRE data that we considered in this work.

SemlaFlow (Irwin et al., 2024) – we apply a mean-squared error loss for ligand coordinates and cross-entropy losses for atom types, bond orders and formal charges. In Appendix C we provide more information on how we handle the case where the model is conditioned on both $\mathcal{P}$ and $\mathcal{I}$.

Additionally, during training we apply self-conditioning (Chen et al., 2023) as a way of reusing the model's previous prediction of l_1 and equivariant optimal transport (Klein et al., 2023) to reduce the transport cost between p_{noise} and p_{data}. Full details of the training setup for self-conditioning and equivariant optimal transport can be found in Irwin et al. (2024).

Generating Novel Ligands Given a protein pocket $\mathcal{P}$ and, optionally, a desired interaction profile Ψ, we can generate samples from the learned data distribution by setting $l_t \leftarrow l_0$ where $l_0 \sim p_{\text{noise}}$ and pushing l_t toward the data distribution by following the learned vector field. Specifically, for molecular coordinates $\mathbf{x}_t$ we follow the vector field $v_t^\theta(\mathbf{x}_t) = \frac{1}{1-t}(\tilde{\mathbf{x}}_1 - \mathbf{x}_t)$ where $\tilde{\mathbf{x}}_1$ is the coordinate component of $\tilde{l}_1 \sim p^\theta_{1|t}(l_1|l_t, \mathcal{P}, \mathcal{I})$. We then integrate the vector field using an Euler solver with step size Δt as follows: $\tilde{\mathbf{x}}_{t+\Delta t} = \mathbf{x}_t + \Delta t\ v_t^\theta(\mathbf{x}_t)$. We refer readers to Campbell et al. (2024) for full details on the integration scheme for discrete types.

B.4 EVALUATION

To maintain consistency across models, we used identical random seeds for training, inference, and data loading. Additionally, we applied the same sampling and evaluation scripts across all models. For each of the 225 test set targets, we generated 100 ligand samples using a standardized size sampling approach. Specifically, we determined native ligand sizes and applied a uniform sampling scheme, allowing for a size deviation of -25% to +10%. This procedure was performed using the same seed across all models to ensure direct comparability.

C INTERACTIONS

In SBDD, understanding how a ligand engages with its target binding site at the atomic level is essential for optimising potency, selectivity, and pharmacological properties (Salentin et al., 2015; Jubb et al., 2016; Bouysset & Fiorucci, 2021). Ligand-pocket interactions, including hydrogen bonds,

hydrophobic contacts, π–π and π–cation stacking, salt bridges, and electrostatic or van der Waals interactions, collectively determine binding affinity and specificity. Consequently, these protein-ligand interactions—or more precisely, a ligand's binding pose—are crucial for assessing biological relevance and activity(Errington et al., 2024). To systematically identify such interactions, protein-ligand interaction fingerprints (PLIFs) are commonly used (Bouysset & Fiorucci, 2021; Errington et al., 2024). PLIFs encode key interaction features, including the interacting protein residue, interaction type, and optionally, the ligand atom involved (Bouysset & Fiorucci, 2021; Errington et al., 2024). In the context of 3D *de novo* ligand generation, an important validation step is to assess whether the generated ligands recapitulate critical interactions known to be essential for activity, as inferred from experimentally validated compounds. Recent studies indicate that deep learning-based docking and co-folding tools perform poorly in recovering key interactions compared to traditional docking methods (Errington et al., 2024). Docking-based approaches such as DiffDock(Corso et al., 2023)—which operate by translating and rotating a valid molecular conformation while adjusting dihedral angles—achieve only a 20% success rate when targeting an interaction recovery rate of at least 50%(Errington et al., 2024). Following Errington et al. (2024), we consider a subset of interaction types in this work extracted using ProLIF (Bouysset & Fiorucci, 2021), including H-bonds (ligand acceptor and ligand donor), π-π stacking, halogen bonds (ligand donor), π-cation (ligand π / protein +), cation-π (ligand + / protein π), anionic (ligand - / protein +), and cationic (ligand + / protein -) interactions. The distribution of these interactions within the SPIRE dataset is shown in Fig. 5. Notably, interaction sparsity is high, with an average of 99.85% of ligand-pocket pairs exhibiting no interactions. Note, the frequency of π-π stacking is inflated as every atom involved in π-π stacking interactions has been counted.

C.1 INTERACTION-AWARE TRAINING AND SAMPLING

Let $\mathbf{X}_p = \{\mathbf{x}_{p,j} \in \mathbb{R}^3 : j = 1, \ldots, n_p\}$ denote the 3D coordinates of the n_p pocket atoms, $\mathbf{X}_l^{(0)} = \{\mathbf{x}_{l,i}^{(0)} \in \mathbb{R}^3 : i = 1, \ldots, n_l\}$ denote the ground-truth (native) 3D coordinates of the n_l ligand atoms, $I \in \{0,1\}^{n_p \times n_l \times d_I}$ be an interaction tensor, where the entry $I_{j,i,k}$ indicates whether pocket atom j and ligand atom i participate in an interaction of type k (with d_I possible interaction channels).

We define a binary mask $M \in \{0,1\}^{n_l}$ by

$$M_i = \mathbb{I}\left\{\sum_{j=1}^{n_p}\sum_{k=1}^{d_I} I_{j,i,k} > 0\right\}, \quad i = 1, \ldots, n_l.$$

This mask partitions the ligand atoms into: fixed atoms $\mathcal{I} = \{i : M_i = 1\}$ and free atoms to be generated $\mathcal{F} = \{i : M_i = 0\}$.

For atom coordinates (same applies to atom types in categorical space), we define a continuous interpolation over time $t \in [0,1]$ between a prior distribution (noise from an isotropic Gaussian) at $t = 0$ and the data distribution at $t = 1$.

For free atoms ($i \in \mathcal{F}$): Let $\mathbf{z}_i$ be a sample from the prior,

$$\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_3).$$

The interpolation is:

$$\mathbf{x}_{l,i}(t) = (1-t)\,\mathbf{z}_i + t\,\mathbf{x}_{l,i}^{(0)}.$$

For fixed atoms ($i \in \mathcal{I}$): Since these atoms are conditioned to remain unchanged, we simply set

$$\mathbf{x}_{l,i}(t) = \mathbf{x}_{l,i}^{(0)} \quad \text{for all } t \in [0,1].$$

Thus,

$$\dot{\mathbf{x}}_{l,i}(t) = \mathbf{0}.$$

We train a FLOWR model F_θ that maps the ligand coordinates (at time t), the pocket coordinates, and the time t to a vector in $\mathbb{R}^3$ for each ligand atom:

$$F_\theta : (\mathbf{X}_l(t),\, \mathbf{X}_p,\, t) \mapsto \mathbb{R}^{n_l \times 3}.$$

The model is trained to match the target vector field:

For free atoms ($i \in \mathcal{F}$):

$$F_\theta\Big(\mathbf{x}_{l,i}(t),\, \mathbf{X}_p,\, t\Big) \approx \dot{\mathbf{x}}_{l,i}(t) = \mathbf{x}_{l,i}^{(0)}$$

and for fixed atoms ($i \in \mathcal{I}$):

$$F_\theta\Big(\mathbf{x}_{l,i}(t),\, \mathbf{X}_p,\, t\Big) \approx \mathbf{0}.$$

At test time, given the pocket $\mathbf{X}_p$ and a known native ligand geometry $\mathbf{X}_l^{(0)}$ (with fixed atoms indicated by M), we generate the free atoms by solving the following ODE:

$$\frac{d}{dt}\mathbf{x}_{l,i}(t) = F_\theta\Big(\mathbf{x}_{l,i}(t),\, \mathbf{X}_p,\, t\Big),$$

with the initial conditions

$$\mathbf{x}_{l,i}(0) = \begin{cases} \mathbf{z}_i, & i \in \mathcal{F}, \\ \mathbf{x}_{l,i}^{(0)}, & i \in \mathcal{I}, \end{cases} \quad \text{and} \quad \mathbf{z}_i \sim p_{\text{prior}}.$$

For free atoms, integration from $t = 0$ to $t = 1$ transforms the prior samples into samples from the data distribution while maintaining consistency with the interaction condition.

To ensure that the model efficiently explores chemical space and generates diverse ligands while being conditioned on pre-specified interaction profiles, we train FLOWR using a mixed unconditional and conditional setting. This approach provides downstream flexibility, allowing sampling to be either unconditional or conditional at inference time, depending on the practitioner's needs. To validate this approach, we compare an unconditional model with a model trained in the mixed unconditional-conditional setting. For the latter, we evaluate both unconditional and conditional sampling modes (Tab. 4).

Table 4: Benchmarking an unconditional FLOWR model with an interaction-conditional model. The latter can be used either for unconditional or interaction-conditional sampling

METRIC	FLOWR	$\text{FLOWR}^{\text{COND}}_{\text{UNCOND}}$	$\text{FLOWR}^{\text{COND}}_{\text{COND}}$
RDKIT-VALIDITY	0.94 ± 0.24	0.93 ± 0.25	0.90 ± 0.30
PB-VALIDITY	0.92 ± 0.24	0.93 ± 0.23	0.90 ± 0.28
PLIF RECOVERY	0.44 ± 0.10	0.44 ± 0.10	0.71 ± 0.15
STRAIN ENERGY	64.61 ± 69.50	64.12 ± 91.85	88.68 ± 123.06
VINA SCORE	-6.36 ± 0.85	-6.45 ± 0.84	-7.01 ± 1.25
VINA SCORE (MINIMIZED)	-6.80 ± 0.82	-6.85 ± 0.81	-7.43 ± 1.23
ANGLESW1	1.34	1.25	1.51
BONDLENGTHSW1 [10^{-2}]	0.4 ± 0.01	0.3 ± 0.01	0.3 ± 0.01
DISTANCE TO NATIVE CENTROID	1.00 ± 0.30	1.02 ± 0.30	0.83 ± 0.38
NOVELTY	0.94 ± 0.23	0.94 ± 0.23	0.93 ± 0.25
UNIQUENESS2D	0.94 ± 0.13	0.94 ± 0.13	0.85 ± 0.25
UNIQUENESS3D	0.50 ± 0.20	0.57 ± 0.25	0.39 ± 0.18
DIVERSITY2D	0.86 ± 0.05	0.86 ± 0.05	0.83 ± 0.08
DIVERSITY3D	0.21 ± 0.12	0.21 ± 0.17	0.07 ± 0.10
SA	0.67 ± 0.13	0.68 ± 0.13	0.67 ± 0.13
QED	0.52 ± 0.21	0.54 ± 0.21	0.51 ± 0.21
RINGS	2.68 ± 1.35	2.74 ± 1.39	3.05 ± 1.35
AROMATIC RINGS	1.52 ± 1.16	1.59 ± 1.16	1.77 ± 1.18
HACCEPTORS	6.67 ± 4.23	6.40 ± 3.95	6.85 ± 4.21
HDONORS	2.52 ± 1.68	2.41 ± 1.65	2.67 ± 1.63
LOGP	0.29 ± 3.31	0.63 ± 3.24	0.67 ± 3.45
MOLWT	350.10 ± 114.00	350.90 ± 112.67	376.70 ± 113.49
LIPINSKI	4.35 ± 1.05	4.41 ± 0.99	4.35 ± 1.05

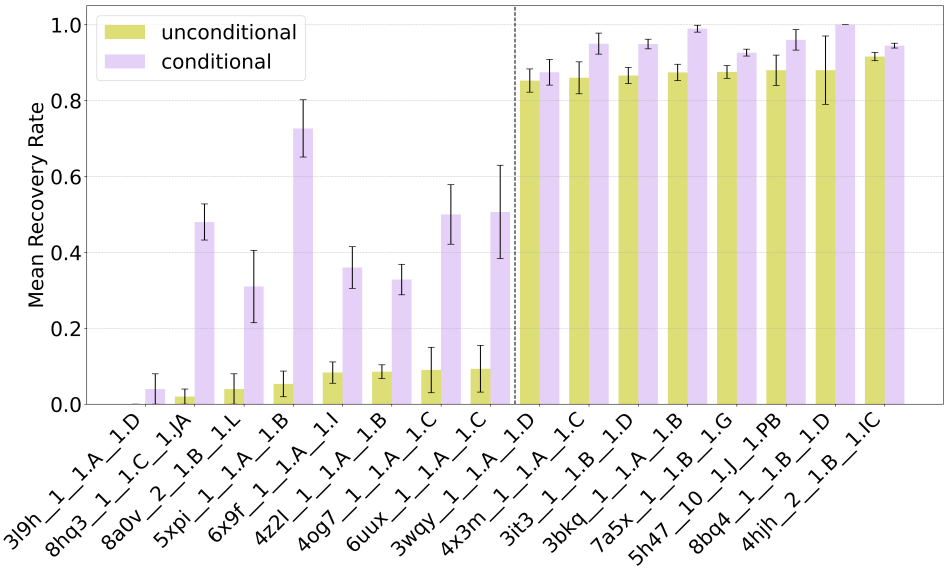


Figure 6: Comparison between unconditional and conditional FLOWR models. We identify eight targets with the lowest (left) and highest (right) average interaction recovery rates under the unconditional FLOWR model. For these selected targets, we compare the performance of the conditional model to assess the impact of conditioning on pocket-ligand interactions.

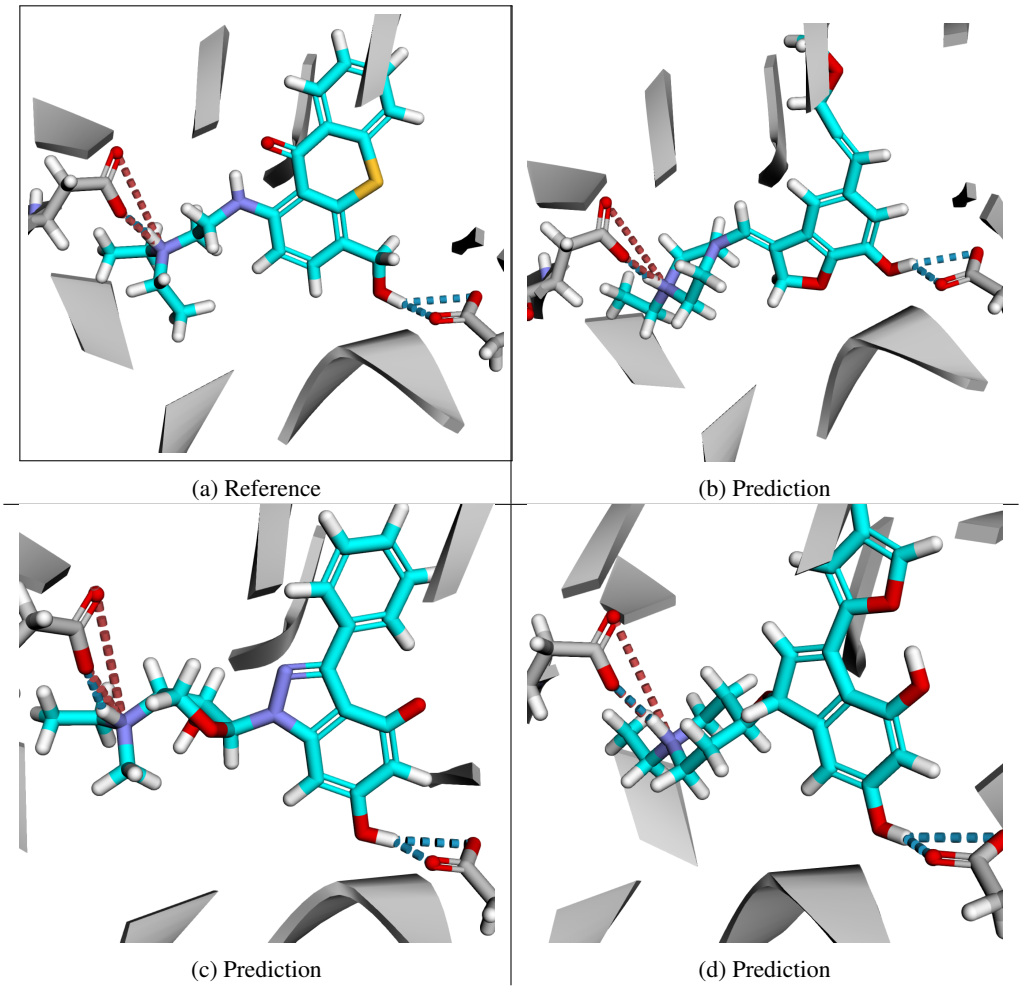
C.2 INTERACTIONS PER TARGET

To better evaluate the effectiveness of the proposed interaction-conditional training and sampling, we compare the unconditional and conditional FLOWR models on a per-target basis. Given that the test set comprises 225 targets, visualizing results for all targets is impractical. Instead, we select M targets with the lowest and with the highest mean interaction recovery rates, as determined by the unconditional model, and compare the corresponding results obtained using the conditional model. This comparison is presented in Fig. 6. Notably, the conditional model consistently improves interaction recovery across targets where the unconditional model struggled to generate ligands with meaningful interactions. Additionally, it achieves significantly better results even for the top-performing targets, demonstrating that interaction-conditional generation effectively enhances ligand design with pre-specified interaction patterns.

Figure 7 presents an example of interaction profiling using the reference ligand of protein 6UUX alongside three randomly selected ligands generated by the interaction-conditional FLOWR model. The reference ligand forms two cationic interactions and one H-bond (ligand donor) interaction with ASP149, as well as two H-bond (ligand donor) interactions with ASP93. Notably, all of these interactions are successfully recovered in the generated ligands.

D ADDITIONAL EXPERIMENTAL RESULTS

Benchmarking newly proposed models and architectures in the context of structure-based drug design requires careful consideration of multiple evaluation aspects. In addition to the results presented in the main text, we provide a broader assessment using various metrics and evaluation settings in the following sections. Specifically, we evaluate the novelty of generated ligands with respect to the training set, as well as the average uniqueness and diversity among the 100 generated ligands per target. To ensure a comprehensive analysis, we consider both SMILES string- and ECFP4-based measures for uniqueness and diversity. Additionally, following Baillif et al. (2024), we extend this analysis to include conformer-based uniqueness and diversity. As indicators of drug-likeness, we report RDKit's Quantitative Estimate of Drug-likeness (QED), the Synthetic Accessibility Score (SAScore) (Ertl & Schuffenhauer, 2009), molecular weight, logP values, and compliance with Lipinski's Rule of Five. Furthermore, we assess model performance under a more restrictive ligand



(a) Reference (b) Prediction

(c) Prediction (d) Prediction

Figure 7: Comparison of reference and predicted ligands on their interaction profiles for the pocket of the protein with PDB id 6uux.

size setting, where ligand sizes are not sampled but fixed to match the sizes of the native ligands. This evaluation provides insights into how the models perform when constrained to a stricter ligand size distribution. Finally, we analyze the impact of reducing the number of inference steps in FLOWR, which allows for further reductions in inference time.

D.1 EVALUATION

In Tab. 5 we report the results comparing PILOT and FLOWR for both settings, without explicit and with explicit hydrogens in training and inference, respectively. On average, PILOT shows higher novelty, uniqueness and diversity values of generated ligands. However, in light of the significantly worse results across distribution and ligand-pocket centric metrics, it is likely that PILOT has a stronger tendency to hallucinate and thus generates physically less plausible, but more diverse structures with higher strains. Regarding RDKit-based ligand property metrics, both models show similar results, while FLOWR shows in general a higher overlap with the test set values indicating better distribution learning capabilities.

Table 5: Benchmark of the proposed FLOWR model against the recent state-of-the-art diffusion-based PILOT model on the SPIRE dataset. We report RDKit- and PoseBusters-validity of generated ligands, the strain energy and the AutoDock-Vina score. We also state the Wasserstein distance of generated ligands for the bond angles and bond lengths distribution to the SPIRE test set. Novelty, uniqueness and diversity measure the capability of the model to explore the chemical space both in 2D and 3D. RDKit's QED evaluation, SAScore, the molecular weight as well as the logP values evaluate druglikeness of generated ligands. All presented values are mean values taken for 100 sampled ligands per test set target. The test dataset comprises 225 test set targets. Ligand sizes were drawn from a uniform distribution around the ground truth ligand size allowing for a deviation of -25% and + 10% with the same random seed for all models.

METRIC	TEST SET	PILOT$^{\text{NO-HS}}$	PILOT$^{\text{WITH-HS}}$	FLOWR$^{\text{NO-HS}}$	FLOWR$^{\text{WITH-HS}}$
RDKIT-VALIDITY	1.00 ± 0.00	0.82 ± 0.39	0.52 ± 0.50	0.94 ± 0.24	0.64 ± 0.48
PB-VALIDITY	0.99 ± 0.02	0.75 ± 0.18	0.47 ± 0.14	0.86 ± 0.21	0.60 ± 0.22
STRAIN ENERGY	30.07 ± 36.96	73.50 ± 64.30	53.07 ± 22.84	64.61 ± 69.50	54.11 ± 33.36
VINA SCORE	-7.69 ± 2.00	-6.06 ± 0.95	-5.00 ± 0.65	-6.36 ± 0.85	-6.48 ± 0.87
VINA SCORE (MINIMIZED)	-7.88 ± 2.00	-6.45 ± 0.95	-5.50 ± 0.66	-6.80 ± 0.82	-6.86 ± 0.87
DISTANCE TO NATIVE CENTROID	-	1.02 ± 0.40	1.53 ± 0.46	1.00 ± 0.30	0.98 ± 0.28
BONDANGLESW1	-	1.71 ± 1.1	2.81 ± 1.3	1.34 ± 0.9	0.82 ± 0.8
BONDLENGTHSW1 [10^{-2}]	-	0.6 ± 0.01	0.1 ± 0.02	0.4 ± 0.01	0.1 ± 0.01
NOVELTY	1.00 ± 0.00	0.99 ± 0.10	1.00 ± 0.00	0.94 ± 0.23	1.00 ± 0.00
UNIQUENESS2D	0.92 ± 0.10	0.99 ± 0.05	1.00 ± 0.02	0.94 ± 0.13	0.97 ± 0.07
UNIQUENESS3D	-	0.66 ± 0.20	0.59 ± 0.19	0.50 ± 0.20	0.55 ± 0.17
DIVERSITY2D	0.92 ± 0.04	0.89 ± 0.03	0.90 ± 0.02	0.86 ± 0.05	0.87 ± 0.06
DIVERSITY3D	-	0.25 ± 0.13	0.13 ± 0.19	0.21 ± 0.12	0.18 ± 0.11
SA	0.66 ± 0.12	0.63 ± 0.12	0.64 ± 0.10	0.67 ± 0.13	0.65 ± 0.10
QED	0.49 ± 0.22	0.51 ± 0.21	0.53 ± 0.18	0.52 ± 0.21	0.53 ± 0.21
RINGS	2.98 ± 1.42	2.52 ± 1.42	1.52 ± 0.98	2.68 ± 1.35	2.64 ± 1.43
AROMATIC RINGS	1.84 ± 1.31	1.12 ± 1.07	1.21 ± 0.95	1.52 ± 1.16	1.59 ± 1.22
HACCEPTORS	7.30 ± 4.49	6.19 ± 3.30	5.46 ± 2.21	6.67 ± 4.23	6.47 ± 3.64
HDONORS	2.62 ± 1.68	2.52 ± 1.65	1.55 ± 1.27	2.52 ± 1.68	2.66 ± 1.58
LOGP	0.29 ± 3.48	0.45 ± 3.08	-0.03 ± 2.33	0.29 ± 3.31	0.34 ± 2.99
MOLWT	390.43 ± 119.82	336.79 ± 107.86	337.30 ± 83.59	350.10 ± 114.00	336.09 ± 108.60
LIPINSKI	4.00 ± 1.34	4.45 ± 0.93	4.73 ± 0.55	4.35 ± 1.05	4.32 ± 1.05

D.2 NUMBER OF INFERENCE STEPS

Unlike diffusion models, empirical studies have shown that flow matching for molecular generation allows for modifying the number of sampling steps during inference without a significant loss in performance (Irwin et al., 2024). To evaluate this property, we benchmarked FLOWR using three different sampling step settings: 30, 50, and 100 (default). In Tab. 6, we summarize the results and compare them with PILOT, which requires 500 denoising steps. As expected, increasing the number of sampling steps generally leads to improved performance. However, even with just 50 sampling steps, FLOWR significantly outperforms PILOT across all evaluated metrics, achieving performance comparable to the 100-step model while being twice as fast. Notably, FLOWR with only 20 sampling steps still achieves performance close to PILOT, while offering a substantial efficiency gain, being on average 30 times faster. These findings highlight the flexibility of flow matching in balancing sampling efficiency and model performance, making it a promising approach for fast and scalable 3D ligand generation.

Table 6: We compare PILOT with FLOWR using different step sizes at sampling.

METRIC	PILOT$^{500\ \text{STEPS}}$	FLOWR$^{20\ \text{STEPS}}$	FLOWR$^{50\ \text{STEPS}}$	FLOWR$^{100\ \text{STEPS}}$
RDKIT-VALIDITY	0.82 ± 0.39	0.85 ± 0.35	0.92 ± 0.27	0.94 ± 0.24
PB-VALIDITY	0.75 ± 0.18	0.73 ± 0.17	0.83 ± 0.20	0.86 ± 0.21
PLIF RECOVERY	0.42 ± 0.31	0.43 ± 0.10	0.44 ± 0.10	0.44 ± 0.10
STRAIN ENERGY	73.50 ± 64.30	98.80 ± 130.14	72.20 ± 72.78	64.61 ± 69.50
VINA SCORE	-6.06 ± 0.95	-6.12 ± 0.83	-6.28 ± 0.82	-6.36 ± 0.85
VINA SCORE (MINIMIZED)	-6.45 ± 0.95	-6.55 ± 0.82	-6.70 ± 0.82	-6.80 ± 0.82
BONDANGLESW1	1.71 ± 1.1	1.97 ± 1.2	1.49 ± 0.9	1.34 ± 0.9
BONDLENGTHSW1 [10^{-2}]	0.6 ± 0.01	0.7 ± 0.02	0.5 ± 0.01	0.4 ± 0.01
DISTANCE TO NATIVE CENTROID	1.02 ± 0.40	0.98 ± 0.29	0.99 ± 0.29	1.00 ± 0.30
NOVELTY	0.99 ± 0.10	0.96 ± 0.20	0.95 ± 0.22	0.94 ± 0.23
UNIQUENESS2D	0.99 ± 0.05	0.96 ± 0.10	0.95 ± 0.12	0.94 ± 0.13
UNIQUENESS3D	0.66 ± 0.20	0.51 ± 0.24	0.50 ± 0.23	0.50 ± 0.20
DIVERSITY2D	0.89 ± 0.03	0.87 ± 0.04	0.86 ± 0.05	0.86 ± 0.05
DIVERSITY3D	0.25 ± 0.13	0.32 ± 0.12	0.25 ± 0.15	0.21 ± 0.12
SA	0.63 ± 0.12	0.64 ± 0.13	0.66 ± 0.13	0.67 ± 0.13
QED	0.51 ± 0.21	0.52 ± 0.21	0.52 ± 0.21	0.52 ± 0.21
RINGS	2.52 ± 1.42	2.70 ± 1.41	2.70 ± 1.37	2.68 ± 1.35
AROMATIC RINGS	1.12 ± 1.07	1.22 ± 1.06	1.43 ± 1.12	1.52 ± 1.16
HACCEPTORS	6.19 ± 3.30	6.51 ± 3.95	6.62 ± 4.16	6.67 ± 4.23
HDONORS	2.52 ± 1.65	2.57 ± 1.66	2.53 ± 1.67	2.52 ± 1.68
LOGP	0.45 ± 3.08	0.28 ± 3.19	0.29 ± 3.25	0.29 ± 3.31
MOLWT	336.79 ± 107.86	344.07 ± 111.97	349.12 ± 114.23	350.10 ± 114.00
LIPINSKI	4.45 ± 0.93	4.38 ± 1.00	4.36 ± 1.04	4.35 ± 1.05
SAMPLING TIME PER POCKET	234.36 +- 75.7	4.02 +- 1.21	7.89 +- 2.34	15.23 +- 4.42

The role of data deduplication We evaluated FLOWR across all dataset versions and report the results in Tab. 7. Interestingly, while overall model performance remains comparable across versions, we observe significant differences in strain energy. We hypothesize that this variation arises from the progressive reduction in available training data in the deduplicated versions. Specifically, the default dataset contains 35,000 complexes, while the RMSD-deduplicated dataset includes 25,000, and the RMSD-sequence-identity-deduplicated version is further reduced to 20,000. These findings highlight the importance of dataset size in improving model performance and reliability. While larger datasets contribute to enhanced overall learning, key aspects such as interaction recovery appear to be well captured even in a low-data regime, provided that dataset quality remains high. This suggests that while increasing dataset size is crucial for further advancements, careful dataset curation can enable models to effectively learn critical molecular interactions even with limited data.

Table 7: We investigate the role of data deduplication. We trained models on three different levels of deduplication: no deduplication, RMSD-based deduplication and RMSD- and sequence ID-based deduplication.

METRIC	FLOWR	FLOWR$^{\text{RMSD}}$	FLOWR$^{\text{RMSD-SEQID}}$
RDKIT-VALIDITY	0.94 ± 0.24	0.93 ± 0.25	0.94 ± 0.24
PB-VALIDITY	0.86 ± 0.21	0.87 ± 0.22	0.86 ± 0.22
PLIF RECOVERY	0.44 ± 0.10	0.44 ± 0.15	0.43 ± 0.14
STRAIN ENERGY	64.61 ± 69.50	84.12 ± 49.69	90.91 ± 53.55
VINA SCORE	-6.36 ± 0.85	-6.49 ± 1.16	-6.37 ± 1.10
VINA SCORE (MINIMIZED)	-6.80 ± 0.82	-6.89 ± 1.11	-6.82 ± 1.07
ANGLESW1	1.34 ± 0.9	1.32 ± 0.8	1.39 ± 0.9
BONDLENGTHSW1 [10^{-2}]	0.4 ± 0.01	0.4 ± 0.01	0.3 ± 0.01
DISTANCE TO NATIVE CENTROID	1.00 ± 0.30	0.99 ± 0.41	1.00 ± 0.40
NOVELTY	0.94 ± 0.23	0.95 ± 0.22	0.95 ± 0.22
UNIQUENESS2D	0.94 ± 0.13	0.94 ± 0.13	0.94 ± 0.12
UNIQUENESS3D	0.50 ± 0.20	0.57 ± 0.20	0.52 ± 0.18
DIVERSITY2D	0.86 ± 0.05	0.86 ± 0.06	0.86 ± 0.05
DIVERSITY3D	0.21 ± 0.12	0.23 ± 0.14	0.24 ± 0.17
SA	0.67 ± 0.13	0.68 ± 0.13	0.68 ± 0.13
QED	0.52 ± 0.21	0.53 ± 0.21	0.53 ± 0.21
RINGS	2.68 ± 1.35	2.75 ± 1.36	2.73 ± 1.34
AROMATIC RINGS	1.52 ± 1.16	1.61 ± 1.18	1.60 ± 1.16
HACCEPTORS	6.67 ± 4.23	6.42 ± 4.02	6.68 ± 4.19
HDONORS	2.52 ± 1.68	2.48 ± 1.64	2.48 ± 1.65
LOGP	0.29 ± 3.31	0.56 ± 3.19	0.43 ± 3.23
MOLWT	350.10 ± 114.00	348.87 ± 113.46	350.06 ± 114.40
LIPINSKI	4.35 ± 1.05	4.40 ± 0.97	4.36 ± 1.02

D.3 SAMPLING WITH FIXED MOLECULE SIZE

While sampling ligand sizes is a common practice in benchmarking SBDD models, it is also valuable to assess model performance when ligand sizes are fixed rather than sampled. In this setting, the model is tasked with generating ligands that match the size of the native ligand, enabling a more direct comparison with the test set distribution. A model that effectively captures the underlying data distribution should, on average, exhibit greater overlap with the test set. In Tab. 8, we compare PILOT and FLOWR, each trained with and without explicit hydrogens, under this fixed-size condition. The results indicate that FLOWR consistently outperforms PILOT in terms of RDKit-validity and PoseBusters-validity, while also achieving a higher overlap with the test set distribution. Notably, FLOWR improves the interaction recovery rate from 44In the case of sampled ligand sizes, the generated molecules tend to be smaller on average than their native counterparts, as the sampling procedure allows for a broader range of reductions in size. Conversely, under the fixed-size condition, models are required to generate larger molecules on average, which may impact performance. Interestingly, while PILOT exhibits a significant drop in validity ($\sim 6\%$), FLOWR maintains comparable results to its sampled-size setting, reinforcing the conclusion that FLOWR is both more effective and more stable at inference.

Table 8: We compare PILOT with FLOWR when the molecule size across targets is set to the native ligand size. We sample 100 ligands per target as before and evaluate on the SPIRE test set.

METRIC	$\text{PILOT}_{\text{FIX}}$	$\text{PILOT}_{\text{FIX}}^{\text{WITH-Hs}}$	$\text{FLOWR}_{\text{FIX}}$	$\text{FLOWR}_{\text{FIX}}^{\text{WITH-Hs}}$
RDKIT-VALIDITY	0.83 ± 0.37	0.57 ± 0.42	0.93 ± 0.26	0.64 ± 0.48
PB-VALIDITY	0.74 ± 0.21	0.08 ± 0.06	0.83 ± 0.21	0.60 ± 0.16
PLIF RECOVERY	0.44 ± 0.10	0.28 ± 0.12	0.47 ± 0.11	0.42 ± 0.09
STRAIN ENERGY	92.75 ± 54.73	62.05 ± 20.33	82.35 ± 65.12	54.11 ± 33.36
VINA SCORE	-6.33 ± 0.77	-4.58 ± 0.99	-6.65 ± 0.95	-6.48 ± 0.87
VINA SCORE (MINIMIZED)	-6.74 ± 0.78	-5.17 ± 0.67	-7.12 ± 0.92	-6.86 ± 0.87
ANGLESW1	1.82	2.00	1.39	0.82
BONDLENGTHSW1	0.8 ± 0.02	0.2 ± 0.02	0.5 ± 0.02	0.1 ± 0.01
DISTANCE TO NATIVE CENTROID	0.99 ± 0.28	1.01 ± 0.31	0.95 ± 0.28	0.98 ± 0.28
NOVELTY	0.99 ± 0.10	1.00 ± 0.01	0.92 ± 0.27	1.00 ± 0.00
UNIQUENESS2D	0.98 ± 0.08	0.99 ± 0.02	0.89 ± 0.22	0.97 ± 0.07
UNIQUENESS3D	0.45 ± 0.06	0.74 ± 0.22	0.39 ± 0.25	0.55 ± 0.17
DIVERSITY2D	0.89 ± 0.03	0.90 ± 0.02	0.84 ± 0.07	0.87 ± 0.06
DIVERSITY3D	0.24 ± 0.07	0.29 ± 0.21	0.17 ± 0.13	0.18 ± 0.11
SA	0.61 ± 0.12	0.36 ± 0.11	0.66 ± 0.13	0.27 ± 0.10
QED	0.49 ± 0.21	0.52 ± 0.18	0.50 ± 0.21	0.53 ± 0.21
RINGS	2.88 ± 1.46	1.56 ± 1.06	3.01 ± 1.33	2.64 ± 1.43
AROMATIC RINGS	1.21 ± 1.12	1.11 ± 0.99	1.66 ± 1.19	1.59 ± 1.22
HACCEPTORS	6.78 ± 3.56	5.49 ± 2.09	7.19 ± 4.55	6.47 ± 3.64
HDONORS	2.69 ± 1.70	1.98 ± 1.54	2.65 ± 1.75	2.66 ± 1.58
LOGP	0.43 ± 3.23	-0.17 ± 2.42	0.31 ± 3.52	0.34 ± 2.99
MOLWT	366.58 ± 111.40	351.58 ± 76.31	378.68 ± 117.28	336.09 ± 108.60
LIPINSKI	4.35 ± 1.05	4.67 ± 0.61	4.24 ± 1.14	4.32 ± 1.05