PRISM: A HYBRID DIFFUSION-REINFORCEMENT LEARNING FRAMEWORK FOR 3D STRUCTURE-BASED *De Novo* DESIGN

Sanaz Kazeminia[1,*], **Lewis Vidler**[2], **Pushkar G. Ghanekar**[3], **Nele Quast**[1] **Garrett M. Morris**[1]

[1]Department of Statistics, University of Oxford, Oxford, UK
[2]Eli Lilly & Company, London, UK
[3]Eli Lilly & Company, Indiana, United States

{sanaz.kazeminia, nele.quast, garrett.morris}@stats.ox.ac.uk
{lewis.vidler, pushkar.g}@lilly.com

ABSTRACT

Structure-based diffusion models offer a promising route for *de novo* 3D ligand generation directly within protein binding sites, but generating stereochemically valid molecules within acceptable molecular property constraints required for drug design remains challenging. Existing approaches typically rely on inference-time guidance, limiting flexibility and preventing medicinal chemists from directly specifying design objectives. We introduce PRISM: **P**ocket **R**einforced **I**terative **S**tructure-based **M**olecular diffusion, a reinforcement learning framework for fine-tuning structure-based diffusion models using Proximal Policy Optimization (PPO). PRISM enables user-defined rewards to be incorporated directly into the generative process. We evaluate PRISM across six well-studied drug targets and systematically study single-objective, multi-objective, and curriculum-based optimization strategies. PRISM consistently improves 3D geometric validity, demonstrating that reinforcement learning can effectively shape diffusion models in continuous coordinate space. Extending to multi-objective rewards highlights how reward design and reward density influence optimization, while a staged curriculum anchored in geometric validity stabilizes training and supports the integration of more complex medicinal chemistry objectives. PRISM is lightweight and practical, requiring only a single GPU and a few hours of training to fine-tune a model toward a desired reward. Together, these results establish reinforcement learning as a flexible and accessible tool for optimising structure-based molecular diffusion, enabling rapid experimentation with custom reward functions for 3D molecular design.

1 INTRODUCTION

Structure-based drug design (SBDD) uses the 3-dimensional structure of a biological target to design compounds that bind with high potency and meet desired pharmacological and molecular property characteristics (Thomas et al., 2017; van Montfort et al., 2017). This poses an inverse design problem: given target properties (for example, binding affinity, drug-likeness, synthetic accessibility (SA), novelty) we must identify molecules that satisfy them - a challenge central to accelerating drug discovery and development which is tedious, slow and expensive (Sadybekov & Katritch, 2023; Sun et al., 2022). Furthermore, we barely explore the vastness of chemical space, which has been estimated to be between 10^{20}-10^{60} (Polishchuk et al., 2013). To address these challenges, deep learning (DL) based generative models for SBDD have garnered heavy interest as powerful alternatives to traditional approaches that rely on the intuition of medicinal chemists (Fu & Chen, 2025; Sadybekov & Katritch, 2023).

Generative diffusion models have emerged as a promising approach for SBDD and are current gold standard for 3D generative design alongside flow matching models (Vost et al., 2025). However,

models frequently produce physically implausible geometries with strained conformations and poor physicochemical properties dissimilar to those of approved drugs (Yang et al., 2025; Harris et al., 2023; Buttenschoen et al., 2024). Most critically, generated molecules also lack meaningful protein-ligand interactions and exhibit problematic functional groups that preclude experimental validation (Harris et al., 2023). Thus, SBDD is a multi-objective optimisation problem which must account for 3D geometric validity, molecular properties alongside potency (Yang et al., 2025; Sanjrani et al., 2025).

Several approaches attempt to steer SBDD diffusion models toward desired properties. Inference-time gradient-based guidance methods, such as TAGMOL (Dorna et al., 2024) apply property gradients during sampling to redirect generation. However, these require training separate regressors for each objective and can produce unstable sampling dynamics when combined. Moreover, because regressor guidance relies on continuous gradients, it struggles to enforce discrete or combinatorial design constraints, such as pharmacophore satisfaction, where gradient signals are sparse or poorly aligned with the underlying generative distribution. IDOLpro (Kadan et al., 2025) guides diffusion latent variables at inference time using differentiable scoring functions to optimise binding affinity and SA. However, this requires scoring functions to be differentiable and re-implemented in PyTorch, limiting flexibility to incorporate arbitrary reward signals without significant engineering effort. In contrast, train-time methods such as Direct Preference Optimization (DPO) (Rafailov et al., 2023) fine-tune model parameters on molecular preference pairs (Schneuing et al., 2025; Cheng et al., 2024). DPO learns by comparing molecule pairs (A is better than B), but drug design objectives rarely permit such absolute rankings - molecular quality is multi-dimensional, and whether molecule A outperforms B depends on context-specific priorities across competing properties. Collectively, these approaches highlight a persistent gap, motivating a framework capable of easily incorporating flexible reward signals and naturally balancing competing objectives within structure-based diffusion models.

Reinforcement learning (RL) provides a principled framework for incorporating arbitrary reward signals without the constraints of differentiability or binary preferences. Policy gradient methods like Proximal Policy Optimization (PPO) (Schulman et al., 2017) have demonstrated success in optimizing SMILES-based generators such as REINVENT (Loeffler et al., 2024), yet their application to 3D structure-based diffusion models has remained unexplored.

In this work we introduce PRISM: **P**ocket **R**einforced **I**terative **S**tructure-based **M**olecular diffusion, a hybrid diffusion–RL framework for structure-based *de novo* molecular generation. We systematically evaluate PRISM under single-objective, multi-objective, and curriculum-based optimization strategies. In the single-objective setting, our stereochemical reward function consistently improved 3D geometric validity across six targets, demonstrating that policy gradients can effectively guide diffusion models toward the manifold of physically plausible molecular structures. Extension to multi-objective rewards revealed strong target dependence and exposed potential failure modes including reward hacking, particularly when reward signals were sparse or weakly coupled to 3D structure. We showed that a staged curriculum, anchoring optimization in geometric validity before introducing more complex objectives, enables explicit control over learning priorities and prevents unintended objective prioritization that occurs when all rewards are introduced simultaneously. Our results establish RL as an effective approach for guiding pre-trained 3D structure-based diffusion models. PRISM offers a principled fine-tuning framework for multi-objective optimization that accommodates arbitrary reward signals, enabling targeted control over geometric validity, binding characteristics, and molecular properties in structure-based design.

2 METHODS

We formulate structure-based molecular generation as a sequential decision-making problem. We apply RL to optimize a pre-trained diffusion model toward user-defined objectives. PRISM extends DiffSBDD (Schneuing et al., 2024), an SE(3)-equivariant diffusion model trained on CrossDocked2020 (Francoeur et al., 2020), finetuning it using PPO with flexibly designed reward functions tailored to specific design objectives.

2.1 PRISM FRAMEWORK

Following Black et al. (2023), we formulate the diffusion sampling process as a Markov Decision Process (MDP) over $T = 500$ denoising timesteps. The protein binding site $p = (\mathbf{x}_p, \mathbf{h}_p)$ represents the all-atom pocket structure with 3D coordinates $\mathbf{x}_p$ and atomic element types $\mathbf{h}_p \in \mathcal{A}^{N_p}$, where $\mathcal{A} = \{\text{C, N, O, S, B, Br, Cl, P, I, F}\}$. The MDP components are: (i) **states** $s_t = (z_t, p, t)$ represent the noised molecular structure $z_t = (\mathbf{x}_t, \mathbf{h}_t)$ with 3D coordinates $\mathbf{x}_t \in \mathbb{R}^{N\times 3}$ and atom types $\mathbf{h}_t \in \mathcal{A}^N$, along with the protein binding site p and diffusion timestep $t \in \{0, 1, \ldots, T\}$; (ii) **actions** $a_t = z_{t-1}$ are the denoising predictions; (iii) **policy** $\pi_\theta(z_{t-1}|z_t, t, p)$ is the diffusion model's learned denoising process; and (iv) **rewards** $R(s_0) = r(z_0, p)$ are assigned at the final timestep and propagated back through the trajectory, since only the fully denoised structure at $t = 0$ is chemically meaningful.

2.2 PPO TRAINING

We apply PPO with the standard clipped objective to fine-tune the policy model (Schulman et al., 2017). To preserve the base model's pocket-conditioning capabilities, we freeze three of the five SE(3)-EGNN layers and apply PPO updates only to timesteps $t \in [200, 500]$. This partial fine-tuning ensures the model retains geometric understanding from the DiffSBDD prior while optimizing for target objectives. The PPO loss prevents overly large policy updates through ratio clipping:

$$L^{\text{CLIP}}(\theta) = \mathbb{E}\left[\min\left(\rho_t(\theta)\hat{A}_t, \text{clip}(\rho_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right)\right], \quad \rho_t(\theta) = \frac{\pi_\theta(z_{t-1}|z_t, t, p)}{\pi_{\theta_{\text{old}}}(z_{t-1}|z_t, t, p)} \tag{1}$$

where $\rho_t(\theta)$ is the importance sampling ratio and ϵ constrains updates. At each timestep, the diffusion model defines a Gaussian reverse transition, allowing us to compute exact log-likelihoods of the sampled denoising steps, which are used to construct the PPO importance ratio.

To stabilize training, advantages $\hat{A}_t$ are computed by normalizing per-molecule rewards over each training batch using a running mean and standard deviation, following the Group Relative Policy Optimization (GRPO) approach (Shao et al., 2024; Black et al., 2023). This standardization yields stable gradient signals across iterations without the use of a trained value network which traditionally requires training.

2.3 REWARD FUNCTIONS

We first investigate a single-objective reward function aimed at correcting the geometric inconsistencies exhibited in the baseline model, DiffSBDD. After establishing geometric competence, we introduce pharmacophore and property-based rewards alongside the maintained geometric reward. This curriculum strategy avoids two failure modes: simultaneous training causes the model to prioritize geometric constraints while ignoring pharmacological objectives, whereas adding pharmacophore and property rewards without geometric anchoring produces physically implausible structures that satisfy objectives through distorted conformations.

2.3.1 INTRODUCING GEOMETRIC VALIDITY

To address the physically implausible geometries often produced by 3D generative models (Buttenschoen et al., 2024), we define a stereochemical reward R_{geom} based on a subset of the PoseBusters test suite. This reward penalizes deviations in bond lengths, bond angles, and internal steric clashes:

$$R_{\text{geom}} = \exp\left(-\lambda \sum_i w_i P_i\right) \tag{2}$$

where $i \in \{\text{length}, \text{angle}, \text{clash}\}$, P_i represents the sum of absolute percent deviations from ideal geometry for each violation type, $\lambda = 2.0$ is the penalty scaling factor, and $w_{\text{length}} = 1.0$, $w_{\text{angle}} = 0.5$, $w_{\text{clash}} = 1.5$ weight the contributions to prioritize steric clash and bond length corrections, which are the most frequent failure modes in SBDD models.

2.3.2 MOLECULAR PROPERTIES REWARD

We score each molecule using Gaussian functions centered on target-specific reference distributions extracted from known binders. For each property j (SA score, H-bond donors/acceptors, aliphatic/aromatic ring counts, rotatable bonds, chiral centers, LogP, fused ring count), we compute:

$$s_j = \exp\left(-\frac{(x_j - \mu_j)^2}{2\sigma_j^2}\right) \cdot w_j, \quad R_{\text{prop}} = \frac{\sum_j s_j}{\sum_j w_j} \tag{3}$$

where s_j is the weighted score for property j, x_j is the molecule's measured property value, μ_j and σ_j are the mean and standard deviation from the target-specific reference distribution, to control gradient width, and w_j weights each property's contribution. w_j for each property was defined based on observed tensions between the different properties the baseline model struggles with. For example, a low likelihood of generating aromatic rings versus a higher likelihood of generating randomly placed H-bond donors and acceptors.

2.3.3 PHARMACOPHORE REWARD

We derive consensus pharmacophoric hotspots from known binders by extracting interaction features and clustering them using DBSCAN (Pedregosa et al., 2011). For each feature type k (aromatic, H-bond acceptor/donor, hydrophobic, ionizable), we score generated molecules based on spatial proximity to high-density cluster centers:

$$R_{\text{pharm}} = \frac{\sum_k w_k s_k}{\sum_k w_k} \tag{4}$$

where s_k is the normalized distance score for feature type k computed via optimal feature-to-cluster matching, and w_k weights each pharmacophore feature's importance. Distance-based scoring uses a quadratic decay within a cut-off radius to provide smooth gradients (feature clustering details in Appendix A1.2).

2.3.4 MULTI-OBJECTIVE OPTIMIZATION AND CURRICULUM LEARNING

We first train PRISM using only the geometric reward R_{geom} to establish a foundation of physical plausibility. After achieving baseline geometric competence, we transition to a multi-objective function combining all reward components:

$$R_{\text{multi}} = \sum_i w_i R_i, \quad \sum_i w_i = 1 \tag{5}$$

where $i \in \{\text{geom}, \text{pharm}, \text{prop}\}$, $w_{\text{geom}} = 0.1$, $w_{\text{pharm}} = 0.6$, $w_{\text{prop}} = 0.3$.

This approach mitigates failure modes where the policy exploits R_{pharm} by generating physically implausible structures or molecules with poor drug-like properties. The retained geometric weight ($w_{\text{geom}} = 0.1$) acts as a regularizer, maintaining physical plausibility while the policy optimizes pharmacophore matching with improved molecular properties.

We trained PRISM on six diverse, well-studied targets: the first bromodomain (BD1) of bromodomain-containing protein 4 (BRD4-BD1), estrogen receptor α (ERα), human immunodeficiency virus type-1 protease (HIV-1-PR), carbonic anhydrase II (CA-II), epidermal growth factor receptor (EGFR), and factor Xa (FXa). For each target, we selected three test structures to capture binding site variability: (i) the highest resolution structure, (ii) a conformationally distinct structure identified through TM-align clustering, if present, and (iii) a randomly selected structure to ensure unbiased coverage. For details on target selection and hyperparameters, see Appendix A1.1.

3 RESULTS & DISCUSSION

Below we present results for (i) single-objective geometric optimization (R_{geom}) and (ii) curriculum-based multi-objective optimization (R_{multi}), where R_{multi} is introduced after geometric competence is achieved with R_{geom}.

Table 1: PoseBusters validity rates (PB-valid, %) for PRISM trained with our geometry reward function compared to DiffSBDD across datasets. All improvements are statistically significant at $p < 0.0001$ (chi-squared test). $N = 15,000$ molecules per method, with $5,000$ molecules for each test structure.

Target	**PB-valid (%)**	
	DiffSBDD	**PRISM**
BRD4-BD1	58.0	**73.9**
CA-II	42.6	**73.1**
EGFR	61.9	**76.8**
ERα	46.6	**77.8**
FXa	56.2	**71.6**
HIV-1-PR	59.2	**75.0**
Mean	54.1	**74.7**

3.1 Single Objective Geometry Reward

PRISM improved the generation of geometrically valid 3D molecules using our geometry reward function through RL, achieving statistically significant improvements over baseline DiffSBDD across all six targets ($p < 0.0001$, chi-squared test; Table 1). Stereochemical validity rates as measured by PoseBusters (denoted PB-valid) increased from 55% to 75% on average, with the largest gains observed in CA-II (30% improvement). These improvements demonstrate that RL can effectively guide diffusion trajectories toward improving the recall of high quality molecular structures without relying on post-hoc filtering.

Analysis of the constituent PoseBusters checks on ligands (Appendix Table A3) show that PRISM substantially reduces bond angle violations, steric clashes, and bond length errors. We did observe a slight increase in aromatic ring flatness failures, likely due to a higher propensity for aromatic ring generation with PRISM. However, global molecular property distributions of the prior remain largely unchanged when using geometry as a single reward, demonstrating that finetuning on our geometric reward function does not affect sample diversity or cause unilateral biases (Appendix Figure A3).

3.2 Multi-objective Optimization and Reward Design

Next, building on successful geometric optimization, we investigated whether PRISM could simultaneously maintain geometric validity while optimizing pharmacophoric feature placement and physicochemical properties. We show that PRISM successfully optimizes multi-objective rewards for BRD4-BD1 and CA-II. However, for FXa, HIV-1-PR, and ERα, performance is highly target-dependent: the model prioritizes different objectives based on reward landscape characteristics, with some targets showing preferential optimization of pharmacophoric features over molecular properties, or minimal change when pharmacophore maps are sparse. Results for all targets can be found in Appendix Section A1.5.

Figure 1 shows PRISM's strongest performance on BRD4-BD1 in terms of pharmacophoric feature placement, while CA-II exhibits an increase in maximum pharmacophoric scores, with the mean remaining comparable to DiffSBDD. For molecular property scores, BRD4 shows only marginal improvement over DiffSBDD, whereas CA-II demonstrates substantial gains, bringing the majority of ligand properties into alignment with the reference ligands. Appendix Figure A6 shows an example BRD4 molecule generated by PRISM, compared to DiffSBDD and a reference ligand.

As shown in Figure 1, FXa most notably exhibited dimensional collapse, as the initial reference ligands provided sparse, non-directional pharmacophoric clusters. This is reflected by the low mean achieved by the reference ligands on the same map, and as a result, the model failed to obtain a stable 3D gradient signal. Consequently, the optimisation defaulted to the densest available signal, namely 2D molecular properties. While this maximised the scalar reward, it led to "reward hacking" where the model favoured low-complexity fragments (low MW, high SA score) that satisfy 2D heuristics but lack the structural complexity required for the binding site. In Appendix Figure A5, the same

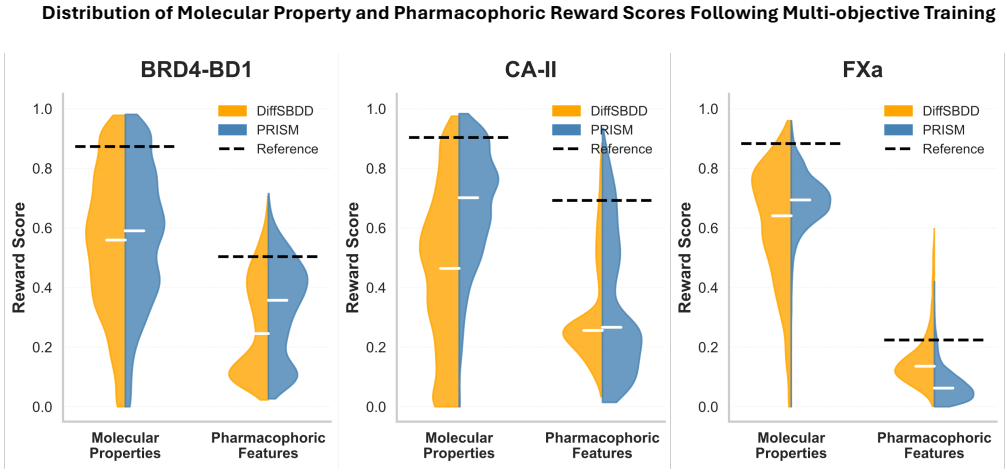

Figure 1: Distribution of molecular property (R_{prop}) and pharmacophore (R_{pharm}) reward scores following multi-objective training. PRISM (blue) and DiffSBDD (orange). Black dashed lines indicate reference means. PRISM achieves higher pharmacophore scores for BRD4-BD1 and CA-II, and consistently outperforms DiffSBDD in molecular property scores across all targets. In contrast, FXa exhibits lower pharmacophore scores following optimization but improved molecular properties. Values are aggregated over $N = 30{,}000$ molecules per method and per test set ($10{,}000$ per test pocket).

Table 2: PoseBusters validity rates (PB-valid, %) following curriculum-based multi-objective training. All improvements are statistically significant at $p < 0.001$ (chi-squared test). $N = 3,000$ molecules per target ($1,000$ per test pocket).

Target	**PB-valid (%)**	
	DiffSBDD	**PRISM**
BRD4-BD1	44.9	**70.0**
CA-II	68.3	**83.4**
EGFR	47.9	**61.3**
ERα	58.4	**83.3**
FXa	39.5	**63.5**
HIV-1-PR	40.4	**64.0**

reward optimisation behaviour can be observed for ERα and, to a lesser extent, HIV-1-PR. Overall, this is a reward design problem and future work for these targets could involve using a specific subset of the pharmacophore map to prevent the policy from receiving confusing signals. Docking scores may also prove somewhat useful or a reliable quantitative structure-activity relationship (QSAR) model as the reward signal to optimise potency and properties rather than pharmacophoric feature hotspots.

While multi-objective optimization exhibits target-dependent behavior in our case studies, PRISM demonstrates that targeted optimization toward specific design objectives can be achieved without sacrificing chemical diversity or novelty. PRISM maintains high novelty ($> 99\%$ across all targets) and internal diversity almost like that of the baseline model to explore chemical space, while simultaneously improving SA scores relative to DiffSBDD (Appendix Table A4). The modest decrease in diversity for some targets reflects policy convergence toward high-reward regions rather than mode collapse.

Curriculum learning maintained geometric validity throughout multi-objective training, though with target-dependent outcomes (Table 2). Some targets improved stereochemical validity (ERα and CA-II to 83%, BRD4-BD1 maintained at 70%), while others showed modest decreases in comparison

to our single objective geometric reward function (HIV-1-PR, FXa, EGFR). We hypothesize this reflects the interplay between reward weights: with pharmacophore scoring weighted at 0.5, versus geometry at 0.1, sparse pharmacophore maps may drive the policy toward feature placement at the expense of geometric constraints. Critically, even targets showing decreased validity maintained physically plausible structures above 60%, demonstrating that the geometric foundation prevents catastrophic collapse despite competing optimization pressures.

4 Conclusion

We introduced PRISM, a hybrid RL-diffusion framework that enables rapid, flexible optimization of 3D structure-based diffusion models through custom reward functions. PRISM accommodates arbitrary, non-differentiable objectives without requiring dedicated regression models or differentiable implementations for each objective. Fine-tuning with PPO achieved substantial improvements in stereochemical validity and alignment with target-specific molecular property distributions while maintaining high chemical diversity and novelty, demonstrating successful exploration-exploitation balance in continuous 3D coordinate space. Our systematic evaluation reveals that sparse 3D signals can lead to dimensional collapse toward more densely rewarded 2D objectives, highlighting reward design as a central challenge for 3D RL-based molecular generation and motivating future work incorporating stronger 3D signals. While demonstrated on DiffSBDD, our approach is model-agnostic and may be applied to other generative structure-based diffusion models. PRISM establishes a practical foundation for computational chemists to rapidly experiment with diverse design objectives, opening new directions for targeted structure-based molecular optimization.

Code and Data Availability

Code for PRISM and instructions to the datasets are publicly available at `https://github.com/SanazKaz/PRISM`.

REFERENCES

RDKit: Open-source cheminformatics. URL https://www.rdkit.org/https://doi.org/10.5281/zenodo.591637.

Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 1 2000. ISSN 0305-1048. doi: 10.1093/NAR/28.1.235. URL https://dx.doi.org/10.1093/nar/28.1.235RCSB.org.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training Diffusion Models with Reinforcement Learning. *12th International Conference on Learning Representations, ICLR 2024*, 5 2023. URL https://arxiv.org/pdf/2305.13301.

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, 2 2024. ISSN 20416539. doi: 10.1039/D3SC04185A. URL https://pubs.rsc.org/en/content/articlehtml/2024/sc/d3sc04185ahttps://pubs.rsc.org/en/content/articlelanding/2024/sc/d3sc04185a.

Xiwei Cheng, Xiangxin Zhou, Yuwei Yang, Yu Bao, Quanquan Gu, and Bytedance Research. Decomposed Direct Preference Optimization for Structure-Based Drug Design. 7 2024. URL https://arxiv.org/pdf/2407.13981.

Peter J.A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J.L. De Hoon. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 6 2009. ISSN 13674803. doi: 10.1093/bioinformatics/btp163. URL https://dx.doi.org/10.1093/bioinformatics/btp163.

Vineeth Dorna, D. Subhalingam, Keshav Kolluru, Shreshth Tuli, Mrityunjay Singh, Saurabh Singal, N. M. Anoop Krishnan, and Sayan Ranu. TAGMol: Target-Aware Gradient-guided Molecule Generation. 6 2024. URL https://arxiv.org/pdf/2406.01650.

Paul G. Francoeur, Tomohide Masuda, Jocelyn Sunseri, Andrew Jia, Richard B. Iovanisci, Ian Snyder, and David R. Koes. 3D Convolutional Neural Networks and a CrossDocked Dataset for Structure-Based Drug Design. *Journal of chemical information and modeling*, 60(9):4200, 9 2020. ISSN 1549960X. doi: 10.1021/ACS.JCIM.0C00411. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC8902699/.

Chen Fu and Qiuchen Chen. The future of pharmaceuticals: Artificial intelligence in drug discovery and development. *Journal of Pharmaceutical Analysis*, pp. 101248, 2 2025. ISSN 2095-1779. doi: 10.1016/J.JPHA.2025.101248. URL https://linkinghub.elsevier.com/retrieve/pii/S2095177925000656.

Charles Harris, Kieran Didi, Arian Rokkum Jamasb, Chaitanya K. Joshi, Simon V Mathis, Pietro Lio, and Tom Leon Blundell. PoseCheck: Generative Models for 3D Structure-based Drug Design Produce Unrealistic Poses, 2023.

Amit Kadan, Kevin Ryczko, Erika Lloyd, Adrian Roitberg, and Takeshi Yamazaki. Guided multi-objective generative AI to enhance structure-based drug design. *Chemical science*, 16(29):13196–13210, 7 2025. ISSN 20416539. doi: 10.1039/d5sc01778e. URL https://pubmed.ncbi.nlm.nih.gov/40463429/.

Hannes H. Loeffler, Jiazhen He, Alessandro Tibo, Jon Paul Janet, Alexey Voronov, Lewis H. Mervin, and Ola Engkvist. Reinvent 4: Modern AI–driven generative molecule design. *Journal of Cheminformatics*, 16(1):1–16, 12 2024. ISSN 17582946. doi: 10.1186/S13321-024-00812-5/METRICS. URL https://jcheminf.biomedcentral.com/articles/10.1186/s13321-024-00812-5http://creativecommons.org/publicdomain/zero/1.0/.

Fabian Pedregosa, Vincent Michel, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Jake Vanderplas, David Cournapeau, Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Bertrand Thirion, Olivier Grisel, Vincent Dubourg, Alexandre Passos, Matthieu Brucher, Perrot Édouardand, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *The Journal of Machine Learning Research*, 12:2825–2830, 11 2011. ISSN 1533-7928. doi: 10.5555/1953048.2078195. URL https://dl.acm.org/doi/pdf/10.5555/1953048.2078195.

P. G. Polishchuk, T. I. Madzhidov, and A. Varnek. Estimation of the size of drug-like chemical space based on GDB-17 data. *Journal of Computer-Aided Molecular Design 2013 27:8*, 27(8): 675–679, 8 2013. ISSN 15734951. doi: 10.1007/s10822-013-9672-4. URL https://link.springer.com/article/10.1007/s10822-013-9672-4.

Chris J. Radoux, Tjelvar S.G. Olsson, Will R. Pitt, Colin R. Groom, and Tom L. Blundell. Identifying Interactions that Determine Fragment Binding at Protein Hotspots. *Journal of Medicinal Chemistry*, 59(9):4314–4325, 5 2016. ISSN 15204804. doi: 10.1021/acs.jmedchem.5b01980. URL /doi/pdf/10.1021/acs.jmedchem.5b01980?ref=article_openPDF.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *Advances in Neural Information Processing Systems*, 36, 5 2023. ISSN 10495258. URL https://arxiv.org/pdf/2305.18290.

Anastasiia V. Sadybekov and Vsevolod Katritch. Computational approaches streamlining drug discovery. *Nature 2023 616:7958*, 616(7958):673–685, 4 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-05905-z. URL https://www.nature.com/articles/s41586-023-05905-z.

Natasha Sanjrani, Damien E. Coupry, Peter Pogány, David S. Palmer, and Stephen D. Pickett. Benchmarking 3D Structure-Based Molecule Generators. *Journal of Chemical Information and Modeling*, 65(15):8006–8021, 8 2025. ISSN 1549960X. doi: 10.1021/ACS.JCIM.5C01020. URL /doi/pdf/10.1021/acs.jcim.5c01020?ref=article_openPDF.

Arne Schneuing, Charles Harris, Yuanqi Du, Kieran Didi, Arian Jamasb, Ilia Igashov, Weitao Du, Carla Gomes, Tom L. Blundell, Pietro Lio, Max Welling, Michael Bronstein, and Bruno Correia. Structure-based drug design with equivariant diffusion models. *Nature Computational Science*, 4(12):899–909, 12 2024. ISSN 26628457. doi: 10.1038/S43588-024-00737-X;SUBJMETA=114,1305,154,631;KWRD=DRUG+DISCOVERY,MACHINE+LEARNING. URL https://www.nature.com/articles/s43588-024-00737-x.

Arne Schneuing, Ilia Igashov, Adrian W. Dobbelstein, Thomas Castiglione, Michael M. Bronstein, and Bruno Correia. Multi-domain Distribution Learning for De Novo Drug Design. *International Conference on Learning Representations*, 2025.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov Openai. Proximal Policy Optimization Algorithms. 2017. URL https://arxiv.org/pdf/1707.06347.

Chenghua Shao, Sebastian Bittrich, Sijian Wang, and Stephen K. Burley. Assessing PDB macromolecular crystal structure confidence at the individual amino acid residue level. *Structure*, 30 (10):1385–1394, 10 2022. ISSN 18784186. doi: 10.1016/j.str.2022.08.004. URL https://doi.org/10.1126/science.abj8754.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. 2 2024. URL https://arxiv.org/pdf/2402.03300.

Duxin Sun, Wei Gao, Hongxiang Hu, and Simon Zhou. Why 90% of clinical drug development fails and how to improve it? *Acta Pharmaceutica Sinica B*, 12(7):3049–3062, 7 2022. ISSN 2211-3835. doi: 10.1016/J.APSB.2022.02.002. URL https://www.sciencedirect.com/science/article/pii/S2211383522000521?via%3Dihub.

Sherine E. Thomas, Vitor Mendes, So Yeon Kim, Sony Malhotra, Bernardo Ochoa-Montaño, Michal Blaszczyk, and Tom L. Blundell. Structural Biology and the Design of New Therapeutics: From HIV and Cancer to Mycobacterial Infections: A Paper Dedicated to John Kendrew. *Journal of Molecular Biology*, 429(17):2677–2693, 8 2017. ISSN 10898638. doi: 10.1016/j.jmb.2017.06.014. URL https://doi.org/10.1038/167929a0.

Rob L.M. van Montfort, Paul Workman, Rob L.M. van Montfort, and Paul Workman. Structure-based drug design: aiming for a perfect fit. *Essays in Biochemistry*, 61(5):431–437, 11 2017. ISSN 00711365. doi: 10.1042/EBC20170052. URL /essaysbiochem/article/61/5/431/78244/Structure-based-drug-design-aiming-for-a-perfecthttps://dx.doi.org/10.1042/EBC20170052.

Marcel L. Verdonk, Paul N. Mortenson, Richard J. Hall, Michael J. Hartshorn, and Christopher W. Murray. Protein Ligand Docking against Non-Native Protein Conformers. *Journal of Chemical Information and Modeling*, 48(11):2214–2225, 2008. ISSN 1549960X. doi: 10.1021/ci8002254. URL /doi/pdf/10.1021/ci8002254?ref=article_openPDF.

Lucy Vost, Yael Ziv, and Charlotte M Deane. Incorporating targeted protein structure in deep learning methods for molecule generation in computational drug design. 2025. doi: 10.1039/d5sc05748e.

John D. Westbrook, Chenghua Shao, Zukang Feng, Marina Zhuravleva, Sameer Velankar, and Jasmine Young. The chemical component dictionary: complete descriptions of constituent molecules in experimentally determined 3D macromolecules in the Protein Data Bank. *Bioinformatics*, 31(8):1274–1278, 4 2015. ISSN 14602059. doi: 10.1093/bioinformatics/btu789. URL https://dx.doi.org/10.1093/bioinformatics/btu789.

Bo Yang, Chijian Xiang, Tongtong Li, Yunong Xu, and Jianing Li. Structure-Based Generation of 3D Small-Molecule Drugs: Are We There Yet? *Journal of medicinal chemistry*, 68(21):22756–22768, 11 2025. ISSN 15204804. doi: 10.1021/ACS.JMEDCHEM.5C01706/SUPPL{_}FILE/JM5C01706{_}SI{_}002.CSV. URL https://pubs.acs.org/doi/abs/10.1021/acs.jmedchem.5c01706.

A1 APPENDIX

A1.1 TARGET SELECTION AND REFINEMENT

A1.1.1 TARGET SELECTION PROTOCOL

For each target, structures were retrieved from the RCSB PDB(Berman et al., 2000) with 100% sequence identity to prevent mutations. Ligands were filtered using the PDB Chemical Component Dictionary (CCD) (Westbrook et al., 2015): ligands were retained with 5-55 heavy atoms, at least one carbon, and composed solely of H, B, C, N, O, F, P, S, Cl, Br, and I. Common non-drug molecules (ions, sugars, amino acids, nucleotides, buffers, solvents, and cofactors) were excluded using established blocklists (Radoux et al., 2016; Shao et al., 2022; Verdonk et al., 2008). Data was prepared for training following steps outlined by Schneuing et al. (2024).

A1.1.2 TEST TARGET SELECTION PROTOCOL

Test structures were selected to capture binding site conformational diversity. Structures were aligned using TM-align and clustered based on TM-score similarity matrices. Although clustering aimed to identify conformationally distinct pockets induced by different ligands, most targets formed a single cluster, indicating limited conformational variation. For each target, we selected three test structures: the highest resolution structure with a ligand bound, a conformationally distinct structure if available, and a randomly selected structure to ensure unbiased coverage. All datasets share 100% sequence similarity with their corresponding test targets by experimental design.

Table A1: Test structures for each target dataset.

Target	**PDB ID**
BRD4-BD1	4WHW, 6FO5, 6XVC
CA-II	3K34, 5N0D, 6RL9
EGFR	3POZ, 4WKQ, 8A27
ERα	2QZO, 4IVY, 5KCT
FXa	1EZQ, 2P3T, 3KL6
HIV-1-PR	1HOS, 2QNN, 3T11

A1.2 DESIGN OF DBSCAN PHARMACOPHORE REWARD MAP

To generate the DBSCAN map, all PDBs for a given target were aligned using BioPython (Cock et al., 2009) to a reference structure. Proteins were then removed and pharmacophoric features (Donor, Acceptor, Aromatic, Hydrophobe, NegIonizable, PosIonizable, ZnBinder, LumpedHydrophobe) were extracted from the ligands using the RDKit feature factory (RDK).

Each feature was represented by its 3D coordinates and clustered independently per feature type using DBSCAN, where a cluster was defined as a region of high feature density consisting of at least 10 features (minimum samples = 10) such that each feature lay within 0.5 Å (ϵ) of at least one other feature in the same cluster. However, for aromatic clustering, the minimum sample count was reduced to 5 due to greater positional variability in aromatic binding across all targets.

Below is an example of how the 3D map is visualised alongside the aligned ligands for CA-II. Multimeric PDB structures were separated into individual chains, and each chain was aligned independently to preserve conformational variation across different chain environments.

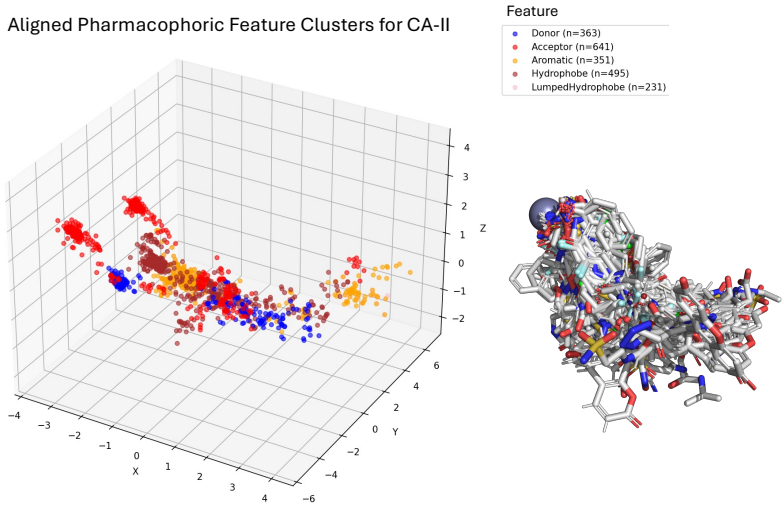

Figure A1: Map used for the pharmacophoric feature reward created using DBSCAN.On the left, an example of a cluster map for CA-II, covering aromatics (yellow), hydrophobic (brown), hydrogen bond acceptors (red) and donor (blue) features. On the right, the aligned ligands from which the clusters were formed are shown alongside the zinc atom sitting in the pocket which drives binding for this target. Atoms colors: nitrogen: blue, oxygens: red, sulfur: yellow, carbon: white

.

A1.3 Hyperparameters & Training Dynamics

The experiments were all run with the same hyperparameters, shown in Table A2 except for number of epochs, which varied between 35-80 depending on the dataset. The most important of these for stability was learning rate, which has to be significantly lower than other PPO and machine learning models during training. The best seed out of 4 (42, 123, 789, 976) was chosen for further training in the curriculum or for generation. We used a single H100 NVIDIA GPU. Wall clock time for training varies depending on the reward function, however, training for the aforementioned rewards never exceeded 6 hours.

Table A2: Hyperparameters used in all reward experiments. A single H100 GPU was used for all runs. Epoch count is variable across datasets and optimisation success.

Category	**Hyperparameter**	**Value**
Data / Throughput	Protein batch size	108
	GPU configuration	$1 \times$ H100
Conditioning / Representation	Conditioning mode	Pocket conditioning
	Pocket representation	Full-atom
PPO Training	Outer epochs (PPO cycles)	35-80
	Inner epochs per rollout	2
	Rollout length (n_{steps})	216
	PPO batch size	108
	Clipping range	0.1
Optimization	Gradient accumulation steps	42
	Training timesteps	300
	Learning rate	1×10^{-5}

Figure A2 shows PRISM optimising DiffSBDD with PPO for the geometry reward as a single objective across different seeds and all datasets.

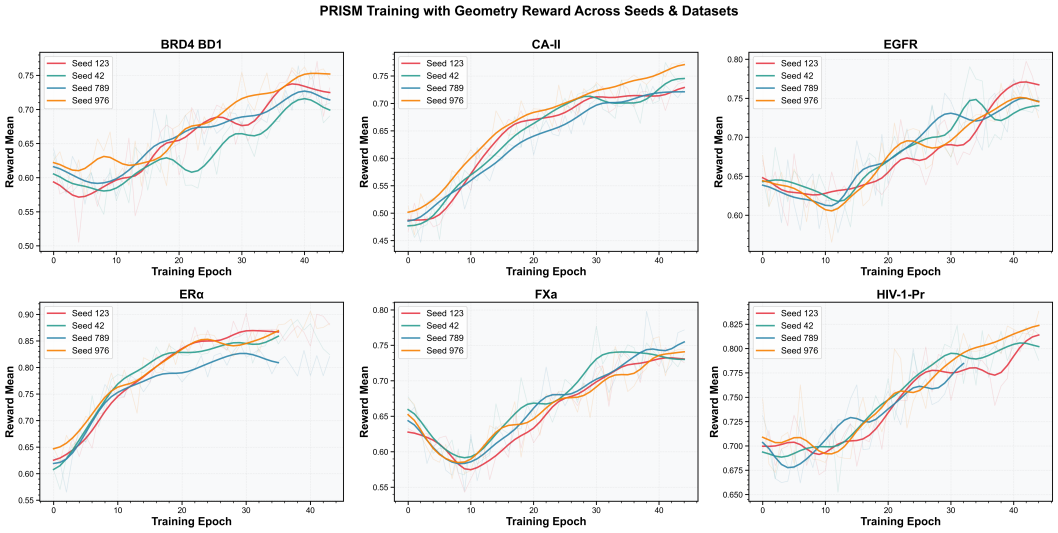

Figure A2: PRISM training across epochs with geometry as a reward for all datasets and multiple seeds. As training progresses, mean reward increases over epochs consistently, showing successful optimisation with RL.

A1.4 Single Objective Geometry Reward: Further Analysis

Table A3 exhibits detailed PoseBuster's breakdown which reveals the model successfully optimised for bond lengths, angles and internal steric clashes.

Table A3: Detailed PoseBusters failure rates by check type. Values show percentage of molecules failing each individual check (molecules can fail multiple checks). $N = 90{,}000$ molecules per method

Check	**Failures**		**Failure Rate (%)**	
	DiffSBDD	PRISM	DiffSBDD	PRISM
Bond angles	29,387	14,891	32.7	16.5
Bond lengths	23,054	7,924	25.6	8.8
Internal steric clash	10,500	3,300	11.7	3.7
Non-aromatic ring flatness	1,336	1,127	1.5	1.3
Double bond flatness	1,183	1,121	1.3	1.2
Aromatic ring flatness	236	272	0.3	0.3

Figure A3 shows little difference between the distribution of the prior (DiffSBDD trained on CrossDocked2020 Francoeur et al. (2020)) and PRISM after training with the single objective geometry reward. However, we note a slight increase in the number of rotatable bonds in all datasets except BRD4-BD1. This likely stems from the fact single bonds require less precision than double, triple or aromatic bonds in length and angles. Additionally, a decrease in overall ring count may be related to the internal steric clash aspect of the geometry reward, as too large, or small rings tend to be culprits for clashes in *de novo* generative models.

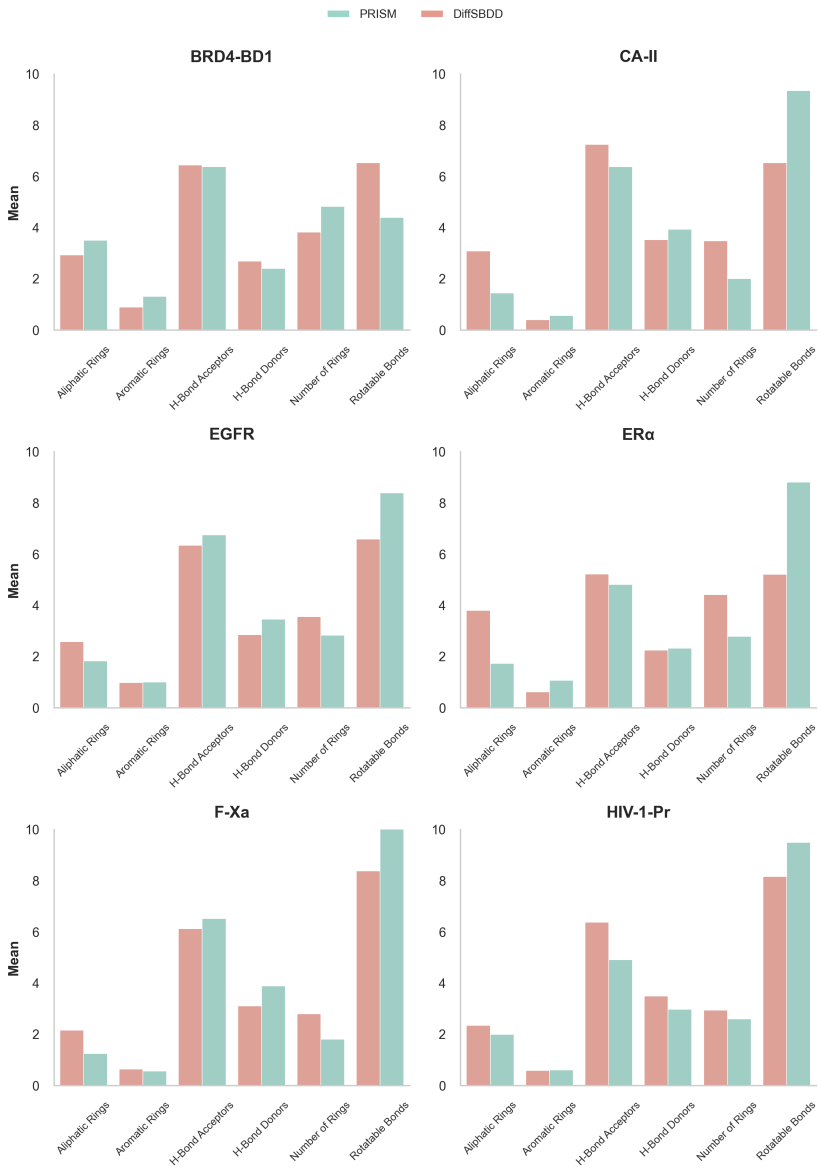

Figure A3: Comparative analysis of molecular property distributions for PRISM vs. DiffSBDD across datasets following training. Mean aliphatic, aromatic rings, hydrogen bond donors and acceptors, number of overall rings as well rotatable bonds are captured. PRISM shows a slight increased propensity for rotatable bonds and makes fewer rings following RL training on single objective geometry reward. $N = 30,000$ per test set per method

A1.5 Multi-objective Optimisation: Further Analysis

Chemical space analysis with t-SNE(A4) demonstrates an observable distributional shift where PRISM-generated molecules occupy the chemical ”neighborhood” of known binders more densely than the DiffSBDD prior for successfully optimised target, CA-II.

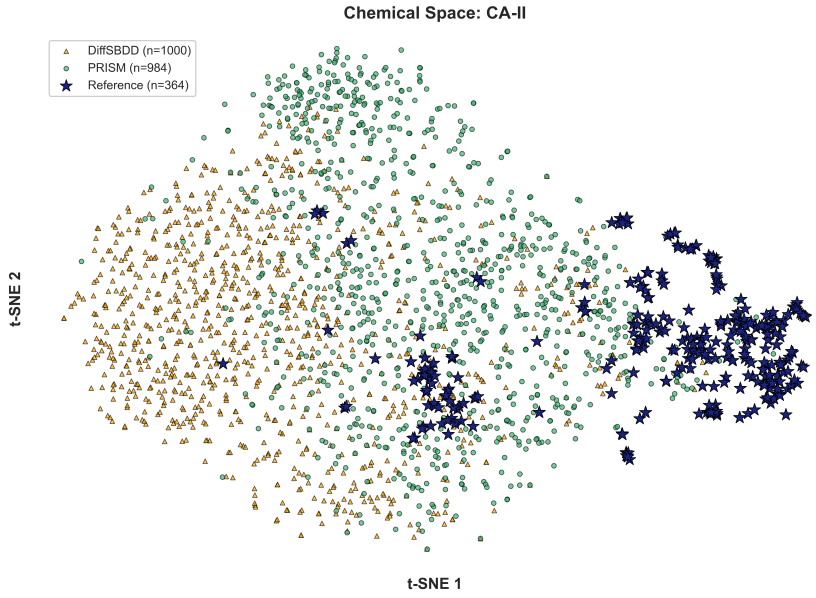

Figure A4: Chemical space projection using PCA-augmented t-SNE. ECFP4 fingerprints (2048-bit, computed with RDKit) were projected into two dimensions to visualize the generative manifold relative to known binders for CA-II test target (PDB ID: 5N0D). Compared to the DiffSBDD prior, PRISM exhibits a clear distributional shift toward the reference chemical space (dark blue stars). $N = 1{,}000$ molecules per method.

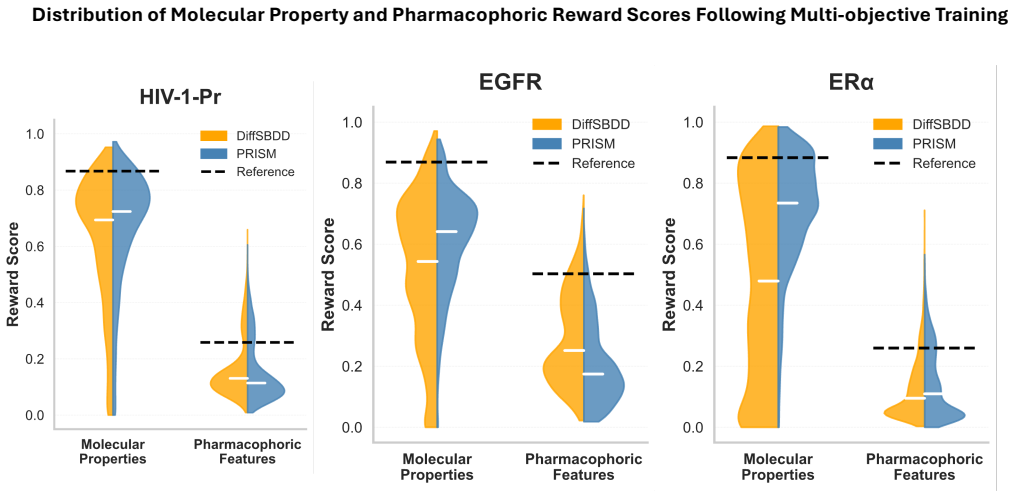

Figure A5: Distribution of molecular property and pharmacophoric reward scores following multi-objective training. PRISM (blue) and DiffSBDD (orange) are shown, with reference means indicated by black dashed lines. PRISM exhibits lower pharmacophoric feature scores but consistently higher molecular property scores compared to DiffSBDD. Performance varies by target, with HIV-1-PR and ERα showing stagnation in pharmacophoric scores, while EGFR exhibits mixed behavior. Values are aggregated across $N = 30{,}000$ molecules per test set, per method.

Table A4: Molecular diversity and drug-likeness metrics. QED: quantitative estimate of drug-likeness. SA: synthetic accessibility (normalized, higher is better). Diversity: mean pairwise Tanimoto distance. Novelty: fraction novel vs training set. PRISM trained with multi-objective optimization (geometry, pharmacophore, molecular properties). $N = 3,000$ molecules per test set, per method.

	QED ↑		**SA ↑**		**Diversity**		**Novelty**	
Target	DiffSBDD	PRISM	DiffSBDD	PRISM	DiffSBDD	PRISM	DiffSBDD	PRISM
BRD4-BD1	0.44	0.44	0.55	0.60	0.60	0.54	1.0	1.0
CA-II	0.46	0.47	0.61	0.63	0.76	0.76	1.0	1.0
EGFR	0.38	0.33	0.52	0.52	0.56	0.49	1.0	1.0
ERα	0.61	0.55	0.56	0.63	0.70	0.67	1.0	1.0
FXa	0.36	0.21	0.57	0.63	0.71	0.76	1.0	1.0
HIV-1-PR	0.25	0.28	0.46	0.45	0.60	0.45	1.0	1.0

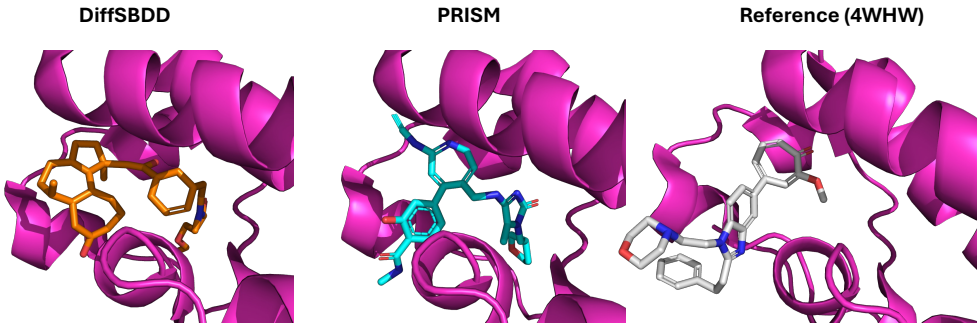

Figure A6: Molecules generated using DiffSBDD (orange), multi-objective optimised PRISM (blue) and the reference ligand (white) shown in BRD4-BD1 test pocket PDB: 4WHW (pink). PRISM shows increased shape, molecular feature similarity as well improved geometric validity compared to DiffSBDD.