Extended Supplementary Information PILOT: Equivariant diffusion for pocket conditioned de novo ligand generation with multi-objective guidance via importance sampling

Julian Cremer[1,3,*], Tuan Le[1,2,*], Frank Noé[2, 4], Djork-Arné Clevert[1], and Kristof T. Schütt[1]

[1]Machine Learning & Computational Sciences, Pfizer Worldwide R&D, Berlin, Germany
[2]Department of Mathematics and Computer Science, Freie Universität Berlin, Germany
[3]Computational Science Laboratory, Universitat Pompeu Fabra, PRBB, Spain
[4]Microsoft Research AI4Science, Microsoft, Berlin, Germany
[*]Corresponding author: [julian.cremer, tuan.le]@pfizer.com

A. Learning curves: From scratch vs fine-tuned

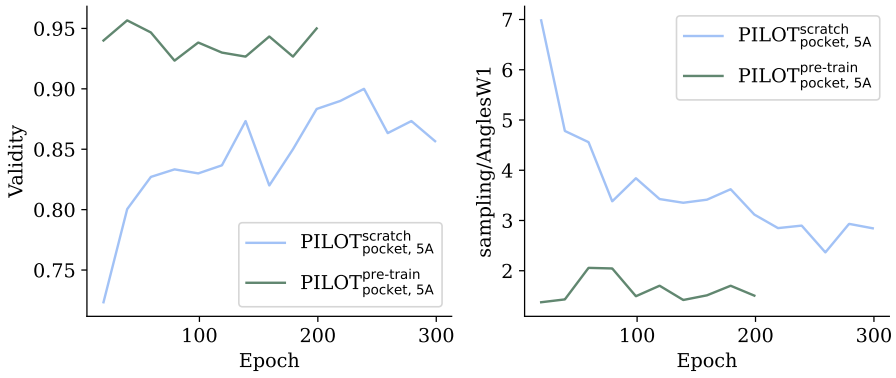

Figure A1: Learning curves comparison for the from scratch trained model against fine-tuned model. For better visibility, we fine-tuned the pre-trained model for 200 epochs to obtain more metrics for visualisation. We show the molecule validity as well as samples/AnglesW1 metric and observe that the fine-tuned model achieves much better metrics already after 20 epochs of training.

We observe that the pre-trained PILOT model achieves faster training convergence compared to the model that is trained from scratch on the CrossDocked dataset. In Figure A1 we show the evaluation curves for both models trained on the 5A PL-complex dataset, i.e., comparing $\text{PILOT}^{\text{scratch}}_{\text{pocket, 5A}}$ against $\text{PILOT}^{\text{pre-train}}_{\text{pocket, 5A}}$. As shown, the pre-trained model

achieves superior metrics compared to the model trained from scratch already in the first evaluation period after 20 epochs of training. Specifically, the molecule validity for the fine-tuned model accomplishes a highest value of 95.67% after 40 epochs while the model trained from scratch only achieves a maximum molecule validity of 90.00% after 240 epochs of training. As the pre-trained model has learned on a vast chemical space from the Enamine Real Diversity, this model achieved to learn general chemistry rules related to valency. Nonetheless, when trained on CrossDocked an extensive distribution shift is expected, since in this scenario, the model inputs a much larger sample in form of a protein-ligand complex. Learning correct geometries how a ligand might fit into a protein pocket is a non-trivial task. Although both models are not explicitly trained to generate a ligand that fit a pocket in a physical sense, like a docking tool, the ligands generated by the fine-tuned model achieves predominant angles distribution metrics compared to the from scratch model. This means that the ligands sampled from the fine-tuned model attain better geometries resembles the angles statistics present in protein-ligand-complexes in the validation set.

B. CrossDocked2020: Analysis

B.1 Pocket size distribution

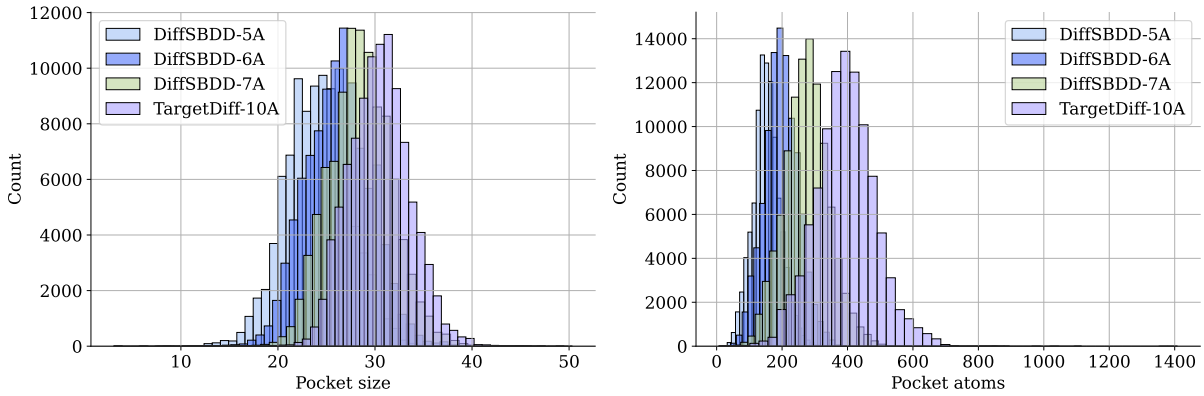

Figure B1: The size distribution for protein pockets in the CrossDocked2020 dataset based on the cutoff radius to create the ligand-pocket complex. The namings DiffSBDD or Targetdiff folowed by xA describe the method to determine the protein pocket with cutoff x. Left: size distribution. Right: distribution of pocket atoms.

To create the Protein-Ligand (PL) complex, the cutoff radius determines which protein atoms should be included next to all ligand atoms, to build the PL complex.

The work by[1] in DiffSBDD creates the PL-complex by computing for each atom in each residue in the protein all pairwise distances to the ligand atoms. As long as one distance from the residues' atom is below the defined cutoff to any ligand atom, the entire residue is included into the protein pocket. Hence choosing such selection through the minimum function potentially creates larger PL complexes, but ensures that all interactions between

Dataset	Mean	Standard Deviation	Median	Skewness
DiffSBDD-5A	24.16	3.35	24.20	0.02
DiffSBDD-6A	26.22	3.33	26.26	0.08
DiffSBDD-7A	28.47	3.25	28.40	0.13
TargetDiff-10A	30.49	3.10	30.47	0.17

Table B1: Statistics of pocket size for different datasets.

Dataset	Mean	Standard Deviation	Median	Skewness
DiffSBDD-5A	152.73	42.04	151.00	0.26
DiffSBDD-6A	198.62	51.90	197.00	0.17
DiffSBDD-7A	274.94	68.32	274.00	0.19
TargetDiff-10A	393.83	90.70	394.00	0.15

Table B2: Statistics of number of pocket atoms for different datasets.

protein-atoms to the ligands are considered.

The PL complex creation in TargetDiff[2] chooses a query point from each residue through the center of mass. Based on this query point the distance to all ligand atoms are computed and the residue with its atoms are included into the PL complex, if any distance between the query point to any ligand atom is below the cutoff. Note that by computing the CoM in the first place, an initial reduction has already been done which can lead so fewer interactions and hence smaller PL complexes. The authors of TargetDiff estimate the pocket size by computing the top 10 farthest pairwise distances of protein atoms. Based on that, they select the median of that as the pocket size for robustness. As the cutoff increases, we observe that the protein pocket also increases as shown in the both panels of Figure B1 for pocket size as well as the number of atoms in the protein pocket. The TargetDiff-10A PL-dataset is particularly large with protein pockets having mean size of 393 atoms. Having larger PL complexes might impede the optimization of the diffusion models since smaller batch sizes and backbones with less trainable parameters are only feasible to fit on a single GPU. Additionally, a smaller cutoff also enables faster training since less message passing steps are required. We believe that a cutoff of 7A is a good trade-off, in that it enables the diffusion model to propagate distant information but also does not fall into the risk of overfitting on a potentially smaller PL complex. The latter is particularly important in the setting of generalization when the diffusion model generates ligands on a new protein target.

B.2 Metrics dependency on ligand size

In this section we show how several metrics are dependent on the ligand size. To this end, we ran six additional sample experiments with $\text{EQGAT}_{\text{diff}}^{\text{pocket, 6A}}$ trained on CrossDocked2020, where ligands are generated without any guidance, i.e., unconditionally with only the protein pocket as context. Figure B2 shows the results. In particular, as de-

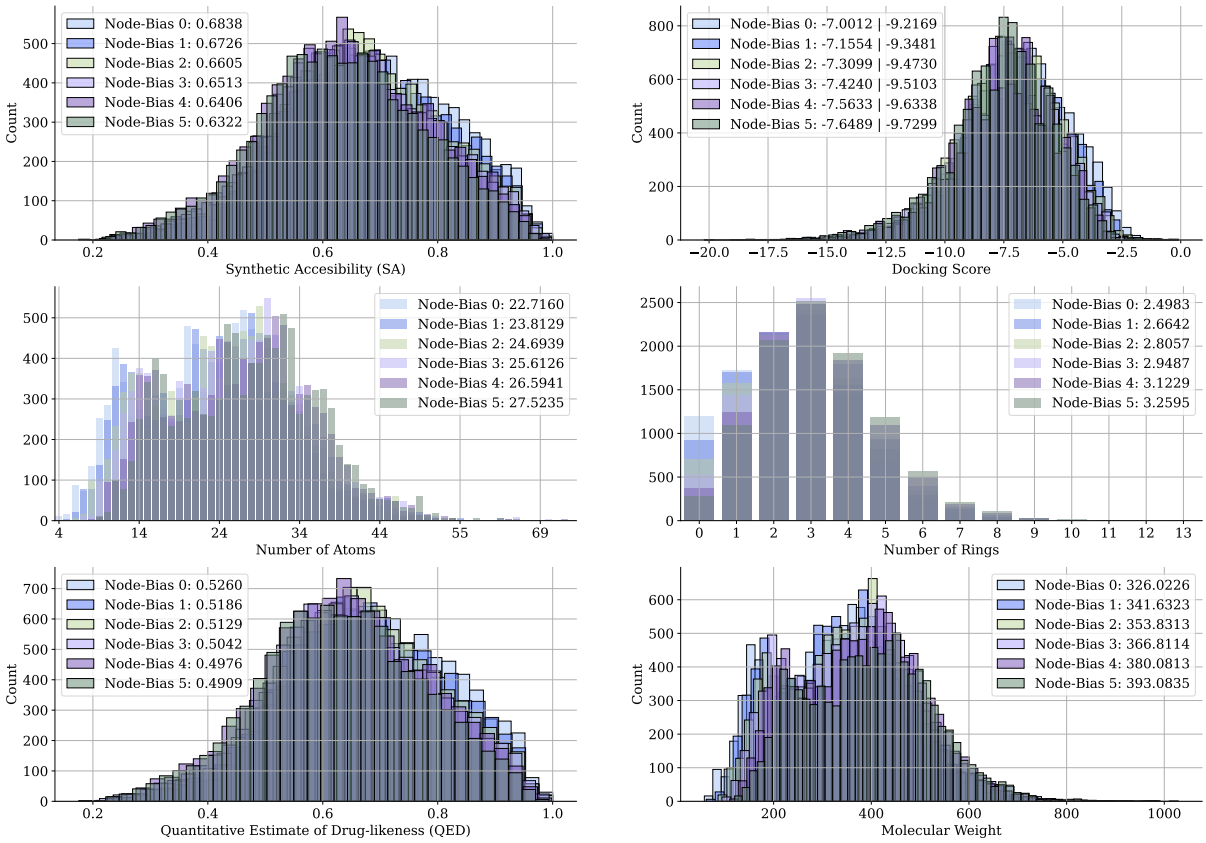

Figure B2: Evaluation metrics for 10,000 ligands where the number of atoms was first sampled from the training set prior and additionally a node bias of n added. Here $n = \{0, 1, 2, 3, 4, 5\}$. The caption shows the mean value across each of the 5 different sets per metric. For the docking scores panel, the first value describes the mean value among all targets, while the second value refers to the top-10% mean value among all targets.

scribed in the main text, larger ligands with more atoms, shown through increasing node bias n tend to have lower synthetic accessibility (SA) score as well as QED and a better (smaller) docking score. Therefore, it is important to take the ligand size distribution into account when evaluating generative models based on docking scores, since the later is negatively correlated with ligand size.

B.3 Ring distribution

To delve deeper into our analysis, we also examine the distribution of ring structures, a known challenge for 3D-based models[3]. The top panel in Fig. B3 illustrates the occurrence of fused and uncommon rings for all models. We observe that TargetDiff, as well as our models, tend to generate more uncommon rings compared to the train and test sets. However, both the SA- and SA-docking-conditional models effectively mitigate this issue by reducing the number of uncommon rings and aligning more closely with the distribution observed in the training and test data.

Consistent with our earlier discussion, the docking-conditional model exhibits a strong propensity for generating numerous rings, including fused and uncommon ones. As depicted in the lower panel of Fig. B3, all models also tend to produce rings that are less

common in drug-like molecules, such as three-, four-, seven-membered, or larger rings. These ring structures are often associated with poor synthetic accessibility, chemical stability, toxicity, or metabolic instability[4–6].

In contrast, five- and six-membered heterocycles containing one or more heteroatoms are considered the gold standard in drug-like molecules[4,6,7], and we observe that these are well represented in the sample space following the training distribution.

Notably, the SA-conditional model effectively regulates the formation of unfavorable ring systems, particularly three- and seven-membered rings. Conversely, the SA-docking-conditional model strikes a reasonable balance, with only a slight increase in seven-membered rings compared to the docking-conditional model, where such rings are more prevalent.

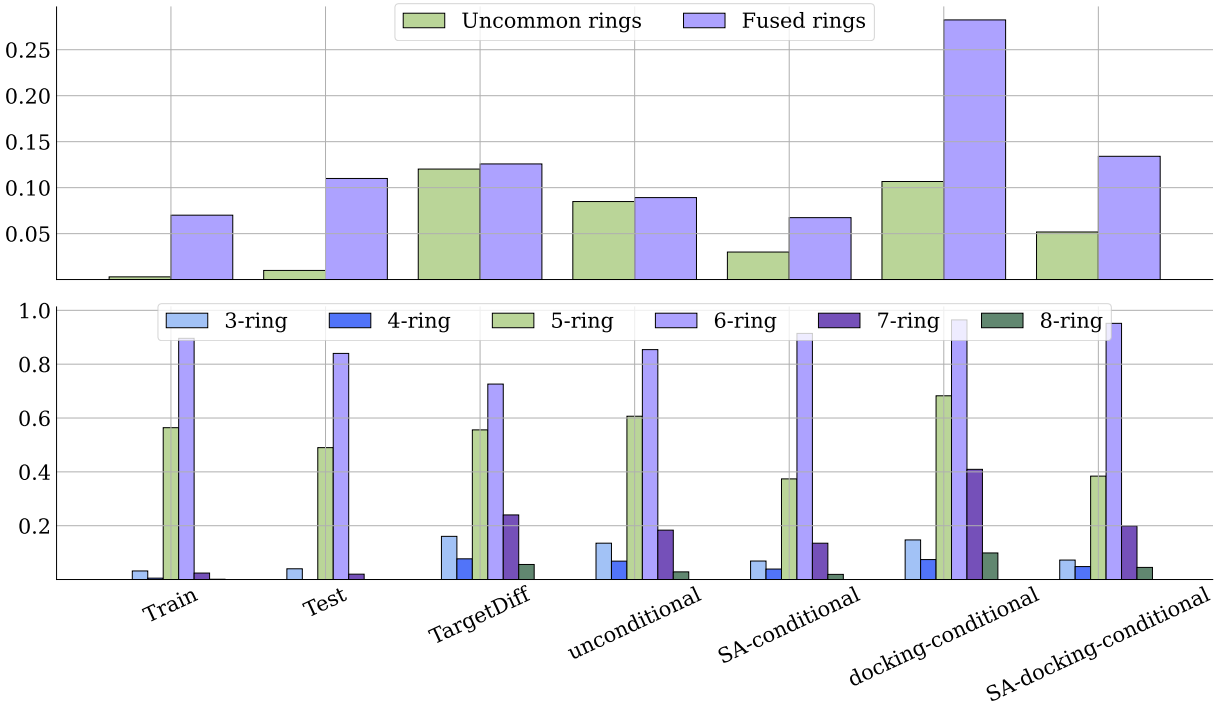

Figure B3: Evaluation of the ring systems in all sampled ligands across test targets. **Top**: Histogram detailing the percentage of uncommon and fused rings for all ligands. **Bottom**: Histogram displaying the distribution of ring sizes from three- to eight-membered rings. Five- and six-membered rings are considered the most drug-like.

C. Hyperparameters for Importance Sampling

In Algorithm 1, we present the property-guided sampling algorithm, which we will further discuss in this section. To perform SA- and docking-score optimization on CrossDocked, we first optimize the SA-score in the population for several iterations. After this initial optimization, we then optimize for docking-score, using a population that has been filtered or biased based on the SA-score guidance. In Table C1, we report the start and end points, as well as the temperature parameter for the unbounded SA-score maximization and unbounded docking-score minimization. Note that the reverse diffusion trajectory

sampling includes $T = 500$ timesteps. We set the population size to 40. This means that for a given protein target P, each batch consists of 40 ligands optimized for high SA-scores and low docking scores. We continue the sampling process until 100 valid ligands are generated.

Property	Start	End	N	Temperature τ
SA-score	0	200	10	0.1
Docking-score	10	250	10	0.1

Table C1: Settings for importance sampling to optimize SA- and docking scores. Start and end columns determine when the importance sampling is performed in the iteration ranging from 1 to 500.

In our experiments, we tried optimizing both SA- and docking score simultaneously by either pointwise adding or multiplying the importance weights $\{w_{k,\text{SA}}\}_{k=1}^{K}$ with $\{w_{k,\text{dock}}\}_{k=1}^{K}$ for each intermediate ligand k for varying time intervals including $(1, 200)$, $(100, 300)$ and $(300, 500)$. We did not see satisfying results where both criteria are optimized in the final ligands, which might be possible to achieve through different hyperparameters including temperature annealing. For this reason, we choose to optimize each property in consecutive order, where the SA-guidance shall act as an initial filter to discard synthetic infeasible (noisy) ligands between steps $(100, 250)$, while the filtered noisy ligands are then guided to minimize docking score in the interval $(300, 400)$. Notice that we do not employ the guidance in iteration in the intervals, but every 10, such that each SA- and docking-guidance include 15 and 10 guidance steps respectively. For sampling, we leveraged an Nvidia A100 with 40GB GPU memory.

C.1 Comparison to Gradient Guidance

We list the metrics for joint optimization fo SA- and docking score in Table C2. For a protein target, we do not sample the number of atoms for generated ligands, but fix the number to the reference ligand to have a fair comparison between importance sampling and classifier guidance.

Table C2: Comparison of importance sampling (IS) vs. classifier guidance (CG) on CrossDocked testset. We generated 100 ligands for each of the 100 pockets in the test set. We report mean time per pocket in minutes, validity, uniqueness and the corresponding property values.

Method	Time [min]	Validity [%]	Uniqueness [%]	SA				Docking			
				Steps	Every-N	λ, τ	Score	Steps	Every-N	λ, τ	Score
IS	3.51	92.29	75.55	$0-200$	10	0.1	0.7531	$150-250$	10	0.1	-7.679
CG	15.02	77.17	64.97	$0-500$	1	0.1	0.8248	$0-500$	1	0.1	-8.4397
CG	3.71	93.18	83.22	$0-200$	10	0.1	0.7205	$150-250$	10	0.1	-7.1523
IS & CG	3.72	92.54	58.79	$0-200$	10	0.1	0.7608	$150-250$	10	0.1	-8.0151
Unconditional	3.01	93.15	83.63	–	–	–	0.711	–	–	–	-7.0248

We experienced GPU memory issues when performing classifier (gradient) guidance because backpropagation is costly, especially for larger protein-ligand complexes, disabling

us to use the default batch size of 40 on an A100 40GB GPU . Reducing the batch size is necessary to avoid out of memory errors, with the disadvantage of increasing the overall running time. We set the batch size to 25 to perform both importance sampling as well as classifier guidance with the same settings on an Nvidia H100 with 80GB GPU memory. For importance sampling, we can easily set a batch size of 40 or even 50 ligands per pocket, achieving an average pocket time of about 3.28 minutes to generate 100 ligands, showing the advantage of importance sampling compared to classifier guidance, to allow for larger ligand batches during generation, effectively reducing runtime.

Since our proposed importance sampling filters for trajectories that either maximize or minimize a property, gradient guidance can also be performed afterwards. To enable both approaches (IS & GC), we choose the same filtering steps to maintain a low computational budget and perform backpropagation to obtain the gradients with respect to atomic coordinates to shift the positions in space after promising candidates were selected based on their importance weights. This results in a generated ligand set with a similar validity of 92.54% but a reduced uniqueness rate of 0.58. The mean SA score reaches 0.76, while the docking score reaches -8.01 as shown in the second last row in Table C2.

D. Kinodata-3D: Analysis

D.1 Correlation

	pIC50	# Rings	# Rotatable bonds	# Atoms	QED	SA
# Rings	0.19					
# Rotatable bonds	0.16	0.15				
# Atoms	0.28	0.66	0.68			
QED	-0.18	-0.53	-0.62	-0.79		
SA	-0.24	-0.37	-0.20	-0.39	0.18	
logP	0.04	0.22	0.13	0.30	-0.39	0.38

Figure D1: Correlation matrix of pIC50s, number of rings, number of atoms, QEDs, and SAs on the Kinodata-3D training set.

We show the correlation matrix on the Kinodata-3D dataset in Figure D1. Similar to the CrossDocked dataset, we observe that metrics like the QED and SA score are negatively correlated with the number of atoms and rings. As opposed to CrossDocked, in the Kinodata-3D a negative correlation of -0.39 between logP and QED is observed while in CrossDocked the correlation amounts to 0.36. One possible explanation for this is that Kinodata-3D consists of experimental kinase-ligand assay data and therefore only considers ligands that covers a smaller chemical space where the logP covers a potentially smaller range.

In Figure D3 we show the ring distributions on the Kinodata-3D dataset next to the distribution that our EQGAT-Diff-Pocket models produced in 4 settings, namely

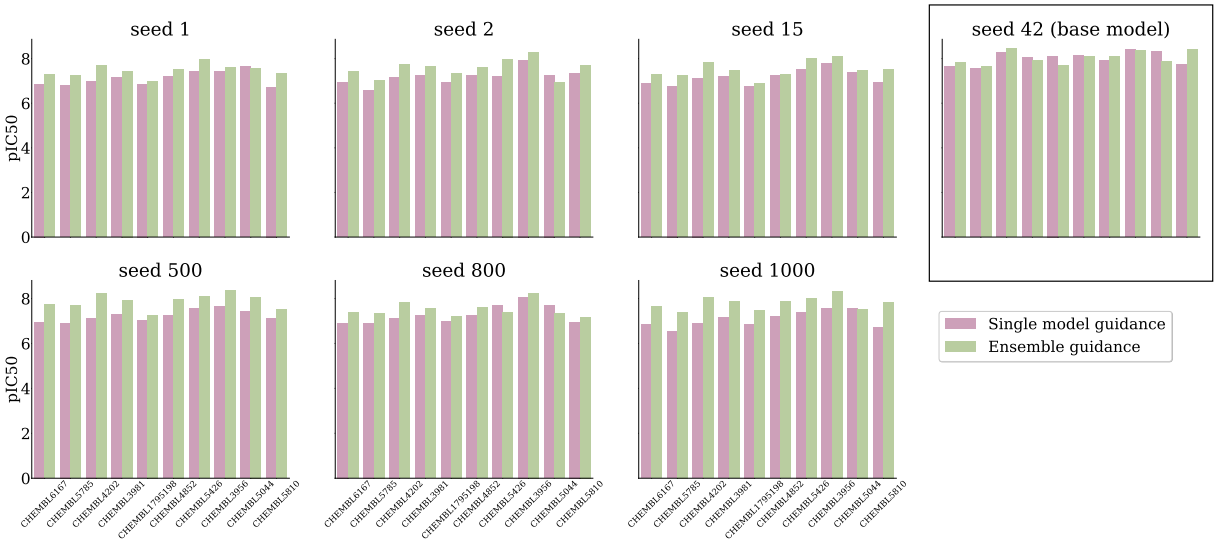

Figure D2: Single model guidance is compared to ensemble guidance. A base model, here the model with seed 42, is used to sample 100 ligands per target. In the case of single model guidance (seed 42 model guides itself), a variety of models trained with different seeds evaluate the sampled ligands with respect to pIC_{50} values. In the case of ensemble guidance, an ensemble of models, here seed 42, 500, and seed 1000, are used for pIC_{50} guidance. Again, all seed models evaluate the sampled ligands with respect to pIC_{50}. We can see that throughout targets and seeds, the ensemble guidance not only works best in terms of pIC_{50} but also in terms of stability and generality. The base model assigns similar pIC_{50} values to its samples for both, single model and ensemble guidance. Nevertheless, across seed models (involved in ensemble guidance or not) the samples taken from the single model guidance exhibit a significantly worse pIC_{50} prediction in contrast to ensemble guidance. Here, all seed models predict similarly high pIC_{50} values suggesting that ensemble guidance leads to a more stable and most importantly more general set of samples.

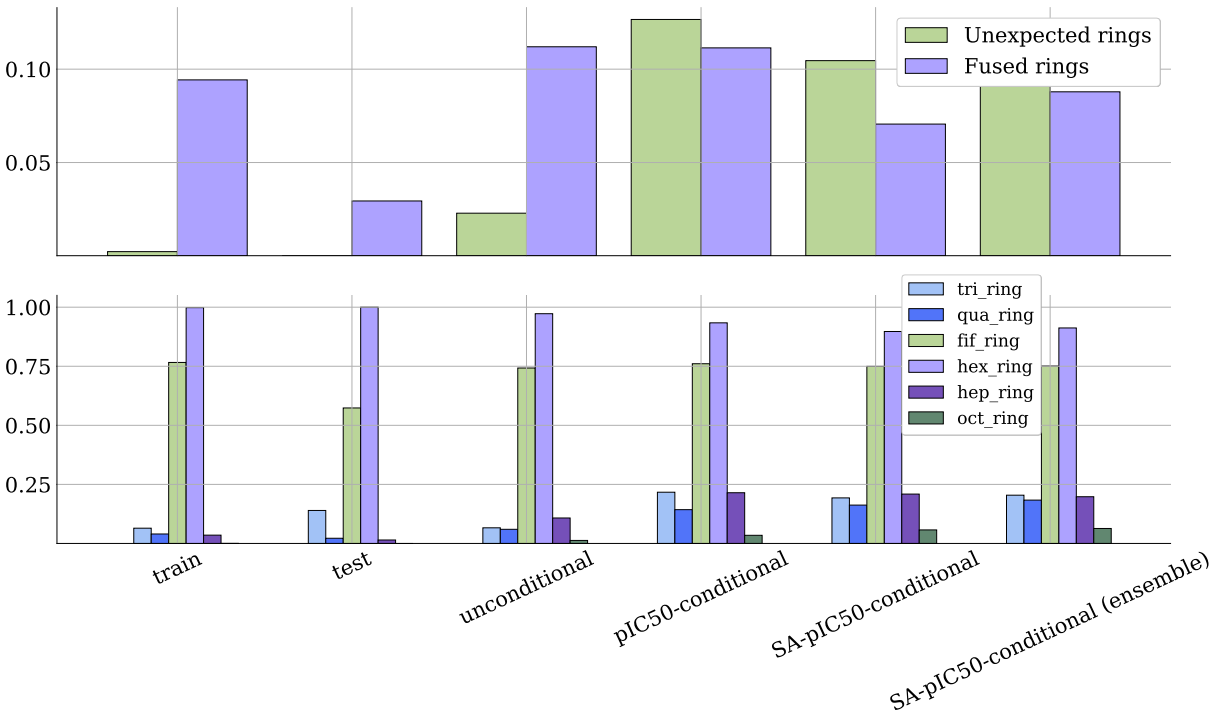

Figure D3: Ring distribution on the Kinodata-3D dataset

the unconditional, single pIC_{50}-conditional, joint SA-pIC_{50}-conditinoal as joint SA-pIC_{50}-conditional ensemble. Similar to CrossDocked, an unconditional model that only generates ligands based on a protein (kinase) pocket as context, the number of fused rings increases, leading to unfavourable molecules that might be difficult to synthesize. When leveraging the importance sampling in the pIC_{50}-conditional generation, we observe that the number of fused rings stays the same but unexpected rings increases. Also, an increase in 4, 7- and 8-membered rings is observed, when we aim in maximizing the pIC_{50} to search for ligands that exhibit high (predicted) binding affinity. One potential reason why that happens might lie in the dataset the pIC_{50} expert model was trained on. As shown in Figure D1 a negative correlation between pIC_{50} and QED as well as SA is observed. Thus, with higher pIC_{50}, ligands tend to become less synthesizable and drug-like according to those metrics. Fortunately, we can leverage our framework and jointly optimize for high pIC_{50} as well as SAscore, which is done in the SA-pIC_{50}-conditional samples. By leveraging expert models for both properties, we can reduce the number of unexpected rings as well as fused rings, although the 4,7- and 8-membered rings are still prevalent but potentially in ligands that according to the SAScore are still better evaluated compared to the pIC_{50}-conditional setting only.

References

[1] A. Schneuing, Y. Du, C. Harris, A. Jamasb, I. Igashov, W. Du, T. Blundell, P. Lió, C. Gomes, M. Welling, M. Bronstein and B. Correia, *Structure-based Drug Design with Equivariant Diffusion Models*, 2023, https://arxiv.org/abs/2210.13695.

[2] J. Guan, W. W. Qian, X. Peng, Y. Su, J. Peng and J. Ma, The Eleventh International Conference on Learning Representations, 2023.

[3] Y. Xia, K. Wu, P. Deng, R. Liu, Y. Zhang, H. Guo, Y. Cui, Q. Pei, L. Wu, S. Xie, S. Chen, X. Lu, S. Hu, J. Wu, C.-K. Chan, S. Chen, L. Zhou, N. Yu, H. Liu, J. Guo, T. Qin and T.-Y. Liu, *Target-aware Molecule Generation for Drug Design Using a Chemical Language Model*, 2024, https://www.biorxiv.org/content/early/2024/01/08/2024.01.08.574635.

[4] R. D. Taylor, M. MacCoss and A. D. G. Lawson, *Journal of Medicinal Chemistry*, 2014, **57**, 5845–5859.

[5] X.-C. Yu, C.-C. Zhang, L.-T. Wang, J.-Z. Li, T. Li and W.-T. Wei, *Organic Chemistry Frontiers*, 2022, **9**, 4757–4781.

[6] A. Rusu, I.-M. Moga, L. Uncu and G. Hancu, *Pharmaceutics*, 2023, **15**, 2.

[7] J. Jampilek, *Molecules*, 2019, **24**, 6.