BLENDING EXPERIMENTAL ELECTRON DENSITY INFORMATION TO PROTEIN-LIGAND INTERACTION

Seonghyun Park[1] **Shuan Chen**[2] **Jun Jin Choong**[2] **Noboru Kuno**[2] **Gyubok Lee**[2]
Kaushalya Madhawa[2] **Keisuke Ozawa**[2] **Hikaru Shindo**[2] **Sungsoo Ahn**[1]
[1]KAIST [2]SB Intuitions
{hyun26,sungsoo.ahn}@kaist.ac.kr

ABSTRACT

Machine learning models for protein–ligand interactions commonly rely on atomic coordinates, which are constructed by fitting molecular models to experimental electron-density maps. However, most models represent each complex as a single fitted structure, which can omit information about conformational heterogeneity, positional uncertainty, and local density support for individual atoms at binding interfaces. To fill this gap, we present **CoDE**, a **Co**ordinate and **D**ensity **E**ncoder, pre-trained on experimental density maps. We represent each binding site on a shared three-dimensional voxel grid combining atomic occupancy, experimental electron density, and its gradient magnitude. Using approximately 100K curated complexes, we pre-train a channel-wise vision transformer to jointly reconstruct masked atomic and density channels, learning a representation that integrates structural and experimental information. We evaluate the frozen encoder on two downstream tasks: (i) binding-affinity regression under leakage-controlled splits and (ii) pocket-conditioned ligand generation. Across these settings, CoDE improves over an identically pre-trained coordinate-only encoder and remains competitive with task-specific baselines, showing that density complements fitted atomic coordinates for protein–ligand interaction modeling and that the density-blended representation is useful for structure-based ligand generation.

1 INTRODUCTION

Modern machine learning for proteins primarily relies on amino acid sequences and three-dimensional atomic coordinates. Sequence models learn evolutionary and functional patterns (Rives et al., 2021; Lin et al., 2023), while structure-based models capture molecular geometry and spatial interactions from atomic coordinates (Jumper et al., 2021; Abramson et al., 2024). Their widespread use is supported by large-scale resources such as UniProt (Consortium, 2019) and the Protein Data Bank (Berman et al., 2000). However, these models often use only a single set of coordinates fitted to X-ray diffraction data, which does not convey how strongly the data support each fitted position or whether alternative conformations are present. This gap is consequential at protein–ligand interfaces, where binding depends on local structure.

Crystallographic electron-density maps offer a complementary view of the same complex, reflecting how well the data constrain each atom. A $2mF_o$-DF_c map combines observed diffraction amplitudes with model-dependent phase estimates and weights (Read, 1986). Its local pattern can reflect occupancy, atomic displacement, and unresolved alternative conformations, alongside resolution and model error (Kuzmanic et al., 2014; Lang et al., 2014). For instance, greater atomic displacement or disorder, reflected in a higher B factor, spreads the local density over a larger volume (Sun et al., 2019). The density map at a protein–ligand interface may therefore help a model interpret both a proposed contact and the structural ambiguity around it (van Zundert et al., 2018).

Yet, experimental electron-density maps remain underused for learning protein–ligand interactions, while established applications mostly concern crystallographic structure determination (Kowiel et al., 2019). Existing molecular models often construct density-like inputs from coordinates through occupancy grids (Ragoza et al., 2022; Pinheiro et al., 2024) or quantum-chemical calculations (Is-

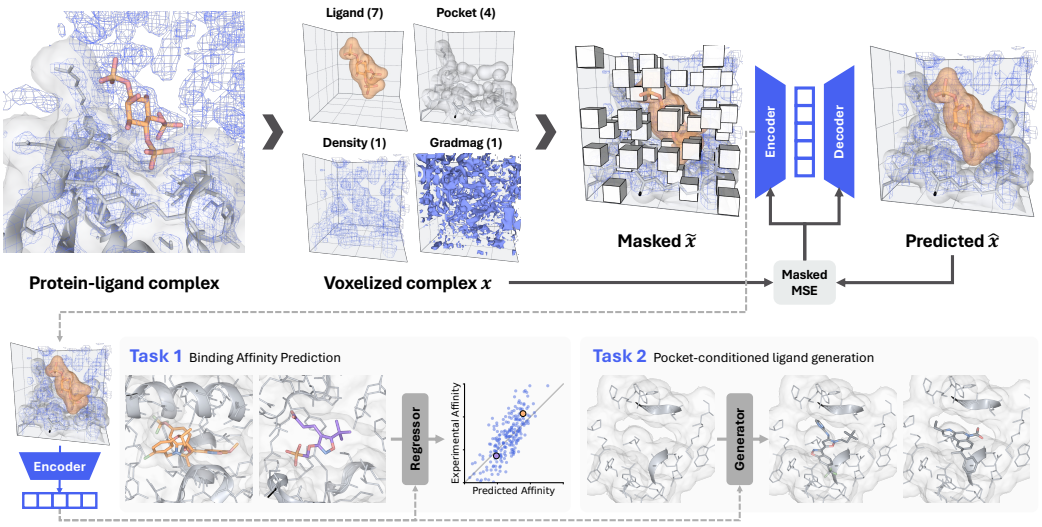


Figure 1: **Overview of pre-training and downstream transfers.** Each protein–ligand complex is voxelized into 13 channels: seven ligand and four pocket heavy-atom occupancy channels, plus experimental density and its gradient magnitude. A Channel-ViT is trained to reconstruct masked regions. For downstream transfer, the frozen encoder is used with a regressor for binding-affinity regression, or fused with a generator for pocket-conditioned ligand generation.

ert et al., 2024; Lin et al., 2025), without incorporating additional crystallographic observations. Recent work evaluates electron density maps as alternatives to atom-based inputs for affinity prediction (Suriana et al., 2025). It remains unclear whether experimental electron-density maps complement the fitted coordinates in representing protein–ligand interactions.

Contribution. To fill this gap, we empirically examine the value of experimental electron density as a complementary input for modeling protein–ligand interactions. We present CoDE, a **Co**ordinate and **D**ensity **E**ncoder, pre-trained by jointly reconstructing masked atomic occupancy, density, and density gradient magnitude on a shared voxel grid, encouraging the representation to encode structural ambiguity or disorder. Curated PLINDER complexes (Durairaj et al., 2024) paired with experimental $2mF_o$-DF_c maps support pre-training, with a matched coordinate-only encoder and channel ablations providing controlled comparisons. We evaluate the frozen representations on two downstream tasks: (i) binding-affinity regression, which probes protein–ligand interaction modeling, and (ii) pocket-conditioned ligand generation, a representative structure-based drug-design task. Both tasks use the reference's experimental density map for the target of interest, allowing us to assess the utility of the representations in scenarios where crystallographic data are available.

In our experiments, we make two observations:

- Electron-density map information is indeed predictive of protein–ligand interactions. CoDE improves over the matched coordinate-only encoder, CoE, in binding-affinity regression and pocket-conditioned ligand generation. When used to condition VoxBind, it also improves docking scores over the original generator and remains competitive with task-specific baselines, while its molecule-quality metrics stay close to VoxBind. Additionally, it produces more consistent local geometry, with gains growing on large rigid fragments.
- The benefit of density is associated with structural disorder. Without being trained to predict B-factors, the frozen CoDE encoder more accurately decodes these values at the complex, residue, and atom levels than CoE. Its affinity-ranking advantage over CoE is larger in higher complex contact B-factors, consistent with density helping most where coordinates are less certain.

2 Related Works

Experimental electron-density maps. An experimental electron-density map is reconstructed from the diffraction intensities of an X-ray experiment, conventionally as a likelihood-weighted $2mF_o$-

DF_c map (Read, 1986). Since the experiment measures no phases, the map combines the observed amplitudes F_o with model-derived amplitudes F_c and phases, weighted by the figure of merit m and the model error D. Because it retains F_o, the map preserves evidence of partial occupancy and alternative conformations that the fitted model omits (Lang et al., 2010; 2014). Prior works use such maps mainly to support the fitting itself, e.g., ligand identification (Kowiel et al., 2019; Karolczak et al., 2025) and local model assessment (Miyaguchi et al., 2021). Only recently, maps have been used as an input for affinity prediction (Suriana et al., 2025), attempting to replace atomic inputs rather than complementing them. Lin et al. (2025) shares a similar approach, tokenizing electron clouds computed from coordinates by quantum chemistry and using hierarchical codebooks. In contrast, we learn directly from experimentally measured maps.

Binding affinity regression. Binding-affinity regression predicts the interaction strength of a protein–ligand complex from experimentally measured binding constants (Wang et al., 2005), with methods differing in the structural detail they consume. *Sequence*-based models use a ligand SMILES and protein sequences without 3D geometry (Öztürk et al., 2018; Huang et al., 2021; Shenoy et al., 2026), whereas *structure*-based models use atomic coordinates through voxel CNNs (Stepniewska-Dziubinska et al., 2018), graph neural networks (Valsson et al., 2025; Kong et al., 2026), or equivariant architectures (Kong et al., 2023; Huang et al., 2024). Since labeled data are limited, self-supervised methods pre-train with denoising (Jin et al., 2023) or interaction-based objectives (Feng et al., 2024; Gao et al., 2024). Since data are scarce and redundant, which distorts evaluation, leakage-controlled splits and benchmarks have been proposed, removing test complexes resembling the training set (Graber et al., 2025; Li et al., 2026; Wei et al., 2026).

Pocket-conditioned ligand generation. Structure-based drug design uses target structures to identify or optimize ligands (Blundell, 1996; Anderson, 2003). Recent pocket-conditioned generators build 3D ligands directly inside a given target pocket. Autoregressive models place atoms or fragments sequentially (Luo et al., 2021; Peng et al., 2022), diffusion models denoise atom types and coordinates (Guan et al., 2023; Schneuing et al., 2024), and voxel models denoise a continuous atom occupancy grid (Pinheiro et al., 2023; 2024). A separate line of approaches supplies the generator with extra information about the target site; DecompDiff uses arm and scaffold priors estimated from a reference ligand (Guan et al., 2024), while others use ligand-shaped electron density and grow molecules into it (Wang et al., 2022; Li et al., 2025). Whereas Wang et al. (2022) use experimental density maps for task-specific ligand generation, we ask whether a representation pre-trained jointly on coordinates and density transfers to both binding-affinity regression and ligand generation.

3 METHOD

CODE represents a protein–ligand complex as a voxel grid combining atomic occupancy with experimental electron density. We pre-train an encoder on this grid by masked reconstruction and transfer its frozen representation to regression and generation tasks. Figure 1 illustrates the full pipeline.

Experimental electron-density map. Our input signal is the experimental electron-density map used to refine each X-ray crystal structure. For each complex in PLINDER (Durairaj et al., 2024), we fetch its $2mF_o$-DF_c map through PDBe (Armstrong et al., 2020) by the PDB ID and normalize it. Though the map is phased with the deposited model, it retains the observed amplitudes and thus evidence of what the model omits. To ensure the ligand is resolved in the map, we keep complexes with a ligand real-space correlation coefficient (Branden & Jones, 1990; Jones et al., 1991, RSCC) of at least 0.8 (Fusani et al., 2020). We also remove every downstream validation and test complex to prevent leakage. Further details are in Appendix A.

Voxel representation of protein–ligand complexes. To place atoms and density in a common frame, we build a binding-site grid centered at the ligand centroid (Pinheiro et al., 2024). The grid has G^3 voxels with $G = 64$ at 0.25 Å spacing, covering a 16 Å cube. It contains the ligand heavy atoms and the protein heavy atoms only within 10 Å of the ligand, cropped to the cube. Each complex then becomes a 13-channel tensor $\mathbf{x}$ as follows:

$$\mathbf{x} = \mathrm{Concat}(\mathbf{a}, \rho, \|\nabla\rho\|) \in \mathbb{R}^{G^3 \times 13}, \quad a_{v,c} = 1 - \prod_{i \in \mathcal{A}_c} (1 - g_{v,i}), \quad g_{v,i} = \exp\left(-\frac{\|\mathbf{r}_v - \mathbf{r}_i\|^2}{(0.93R_i)^2}\right). \tag{1}$$

Table 1: **Binding-affinity prediction on LP-PDBBind *test* and *CL3-filtered test* sets.** Seq. and Struc. denote sequence and structure inputs; structural inputs include coordinates and their chemical features. Results are averaged over five seeds. Affinity-leaked cases are marked with †, included for context only. **Bold** and underlining indicate the best and second-best results; overlapping intervals are marked together. N/A denotes unavailable metrics, and **Pre.** denotes pre-training.

			Test (N=1320)			CL3 filtered Test (N=733)		
Input	**Method**	**Pre.**	**Pearson r ↑**	**Spearman ρ ↑**	**RMSE ↓**	**Pearson r ↑**	**Spearman ρ ↑**	**RMSE ↓**
Seq.	DeepDTA	✗	0.535 ± 0.022	0.500 ± 0.023	1.510 ± 0.026	0.513 ± 0.026	0.481 ± 0.022	1.505 ± 0.028
	MolTrans	✗	0.494 ± 0.030	0.461 ± 0.026	1.567 ± 0.056	0.415 ± 0.131	0.404 ± 0.116	1.585 ± 0.076
	HonestAffinity	✓	0.472 ± 0.040	0.459 ± 0.046	1.582 ± 0.034	0.466 ± 0.030	0.436 ± 0.030	1.563 ± 0.034
	Nesso-1†	✓	0.670	0.663	1.410	0.673	0.659	1.367
Struc.	HBGSA	✗	0.567 ± 0.016	0.542 ± 0.022	1.703 ± 0.191	0.568 ± 0.015	0.548 ± 0.011	1.586 ± 0.201
	EGNN	✗	0.593 ± 0.008	0.583 ± 0.007	1.433 ± 0.014	0.595 ± 0.004	0.561 ± 0.005	1.403 ± 0.010
	$\text{EGNN}_{\text{TargetDiff}}$	✗	0.595 ± 0.007	0.584 ± 0.007	1.430 ± 0.013	0.590 ± 0.003	0.556 ± 0.005	1.410 ± 0.004
	GET	✗	0.593 ± 0.006	0.587 ± 0.006	1.430 ± 0.008	0.594 ± 0.003	0.559 ± 0.002	1.400 ± 0.006
	CheapNet	✗	0.628 ± 0.003	0.604 ± 0.003	1.417 ± 0.014	0.592 ± 0.024	0.551 ± 0.024	1.462 ± 0.041
	IPNet	✗	0.593 ± 0.022	0.569 ± 0.024	1.437 ± 0.041	0.556 ± 0.032	0.526 ± 0.020	1.440 ± 0.037
	IPNet†	✓	0.586 ± 0.006	0.554 ± 0.006	1.477 ± 0.019	0.584 ± 0.024	0.533 ± 0.028	1.433 ± 0.026
	AEV-PLIG	✗	0.552 ± 0.017	0.527 ± 0.015	1.593 ± 0.032	0.515 ± 0.016	0.476 ± 0.020	1.574 ± 0.031
	DSMBind	✓	0.308	0.363	N/A	0.464	0.480	N/A
	BindNet	✓	0.526 ± 0.025	0.507 ± 0.019	1.583 ± 0.105	0.487 ± 0.022	0.463 ± 0.018	1.567 ± 0.039
	ProFSA	✓	0.623 ± 0.005	0.594 ± 0.006	1.518 ± 0.012	0.645 ± 0.008	0.609 ± 0.009	1.370 ± 0.016
	CoE	✗	0.563 ± 0.023	0.553 ± 0.012	1.524 ± 0.058	0.567 ± 0.032	0.545 ± 0.017	1.484 ± 0.058
		✓	0.593 ± 0.009	0.560 ± 0.008	1.478 ± 0.032	0.620 ± 0.010	0.572 ± 0.011	1.401 ± 0.030
Struc., Density	CoDE (Ours)	✗	0.567 ± 0.017	0.557 ± 0.009	1.500 ± 0.042	0.570 ± 0.022	0.545 ± 0.013	1.457 ± 0.043
		✓	**0.654** ± 0.005	**0.641** ± 0.004	**1.364** ± 0.012	**0.667** ± 0.005	**0.638** ± 0.006	**1.314** ± 0.017

The occupancy $\mathbf{a} \in \mathbb{R}^{G^3 \times 11}$ has seven ligand (C, O, N, S, F, Cl, P) and four pocket (C, O, N, S) atom-type channels. Each channel c is the combination of Gaussian blobs $g_{v,i}$ over its heavy atoms $i \in \mathcal{A}_c$ (Orlando et al., 2022). A blob $g_{v,i}$ decays with the distance from the atom center $\mathbf{r}_i$ to the voxel center $\mathbf{r}_v$, at a width set by the element-specific van der Waals radius R_i. The density channel ρ is the map tri-linearly interpolated at each voxel center. Its gradient magnitude $\|\nabla\rho\|$ is computed on the same grid (Suriana et al., 2025). Further details are in Appendix A.

Masked pre-training. We then learn representations of the grid without labels, following masked image modeling (He et al., 2022; Xie et al., 2022). A Channel-ViT (Bao et al., 2023) encoder f_θ splits the grid into $(G/P)^3$ blocks of P^3 voxels, with $P = 8$ corresponding to a 2 Å cube. We mask 75% of the blocks and broadcast the mask to a voxel-level mask $\mathbf{M} \in \{0,1\}^{G^3}$. The encoder only sees the visible input $\widetilde{\mathbf{x}} = \mathbf{x} \odot (1 - \mathbf{M})$, with masked blocks zeroed rather than dropped. A lightweight reconstruction MLP head h_ψ then reconstructs all channels with a loss $\mathcal{L}$ as follows:

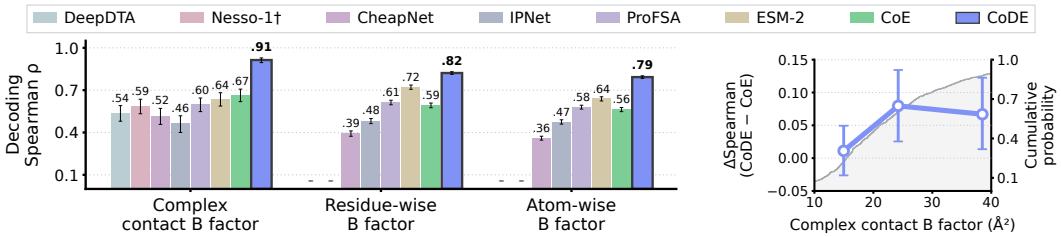
$$\mathcal{L}(\theta,\psi) = \frac{1}{|\mathcal{M}|C} \sum_{v\in\mathcal{M}} \sum_{c=1}^{C} w_c \omega_{c,v} \left(\widehat{x}_{v,c} - x_{v,c}\right)^2 , \qquad \widehat{\mathbf{x}} = h_\psi(f_\theta(\widetilde{\mathbf{x}})) , \tag{2}$$

where $\mathcal{M} = \{v : M_v = 1\}$ denotes the masked voxels. The channel weights w_c balance sparse element frequencies and density channels, while the voxel weight $\omega_{c,v}$ emphasizes occupied voxels. Additionally, we augment the entire grid by random rotation for robustness (Pinheiro et al., 2024).

Downstream task transfer. After pre-training, we freeze the encoder f_θ including its voxel projection, and discard the reconstruction head h_ψ. For affinity regression, we mean-pool the final patch tokens and fit an MLP regressor on them. For pocket-conditioned ligand generation, the voxel generator is unchanged and fuses the encoder's voxel output with VoxBind's features through a zero-initialized convolution, before the denoising U-Net. Only the generator and this fusion layer are trained, while the encoder stays frozen. For further details, refer to Appendix B.

4 Experiment Results

We validate the pre-trained representations with downstream tasks on protein–ligand complexes, spanning both discriminative and generative settings: binding affinity regression and pocket-



(a) Decoding structural uncertainty from representations (b) Binding-affinity improvement

Figure 2: **Structural uncertainty in representations and affinity prediction. (a) Decoding structure uncertainty.** B-factors at complex, residue, and atom levels are decoded from the frozen representation of the CL3 test complexes. DeepDTA and Nesso-1 appear only at the complex level, since their representations are not tied to residue and atom levels. **(b) Binding-affinity improvement.** We split the test complexes into three complex contact B-factor bins (< 20, $[20, 30)$, ≥ 30 Å^2), and show the CoDE−CoE Spearman gain. Results are averaged over five seeds, and points are placed at bin medians. The gray curve is the cumulative distribution of complex contact B-factor.

conditioned ligand generation. Both tasks use a reference density map of each target, which is a mild assumption since it exists for targets of interest, even when bound to a different ligand. Throughout, we compare against a coordinate-only variant of CoDE, coined CoE, that is identical except for the density and its gradient magnitude input, which isolates the contribution of the density information. For details on experiments and additional results, refer to Appendices A to E.

4.1 Binding affinity regression

Setup. We first use protein–ligand binding affinity regression as a controlled test of whether experimental density adds information beyond atomic coordinates. The goal of this experiment is not to establish a practical affinity predictor, but rather to determine whether density encodes complementary information. We use complex data from PDBBind v2020 (Liu et al., 2015), keeping only those with a crystallographic density map and affinity measured as K_d or K_i. We exclude validation and test complex data from the PLINDER-based pre-training set. We adopt the LP-PDBBind split (Li et al., 2026) to reduce leakage from complex-level interaction similarity, and report results on both the full test set and its CL3 subset, restricting evaluation to higher-quality complexes.

Evaluation. We measure linear and rank agreement by the Pearson (r) and Spearman (ρ) correlations, and the predictive error by RMSE. Results are averaged over five seeds, except for zero-shot pre-trained scorers with single predictions. Other results come from our own runs under identical protocols: pre-trained encoders use the same frozen-probe protocol as CoDE with their own probing heads, while the remaining methods are trained end-to-end with their own heads. Some baselines pre-train on data that overlaps with the LP-PDBBind test split, e.g., the co-folding model Nesso-1 (Shenoy et al., 2026) and affinity-supervised IPNet (Huang et al., 2024). We mark them with † and include them for context rather than as leakage-controlled comparisons.

Main results. Table 1 shows that CoDE improves over CoE on both test sets, and the pre-trained CoE itself is competitive with structure-based baselines. CoDE is comparable to the strongest structure-based baseline, ProFSA (Gao et al., 2024). The sequence-based co-folding model Nesso-1 (Shenoy et al., 2026) reaches similar correlation without any structural input, though its training data overlaps with the test split of LP-PDBBind. Without pre-training, CoDE and CoE perform comparably to weaker structure-based baselines, indicating that the pre-training provides transferable representations. The larger gain of CoDE further supports that density provides complementary information. Overall, we read these results as evidence that experimental electron density carries transferable information on binding.

Density helps in uncertain structures. We attribute the density gain to structural uncertainty, which we quantify through crystallographic B factors as a proxy for positional uncertainty (higher values indicate less certain atomic positions). To test what the representation encodes, we keep the encoder frozen and train lightweight MLPs to decode B factors at the complex, residue, and atom levels. In Figure 2a, CoDE decodes B factors more accurately than CoE and all baselines at every level. This

Table 3: **Pocket-conditioned ligand generation results on the CrossDocked2020 benchmark** for 79 test pockets with experimental density maps. Vina evaluation metrics are taken over all of the generated molecules as in prior works. High aff. denotes the per-pocket percentage of the generated molecules with a better Vina docking score than the reference, aggregated over pockets, and n denotes the number of valid molecules, counted as distinct per pocket for VoxBind variants.

		Vina evaluation								Sample quality							
		Score ↓		Min ↓		Dock ↓		High aff. % ↑		QED ↑		SA ↑		Div. ↑		# atoms / mol	
Method	n	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.
Reference	79	-6.21	-6.10	-6.66	-6.15	-7.32	-6.99	-	-	0.46	0.46	0.72	0.75	-	-	22.8	21.0
AR	7,655	-5.76	-5.58	-6.16	-5.85	-6.73	-6.54	41.8	37.0	0.50	0.50	0.63	0.62	0.71	0.71	17.2	16.0
Pocket2Mol	7,772	-5.06	-4.78	-6.47	-5.90	-7.27	-6.79	51.2	51.0	**0.57**	**0.57**	**0.74**	**0.75**	0.72	0.78	18.5	16.0
DiffSBDD	7,720	-2.70	-4.62	-4.89	-5.41	-6.71	-6.78	41.3	36.4	0.49	0.50	0.63	0.62	**0.80**	**0.80**	20.2	20.0
DecompDiff	6,427	-5.00	-4.97	-5.90	-5.70	-6.89	-6.77	49.8	46.8	0.49	0.50	0.67	0.65	0.71	0.72	20.9	21.0
TargetDiff	7,287	-5.27	-5.92	-6.42	-6.46	-7.48	-7.51	54.8	52.7	0.49	0.50	0.61	0.60	0.75	0.73	22.2	22.0
FuncBind	7,895	-5.75	-5.57	-6.40	-6.09	-7.27	-7.13	49.4	46.0	0.50	0.51	0.66	0.65	0.70	0.71	19.0	18.0
VoxBind$_{\sigma=0.9}$	7,887	-6.15	-6.83	-7.12	-7.22	-7.97	-8.05	61.7	65.7	0.54	0.55	0.70	0.71	0.76	0.76	24.0	25.0
+ CoE	5,911	-6.18	-6.07	-6.98	-6.80	-7.73	-7.58	58.5	66.0	0.43	0.44	0.58	0.58	0.56	0.54	24.4	25.0
+ CoDE	7,873	**-6.58**	**-7.20**	**-7.64**	**-7.70**	**-8.48**	**-8.51**	**68.1**	**78.0**	0.52	0.53	0.68	0.68	0.71	0.70	24.9	25.0

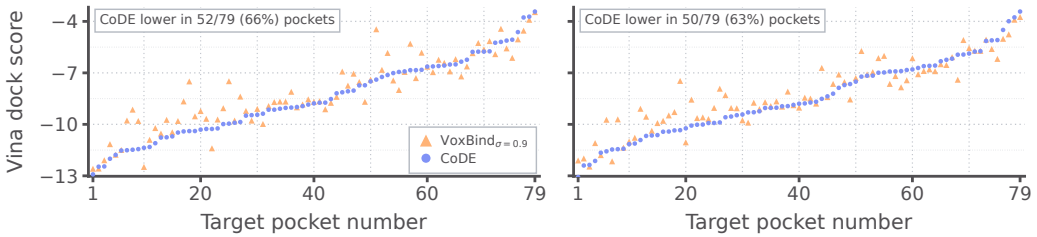


Figure 3: **Mean (left) and median (right) Vina docking scores** for generated molecules over 79 test pockets with experimental density maps, sorted by the CoDE Vina docking score. CoDE achieves better mean and median Vina docking scores than VoxBind for 66% and 63% of the pockets.

is partly expected, since B factors are refined against the map, but it confirms thatCoDE retains this information. Figure 2b relates this to binding affinity, where CoDE ranks better than CoE in the two higher B-factor bins while the lowest bin shows no difference. Density therefore helps most where the fitted coordinates are least certain, though this association is exploratory rather than causal. Further details are described in Appendix B.

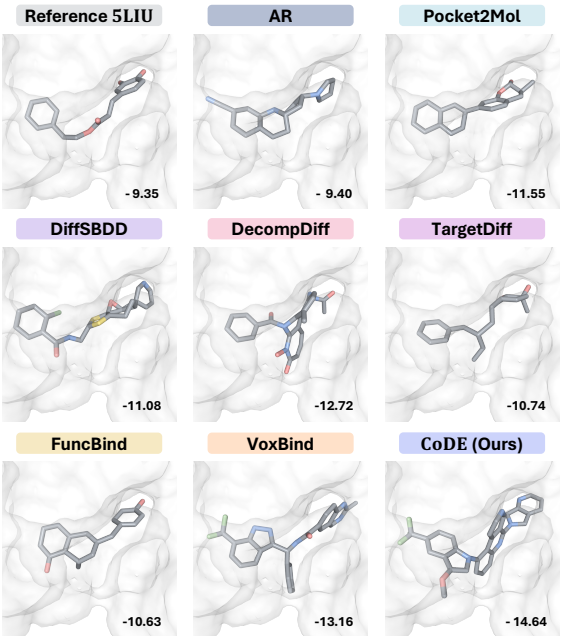
Design choices. We also isolate the density-aware representations on the CL3 subset in Table 2. Both density channels contribute, and neither alone recovers the full gain; removing either the density or its gradient magnitude falls below CoDE while staying above CoE. The gain also requires the experimental signal at the input. Supplying density only as a reconstruction target, replacing the experimental map with random values, and using gradient magnitudes derived from atom blobs ($\|\nabla a\|$) all stay below CoDE. Together, these results suggest that the benefit comes from the experimental density, providing information complementary to coordinates.

Table 2: **Design choice results**. **Bold** marks best, D and G denote density and gradmag.

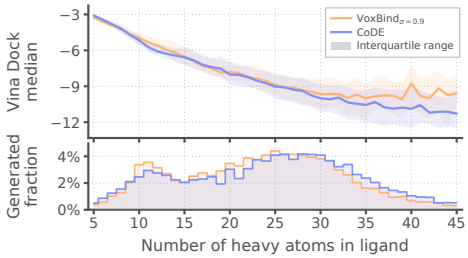
Design choices	Pearson r ↑	Spearman ρ ↑	RMSE ↓
CoDE (Ours)	**0.667** $\pm$ 0.005	**0.638** $\pm$ 0.006	**1.314** $\pm$ 0.017
- w.o. D, G (CoE)	0.620 $\pm$ 0.010	0.572 $\pm$ 0.011	1.401 $\pm$ 0.030
- CoE + D recon.	0.597 $\pm$ 0.021	0.562 $\pm$ 0.015	1.421 $\pm$ 0.048
- w.o. D channel	0.652 $\pm$ 0.005	0.621 $\pm$ 0.005	1.326 $\pm$ 0.010
- w.o. G channel	0.644 $\pm$ 0.008	0.605 $\pm$ 0.006	1.347 $\pm$ 0.016
- w. noisy D,G	0.627 $\pm$ 0.008	0.586 $\pm$ 0.010	1.380 $\pm$ 0.035
- w. $\|\nabla a\|$	0.638 $\pm$ 0.015	0.603 $\pm$ 0.015	1.377 $\pm$ 0.059

4.2 Pocket-conditioned ligand generation

Setup. Next, we evaluate pocket-conditioned ligand generation on the CrossDocked2020 benchmark (Francoeur et al., 2020), augmenting VoxBind (Pinheiro et al., 2024) with representations from CoDE. We pass the reference complex based on the PDB ID to the frozen encoder, producing features per P^3-voxel block. These features are fused with VoxBind's features through a zero-initialized convolution. All other components follow the original VoxBind. We adopt established splits and preprocessing protocols (Luo et al., 2021; Guan et al., 2023; Schneuing et al., 2024) and use 78,512 training data points and 79 test pockets with experimental maps for CoDE.



(a) Visualization of generated ligands



(b) Docking score by number of heavy atoms

Method	Novelty ↑ Mol. %	Novelty ↑ Scaf. %	SNN ↓
AR	**99.1**	79.3	0.297
Pocket2Mol	96.5	69.7	0.349
DiffSBDD	98.6	79.6	0.286
DecompDiff	96.9	76.2	0.335
TargetDiff	98.0	82.3	**0.281**
FuncBind	93.8	70.7	0.339
VoxBind$_{\sigma=0.9}$	97.2	80.6	0.350
+ CoE	98.3	82.2	0.343
+ CoDE	98.8	**84.8**	0.301

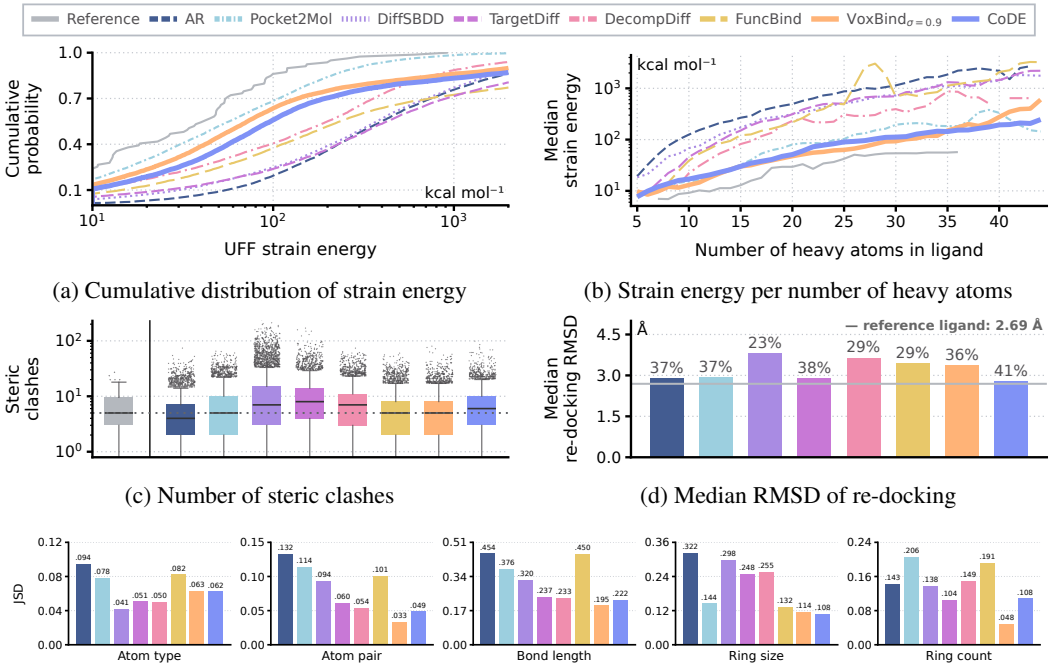
(c) Novelty against reference ligand

Figure 4: **(a) Qualitative examples of the generated ligands**. For pocket 5LIU, the best Vina-scoring ligand from each method is shown, with its docking score at the bottom right. **(b) Median Vina docking score by the number of heavy atoms**, with the distribution of generated ligand sizes at the bottom right. Docking scores do not simply become better proportionally to the molecule size, yet VoxBind + CoDE lowers them across atom counts. **(c) Novelty against the reference ligands**, measured by molecule-level and scaffold-level novelty and nearest-neighbor similarity.

Evaluation. Following Guan et al. (2023), we aim to generate 100 ligands for each pocket with every method and ask three questions of the generated ligands. First, do they bind better to the pocket? Stronger binding is the goal of pocket-conditioned ligand generation; therefore, we dock the ligands with AutoDock Vina (Eberhardt et al., 2021) and compare Vina scores and sample quality. Second, do the gains reflect new molecules rather than copies of the reference ligand? Since CoDE conditions on the reference complex, higher scores would be trivial if it simply replayed the reference, so we measure the novelty and nearest-neighbor similarity against the training set. Third, are the molecules physically plausible? A better docking score is meaningful only for a realistic pose; we check molecular geometry with PoseCheck (Harris et al., 2023), validity with PoseBusters (Buttenschoen et al., 2024), and geometry of the rigid fragments (Guan et al., 2023). For details and additional results on each evaluation, refer to Appendices D and E.

Density conditioning improves docking scores. As shown in Table 3, conditioning on experimental electron density results in ligands that dock better across pockets, while keeping sample quality close to VoxBind, which we report for context rather than as a ranking (Guan et al., 2023). This gain is broad rather than driven by a few pockets: Figure 3 shows a better mean and median docking score on a majority of pockets. As an example, Figure 4a visualizes the best-scoring ligand of each method for 5LIU. Also, CoDE's gain is not a size effect; while CoDE generates slightly larger molecules (# atoms/mol in Table 3), Figure 4b shows that larger molecules do not always dock better in VoxBind, yet CoDE's gain holds across heavy atom counts. Notably, CoE repeatedly generates the same ligands, leaving fewer distinct molecules within the same attempt budget (n in Table 3), whereas density conditioning avoids this.

Generated ligands do not copy the reference. We further show that CoDE does not simply replay the ligand from the reference, similar to DecompDiff (Guan et al., 2024). Against the CrossDocked2020 training split, we report novelty of the generated ligands at two levels: (i) the fraction of canonical SMILES (Weininger, 1988) and Bemis-Murcko scaffolds (Bemis & Murcko, 1996) that never appear in training (Polykovskiy et al., 2020; Brown et al., 2019), and (ii) the similar-



(a) Cumulative distribution of strain energy

(b) Strain energy per number of heavy atoms

(c) Number of steric clashes

(d) Median RMSD of re-docking

(e) JSD to reference ligands regarding molecular properties

Figure 5: **PoseCheck evaluation of generated molecule poses.** Method colors are consistent across panels, with reference ligands shown in gray. **(a) Strain energy distributions** shown by the empirical cumulative distribution function. **(b) Median strain energy** as a function of the number of heavy atoms. **(c) Number of steric clashes.** Boxes denote the interquartile range, center lines denote the medians, whiskers extend to 1.5 times the interquartile range, and outliers are shown as individual points. **(d) Median re-docking RMSD.** Labels above the bars indicate the percentage of poses with RMSD within 2 Å, and the gray horizontal line marks the reference median of 2.69 Å. **(e) Jensen-Shannon distance (JSD) to reference ligands** for atom types, heavy-atom pair distances, bond lengths, ring sizes, and ring counts.

ity to the nearest neighbor (SNN) (Walters & Murcko, 2020) by the Tanimoto similarity (Bajusz et al., 2015) of each generated molecule to its closest training molecule. Figure 4c shows that every method maintains high molecule/scaffold novelty and low SNN. Despite conditioning on the reference complex, CoDE remains at least as novel as VoxBind, indicating that density information supplies pocket context rather than a template to copy. We additionally find low similarity between each generated ligand and the reference ligand of the test pocket; see Appendix D.4 for details.

Docking improved with realistic geometry. Since better docking scores are meaningful with realistic poses, we check the generated geometry with PoseCheck (Harris et al., 2023) following Guan et al. (2023); Pinheiro et al. (2024). Figure 5 reports three pose-level geometric measures, where CoDE shows comparable numbers to VoxBind, sometimes better. *Strain energy*, the universal force-field (Rappé et al., 1992, UFF) relaxed energy a pose has above a stable conformer pose, detects internally distorted molecules. Figure 5a shows that CoDE matches VoxBind and places ligands at lower strain than other baselines. This holds at fixed ligand size, i.e., the number of heavy atoms, as shown in Figure 5b. *Steric clashes*, the number of ligand atoms overlapping with protein atoms, detect poses the pocket cannot accommodate. CoDE is comparable to VoxBind and the reference ligands, while DiffSBDD (Schneuing et al., 2024) and TargetDiff (Guan et al., 2023) show much heavier tails. *Redocking RMSD*, how far Vina moves a pose when docking the molecule, measures the stability of the generated pose. As shown in Figure 5d, CoDE attains the lowest median among the methods and is closest to the reference.

Beyond individual poses, we also compare the generated molecules against the reference distribution with the Jensen-Shannon distance (Lin, 1991, JSD) over atom, bond, and ring statistics (Guan et al., 2023). Figure 5e shows that CoDE broadly follows the reference, although a lower JSD indicates

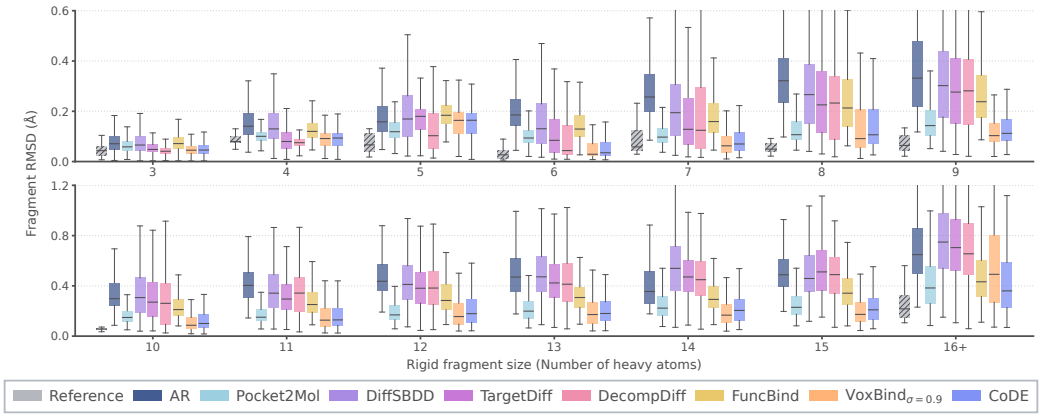


Figure 6: **Rigid fragment consistency results.** Median rigid-fragment RMSD between generated and force-field-relaxed pose after fragment-wise alignment, grouped by the number of heavy atoms. Boxes indicate the interquartile range, horizontal lines denote medians, and whiskers extend to 1.5 times the interquartile range. Lower RMSD indicates local geometry closer to a relaxed pose.

closeness to the reference rather than a better molecule. Interestingly, CoDE generates more rings than the reference ligands, even relative to their size, which raises its ring-count JSD.

Rigid fragments keep their relaxed geometry. CoDE reproduces the local geometry of its rigid fragments as well as VoxBind, and is best at the largest fragments. We follow Guan et al. (2023) and focus on rigid fragments, whose geometry is fixed by chemistry rather than by conformational choice, and report the RMSD between its generated pose and its force-field-relaxed pose, where lower is more physically plausible. Figure 6 shows that most methods place small fragments close to their relaxed geometry, and RMSD grows with fragment size. In contrast, CoDE stays with VoxBind while most baselines drift upward, and the margin widens along the rigid fragment size. While the improved consistency of large rigid fragments does not imply improved macrocycle generation, we additionally provide a macrocycle analysis in Appendix E.4.

Comparable chemical validity to VoxBind. Furthermore, conditioning on experimental electron density yields a comparable chemical validity to VoxBind. We screen the generated ligands without any relaxation with PoseBusters (Buttenschoen et al., 2024) in its `dock` configuration, and report the fraction that passes all checks. Figure 7 shows that CoDE is comparable to VoxBind and passes most categories at near-reference rates, with ring pucker as the main exception. Our post hoc analysis suggests that bond-order and hydrogen assignment in the OpenBabel-based reconstruction pipeline (O'Boyle et al., 2011) contribute to many ring non-flatness failures. Reassigning eligible rings as aromatic reduces the ring non-flatness failure rate while preserving heavy atom coordinates; further details are provided in Appendix E.8.

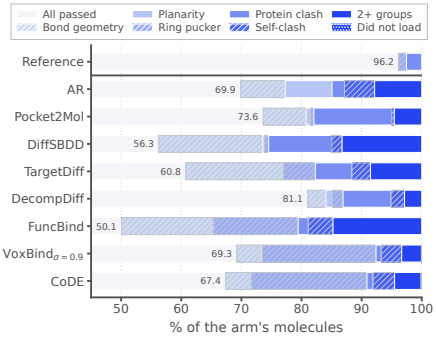


Figure 7: **Chemical validity by PoseBusters**, split by category (colors) with all-passed rate labeled; axis starts at 50%.

5 Conclusion

In this work, we introduced CoDE, a voxel encoder jointly pre-trained on atomic occupancy and experimental $2mF_o$-DF_c maps by masked reconstruction. Across frozen-probe affinity regression and pocket-conditioned ligand generation, CoDE improves over a matched coordinate-only encoder and stays competitive with task-specific baselines, and its advantage is largest where the fitted coordinates are least certain. These results support experimental electron density as a transferable complement to coordinates for learning protein–ligand interactions.

AI USE STATEMENT

We used generative AI tools, including OpenAI Codex and Claude Code, to assist with data processing, code implementation, model training, and evaluation, preparation of figures and tables, and drafting and polishing text. These tools were used under specific instructions reflecting the authors' ideas and intentions. We did not use generative AI tools to originate the core research hypothesis or methodology, formulate scientific claims, construct conclusions, or construct mathematical proofs. All AI-assisted work was reviewed by the authors: generated code and configurations were checked, experimental outputs and analysis were verified, and figures, tables, and text suggested were reviewed by the authors. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work focuses on computational modeling of protein-ligand interactions using structural and experimental electron-density information. As with other computational methods for molecular modeling and drug discovery, the proposed approach could potentially be applied to the design or evaluation of harmful compounds. We therefore encourage responsible use of the method and independent validation of computational predictions before experimental or practical use. All datasets used in this work are subject to their respective licenses and terms of use.

REPRODUCIBILITY STATEMENT

The code for data processing, pre-training, and downstream evaluation will be publicly released at https://github.com/seonghyun26/CoDE, with configuration described in the appendices. Additionally, the model architecture, input representation, preprocessing procedures, training settings, evaluation protocols, and hyperparameters used for the experiments are reported in the appendices. We specify all evaluation procedures used for downstream tasks. Reproducing experiments may additionally require access to the corresponding original data sources and compliance with their respective licensing conditions.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024. 1

Amy C Anderson. The process of structure-based drug design. *Chemistry & biology*, 10(9):787–797, 2003. 3

David R Armstrong, John M Berrisford, Matthew J Conroy, Aleksandras Gutmanas, Stephen Anyango, Preeti Choudhary, Alice R Clark, Jose M Dana, Mandar Deshpande, Roisin Dunlop, et al. Pdbe: improved findability of macromolecular structure data in the pdb. *Nucleic acids research*, 48(D1):D335–D343, 2020. 3

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 21

Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, 7(1):20, 2015. 8, 29

Yujia Bao, Srinivasan Sivanandan, and Theofanis Karaletsos. Channel vision transformers: An image is worth 1 x 16 x 16 words. *arXiv*, 2023. doi: 10.48550/arxiv.2309.16108. 4, 21

Guy W Bemis and Mark A Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of medicinal chemistry*, 39(15):2887–2893, 1996. 7, 29

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1): 235–242, 2000. 1

G Richard Bickerton, Gaia V Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L Hopkins. Quantifying the chemical beauty of drugs. *Nature chemistry*, 4(2):90–98, 2012. 28

Tom L Blundell. Structure-based drug design. *Nature*, 384(6604):23, 1996. 3

Carl-Ivar Branden and T Alwyn Jones. Between objectivity and subjectivity. *Nature*, 343(6260): 687, 1990. 3

Nathan Brown, Marco Fiscato, Marwin HS Segler, and Alain C Vaucher. Guacamol: benchmarking models for de novo molecular design. *Journal of chemical information and modeling*, 59(3): 1096–1108, 2019. 7, 29

Martin Buttenschoen, Garrett M Morris, and Charlotte M Deane. Posebusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical science*, 15(9):3130–3139, 2024. 7, 9, 31

UniProt Consortium. Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research*, 47 (D1):D506–D515, 2019. 1

Todd J Dolinsky, Paul Czodrowski, Hui Li, Jens E Nielsen, Jan H Jensen, Gerhard Klebe, and Nathan A Baker. Pdb2pqr: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic acids research*, 35(suppl_2):W522–W525, 2007. 27

Janani Durairaj, Yusuf Adeshina, Zhonglin Cao, Xuejin Zhang, Vladas Oleinikovas, Thomas Duignan, Zachary McClure, Xavier Robin, Gabriel Studer, Daniel Kovtun, et al. Plinder: The protein-ligand interactions dataset and evaluation resource. *BioRxiv*, pp. 2024–07, 2024. 2, 3, 18

Alexandre Duval, Simon V Mathis, Chaitanya K Joshi, Victor Schmidt, Santiago Miret, Fragkiskos D Malliaros, Taco Cohen, Pietro Lio, Yoshua Bengio, and Michael Bronstein. A hitchhiker's guide to geometric GNNs for 3d atomic systems. *arXiv*, 2023. doi: 10.48550/arxiv.2312.07511. 25

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, 2021. 7, 27

Bradley Efron. Better bootstrap confidence intervals. *Journal of the American statistical Association*, 82(397):171–185, 1987. 27

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):8, 2009. 28

Shikun Feng, Minghao Li, Yinjun Jia, Wei-Ying Ma, and Yanyan Lan. Protein-ligand binding representation learning from fine-grained interactions. In *International Conference on Learning Representations*, volume 2024, pp. 10625–10640, 2024. 3, 25

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, 2020. 6, 18

Lucia Fusani, David S Palmer, Don O Somers, and Ian D Wall. Exploring ligand stability in protein crystal structures using binding pose metadynamics. *Journal of Chemical Information and Modeling*, 60(3):1528, 2020. 3, 19

Bowen Gao, Yinjun Jia, Yuanle Mo, Yuyan Ni, Wei-Ying Ma, Zhi-Ming Ma, and Yanyan Lan. Self-supervised pocket pretraining via protein fragment-surroundings alignment. In *International Conference on Learning Representations*, volume 2024, pp. 25246–25268, 2024. 3, 5, 26

David Graber, Peter Stockinger, Fabian Meyer, Siddhartha Mishra, Claus Horn, and Rebecca Buller. Resolving data bias improves generalization in binding affinity prediction. *Nature Machine Intelligence*, 7(10):1713–1725, 2025. 3

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 *International Conference on Learning Representations*, volume 2023, 2023. doi: 10.48550/arxiv.2303.03543. 3, 6, 7, 8, 9, 18, 25, 26, 27, 28, 31, 32

Jiaqi Guan, Xiangxin Zhou, Yuwei Yang, Yu Bao, Jian Peng, Jianzhu Ma, Qiang Liu, Liang Wang, and Quanquan Gu. Decompdiff: diffusion models with decomposed priors for structure-based drug design. *arXiv preprint arXiv:2403.07902*, 2024. 3, 7, 18, 26

Thomas A Halgren. Merck molecular force field. i. basis, form, scope, parameterization, and performance of mmff94. *Journal of computational chemistry*, 17(5-6):490–519, 1996. 32

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. 7, 8, 30

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022. 4

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 21

Kexin Huang, Cao Xiao, Lucas M Glass, and Jimeng Sun. Moltrans: molecular interaction transformer for drug–target interaction prediction. *Bioinformatics*, 37(6):830–836, 2021. 3, 25

Zhilin Huang, Ling Yang, Xiangxin Zhou, Zhilong Zhang, Wentao Zhang, Xiawu Zheng, Jie Chen, Yu Wang, Bin Cui, and Wenming Yang. Protein-ligand interaction prior for binding-aware 3d molecule diffusion models. In *International Conference on Learning Representations*, volume 2024, pp. 18453–18474, 2024. 3, 5, 25

Clemens Isert, Kenneth Atz, Sereina Riniker, and Gisbert Schneider. Exploring protein–ligand binding affinity prediction with electron density-based geometric deep learning. *RSC advances*, 14(7):4492–4502, 2024. 1

Wengong Jin, Xun Chen, Amrita Vetticaden, Siranush Sarzikova, Raktima Raychowdhury, Caroline Uhler, and Nir Hacohen. Dsmbind: Se (3) denoising score matching for unsupervised binding energy prediction and nanobody design. *bioRxiv*, pp. 2023–12, 2023. 3, 25

T Alwyn Jones, J-Y Zou, SW t Cowan, and Morten Kjeldgaard. Improved methods for building protein models in electron density maps and the location of errors in these models. *Foundations of crystallography*, 47(2):110–119, 1991. 3

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021. 1, 18

Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Foundations of Crystallography*, 32(5):922–923, 1976. 19, 32

Jacek Karolczak, Anna Przybyłowska, Konrad Szewczyk, Witold Taisner, John M Heumann, Michael HB Stowell, Michał Nowicki, and Dariusz Brzezinski. Ligand identification in cryoem and x-ray maps using deep learning. *Bioinformatics*, 41(1):btae749, 2025. 3

Matthieu Kirchmeyer, Pedro O Pinheiro, Emma Willett, Karolis Martinkus, Joseph Kleinhenz, Emily Makowski, Andrew Watkins, Vladimir Gligorijevic, Richard Bonneau, and Saeed Saremi. Unified all-atom molecule generation with neural fields. *Advances in Neural Information Processing Systems*, 38:69643–69674, 2026. 26

Junxiao Kong, Chupei Tang, Di Wang, Jixiu Zhai, Yi He, Moyu Tang, and Tianchi Lu. Hbgsa: Hydrogen bond graph with self-attention for drug-target binding affinity prediction. *arXiv preprint arXiv:2604.23115*, 2026. 3, 25

Xiangzhe Kong, Wenbing Huang, and Yang Liu. Generalist equivariant transformer towards 3d molecular interaction learning. *arXiv preprint arXiv:2306.01474*, 2023. 3, 25

Marcin Kowiel, Dariusz Brzezinski, Przemyslaw J Porebski, Ivan G Shabalin, Mariusz Jaskolski, and Wladek Minor. Automatic recognition of ligands in electron density by machine learning. *Bioinformatics*, 35(3):452–461, 2019. 1, 3

Antonija Kuzmanic, Navraj S Pannu, and Bojan Zagrovic. X-ray refinement significantly underestimates the level of microscopic heterogeneity in biomolecular crystals. *Nature communications*, 5(1):3220, 2014. 1

Greg Landrum, Paolo Tosco, Brian Kelley, Ricardo Rodriguez, David Cosgrove, Riccardo Vianello, Peter Gedeck, Gareth Jones, Eisuke Kawashima, Dan Nealschneider, et al. rdkit/rdkit: 2025_03_1 (q1 2025) release. *Zenodo*, 2025. 28

P Therese Lang, Ho-Leung Ng, James S Fraser, Jacob E Corn, Nathaniel Echols, Mark Sales, James M Holton, and Tom Alber. Automated electron-density sampling reveals widespread conformational polymorphism in proteins. *Protein Science*, 19(7):1420–1431, 2010. 3

P Therese Lang, James M Holton, James S Fraser, and Tom Alber. Protein structural ensembles are revealed by redefining x-ray electron density noise. *Proceedings of the National Academy of Sciences*, 111(1):237–242, 2014. 1, 3

Jie Li, Xingyi Guan, Oufan Zhang, Kunyang Sun, Yingze Wang, Dorian Bagni, and Teresa Head-Gordon. Leak proof pdbbind: a reorganized data set of protein–ligand complexes for more generalizable binding affinity prediction. *The Journal of Physical Chemistry B*, 130(2):730–740, 2026. 3, 5, 18

Mingyu Li, Kun Song, Jixiao He, Mingzhu Zhao, Gengshu You, Jie Zhong, Mengxi Zhao, Arong Li, Yu Chen, Guobin Li, et al. Electron-density-informed effective and reliable de novo molecular design and optimization with ed2mol. *Nature Machine Intelligence*, 7(8):1355–1368, 2025. 3

Hyukjun Lim, Sun Kim, and Sangseon Lee. Cheapnet: Cross-attention on hierarchical representations for efficient protein-ligand binding affinity prediction. In *International Conference on Learning Representations*, volume 2025, pp. 89010–89043, 2025. 25

Haitao Lin, Odin Zhang, Jia Xu, Yunfan Liu, Zheng Cheng, Lirong Wu, Yufei Huang, Zhifeng Gao, and Stan Z Li. Tokenizing electron cloud in protein-ligand interaction learning. *arXiv preprint arXiv:2505.19014*, 2025. 2, 3

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991. 8, 31, 40

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. 1, 25

Shengchao Liu, Hongyu Guo, and Jian Tang. Molecular geometry pretraining with se (3)-invariant denoising distance matching. *arXiv preprint arXiv:2206.13602*, 2022. 25

Zhihai Liu, Yan Li, Li Han, Jie Li, Jie Liu, Zhixiong Zhao, Wei Nie, Yuchen Liu, and Renxiao Wang. Pdb-wide collection of binding data: current status of the pdbbind database. *Bioinformatics*, 31 (3):405–412, 2015. 5, 18, 25

Shitong Luo, Jiaqi Guan, Jianzhu Ma, and Jian Peng. A 3d generative model for structure-based drug design. *Advances in neural information processing systems*, 34:6229–6239, 2021. 3, 6, 18, 26

Ikuko Miyaguchi, Miwa Sato, Akiko Kashima, Hiroyuki Nakagawa, Yuichi Kokabu, Biao Ma, Shigeyuki Matsumoto, Atsushi Tokuhisa, Masateru Ohta, and Mitsunori Ikeguchi. Machine learning to estimate the local quality of protein crystal structures. *Scientific Reports*, 11(1):23599, 2021. 3

Garrett M Morris, Ruth Huey, William Lindstrom, Michel F Sanner, Richard K Belew, David S Goodsell, and Arthur J Olson. Autodock4 and autodocktools4: Automated docking with selective receptor flexibility. *Journal of computational chemistry*, 30(16):2785–2791, 2009. 27

Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. Open babel: An open chemical toolbox. *Journal of cheminformatics*, 3(1):33, 2011. 9, 45

Gabriele Orlando, Daniele Raimondi, Ramon Duran-Romaña, Yves Moreau, Joost Schymkowitz, and Frederic Rousseau. Pyuul provides an interface between biological structures and deep learning algorithms. *Nature communications*, 13(1):961, 2022. 4, 21

Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018. 3, 25

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, 2025. 25

Xingang Peng, Shitong Luo, Jiaqi Guan, Qi Xie, Jian Peng, and Jianzhu Ma. Pocket2mol: Efficient molecular sampling based on 3d protein pockets. In *International conference on machine learning*, pp. 17644–17655. PMLR, 2022. 3, 26

Pedro O Pinheiro, Joshua Rackers, Joseph Kleinhenz, Michael Maser, Omar Mahmood, Andrew Martin Watkins, Stephen Ra, Vishnu Sresht, and Saeed Saremi. 3d molecule generation by denoising voxel grids. *arXiv*, 2023. doi: 10.48550/arxiv.2306.07473. 3, 18

Pedro O Pinheiro, Arian Jamasb, Omar Mahmood, Vishnu Sresht, and Saeed Saremi. Structure-based drug design by denoising voxel grids. *arXiv preprint arXiv:2405.03961*, 2024. 1, 3, 4, 6, 8, 23, 26, 27, 31

Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, et al. Molecular sets (moses): a benchmarking platform for molecular generation models. *Frontiers in pharmacology*, 11:565644, 2020. 7, 29

Matthew Ragoza, Tomohide Masuda, and David Ryan Koes. Generating 3d molecules conditional on receptor binding sites with deep generative models. *Chemical science*, 13(9):2701–2713, 2022. 1

Anthony K Rappé, Carla J Casewit, Kent S Colwell, William A Goddard III, and W Mason Skiff. Uff, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American chemical society*, 114(25):10024–10035, 1992. 8

Randy J Read. Improved fourier coefficients for maps using phases from partial structures with errors. *Foundations of Crystallography*, 42(3):140–149, 1986. 1, 3

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the national academy of sciences*, 118(15):e2016239118, 2021. 1

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010. 29

Diogo Santos-Martins, Yiran He, Jerome Eberhardt, Parnika Sharma, Niccolò Bruciaferri, Matthew Holcomb, Manuel A Llanos, Althea Hansel-Harris, Allison Pearl Barkdull, Andreas Frank Tillack, et al. Meeko: molecule parametrization and software interoperability for docking and beyond. *Journal of chemical information and modeling*, 65(24):13045–13050, 2025. 27

Arne Schneuing, Charles Harris, Yuanqi Du, Kieran Didi, Arian Jamasb, Ilia Igashov, Weitao Du, Carla Gomes, Tom L Blundell, Pietro Lio, et al. Structure-based drug design with equivariant diffusion models. *Nature Computational Science*, 4(12):899–909, 2024. 3, 6, 8, 18, 26

Nikhil Shenoy, David Errington, Emmanuel Bengio, Kacper Kapuśniak, Kerstin Klaeser, Yui Tik Pang, Vladimir Radenkovic, Prudencio Tossou, Therence Bois, Andrew Wedlake, and Francesco Di Giovanni. Nesso-1: Accelerating open-source binding affinity predictions. *bioRxiv*, 2026. doi: 10.64898/2026.08.01.742196. URL https://www.biorxiv.org/content/10.64898/2026.08.01.742196v1. 3, 5, 25

Marta M Stepniewska-Dziubinska, Piotr Zielenkiewicz, and Pawel Siedlecki. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics*, 34(21):3666–3674, 2018. 3

Minyi Su, Qifan Yang, Yu Du, Guoqin Feng, Zhihai Liu, Yan Li, and Renxiao Wang. Comparative assessment of scoring functions: the casf-2016 update. *Journal of chemical information and modeling*, 59(2):895–913, 2019. 18

Zhoutong Sun, Qian Liu, Ge Qu, Yan Feng, and Manfred T Reetz. Utility of b-factors in protein science: interpreting rigidity, flexibility, and internal motion and engineering thermostability. *Chemical reviews*, 119(3):1626–1665, 2019. 1

Patricia Suriana, Joshua A Rackers, Ewa M Nowara, Pedro O Pinheiro, John M Nicoloudis, and Vishnu Sresht. Beyond atoms: Evaluating electron density representation for 3d molecular learning. *arXiv preprint arXiv:2511.21900*, 2025. 2, 3, 4

Raphael JL Townshend, Martin Vögele, Patricia Suriana, Alexander Derry, Alexander Powers, Yianni Laloudakis, Sidhika Balachandar, Bowen Jing, Brandon Anderson, Stephan Eismann, et al. Atom3d: Tasks on molecules in three dimensions. *arXiv preprint arXiv:2012.04035*, 2020. 27

Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010. 27

Ísak Valsson, Matthew T Warren, Charlotte M Deane, Aniket Magarkar, Garrett M Morris, and Philip C Biggin. Narrowing the gap between machine learning scoring functions and free energy perturbation using augmented data. *Communications Chemistry*, 8(1):41, 2025. 3, 25

Gydo CP van Zundert, Brandi M Hudson, Saulo HP de Oliveira, Daniel A Keedy, Rasmus Fonseca, Amelie Heliou, Pooja Suresh, Kenneth Borrelli, Tyler Day, James S Fraser, et al. qfit-ligand reveals widespread conformational heterogeneity of drug-like molecules in x-ray electron density maps. *Journal of Medicinal Chemistry*, 61(24):11183, 2018. 1

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 21

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020. 31

W Patrick Walters and Mark Murcko. Assessing the impact of generative ai on medicinal chemistry. *Nature biotechnology*, 38(2):143–145, 2020. 8, 29

Junmei Wang, Piotr Cieplak, and Peter A Kollman. How well does a restrained electrostatic potential (resp) model perform in calculating conformational energies of organic and biological molecules? *Journal of computational chemistry*, 21(12):1049–1074, 2000. 27

Lvwei Wang, Rong Bai, Xiaoxuan Shi, Wei Zhang, Yinuo Cui, Xiaoman Wang, Cheng Wang, Haoyu Chang, Yingsheng Zhang, Jielong Zhou, et al. A pocket-based 3d molecule generative model fueled by experimental electron density. *Scientific reports*, 12(1):15100, 2022. 3

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. 3, 18

Junhao Wei, Baili Lu, Zhenhong Peng, Wanyan Li, Zhirong Huang, Yanxiao Li, Yifu Zhao, Dexing Yao, Haochen Li, Xudong Ye, et al. Honestaffinity: Leak-aware evaluation of protein and pocket priors for binding affinity prediction. *arXiv preprint arXiv:2606.03422*, 2026. 3, 18, 25

David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988. 7, 29

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9643–9653. IEEE, 2022. 4

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. In *The eleventh international conference on learning representations*, 2023. 25

APPENDIX

A Data processing

A.1 Dataset sources . . . 18

A.2 Preprocessing procedure . . . 19

B CoDE details

B.1 Architecture . . . 21

B.2 Pre-training configuration . . . 22

B.3 Downstream task adaptation . . . 23

B.4 Motivational observation configuration . . . 23

B.5 Ablation experiments configuration . . . 24

C Baseline details

C.1 Binding affinity regression . . . 25

C.2 Pocket-conditioned ligand generation . . . 26

D Evaluation details

D.1 Binding affinity regression . . . 27

D.2 Vina docking . . . 27

D.3 Generation quality . . . 28

D.4 Novelty computation . . . 29

D.5 PoseCheck . . . 30

D.6 Molecular structure . . . 31

D.7 PoseBusters . . . 31

D.8 Rigid-fragment consistency . . . 32

E Additional results

E.1 Binding affinity regression . . . 33

E.2 Vina docking . . . 34

E.3 Qualitative visualization . . . 36

E.4 Macrocycles in ligand generation . . . 37

E.5 PoseCheck . . . 38

E.6 Molecular structure . . . 40

E.7 Rigid fragment consistency . . . 43

E.8 PoseBusters . . . 44

A Data processing

A.1 Dataset sources

We use five datasets in this work, summarized in Table 4. PLINDER is used for pre-training, and PDBBind and CASF-2016 for binding-affinity regression with the LP-PDBBind split, and CrossDocked 2020 for pocket-conditioned generation. All electron density maps are obtained from PDBe.

Table 4: **Datasets used in this work.** Split counts are reported as train/validation/test.

Dataset	Source size	Used after preprocessing	Role
PLINDER	449,383 systems (1,357,906 instances)	112,000 / 100 / -	Encoder pre-training
PDBBind v2020	19,443 complexes	12,674 voxelized complexes	Source of affinity-labelled complexes
LP-PDBBind	18,795 split-assigned complexes	3,850 / 817 / 1,320	Affinity regression
CASF-2016	285 complexes	- / - / 92	Clean held-out affinity evaluation
CrossDocked2020	22.5M poses (processed into 99,881 / / 100)	78,512 / 100 / 79	Pocket-conditioned ligand generation

PLINDER. PLINDER (Durairaj et al., 2024) is our pre-training source. It curates the Protein Data Bank (PDB) as of 2024-04-09 into 449383 protein–ligand interaction (PLI) systems, each annotated with over 500 fields and with pocket-, ligand-, and interaction-level similarity. It also links 98473 experimental *apo* structures and 205300 AlphaFold2 (Jumper et al., 2021) predictions to their *holo* entries. We use only holo X-ray systems, drawn from all PLINDER splits, since our downstream tasks define their own test sets. Every downstream validation and test complex was excluded.

PDBBind 2020. PDBBind (Wang et al., 2005; Liu et al., 2015) provides our affinity labels for the binding affinity regression downstream task. It pairs protein–ligand crystal structures with measured binding constants ($K_d, K_i, \mathrm{IC}_{50}$). The 2020 release consists of three sets: (i) the general set consisting of 19443 complexes, (ii) the refined set consisting of 5316 complexes that additionally satisfy stricter criteria such as a crystallographic resolution of 2.5 Å or better and K_i, K_d only records, and (iii) a core set of 285 complexes. The general, refined, core set was designed around structure quality rather than around similarity control. We therefore do not use the conventional split that trains on the refined set excluding the core set and evaluates on the core set.

LP-PDBBind. LP-PDBBind (Li et al., 2026) re-splits PDBBind v2020 so that test complexes are dissimilar to training complexes. It showed that the standard sets share highly similar proteins and ligands, so gains on the core set overstate generalization. LP-PDBBind bounds protein and ligand similarity across 11513 training, 2422 validation, and three nested test tiers of 4286 (CL1), 2179 (CL2), and 1348 (CL3) complexes. CL1 removes covalent binders, ligands with rare elements, and complexes with steric clashes; CL2 keeps consistent K_d/K_i measurements; and CL3 keeps only refined- and core-set complexes. We further drop entries that RDKit cannot parse.

CASF-2016. CASF-2016 (Su et al., 2019) serves as an additional held-out affinity test. It is built on the PDBBind core set, with 57 target clusters of five ligands that bind the same protein across a wide affinity range. Following Wei et al. (2026), we probe learned representations, restricted to the 92 complexes that have a density map and appear neither in pre-training nor in the LP-PDBBind training or validation split.

CrossDocked2020. CrossDocked2020 (Francoeur et al., 2020) is our pocket-conditioned ligand generation benchmark. It docks ligands into similar binding pockets across the PDB, yielding 22.5 million poses grouped by binding-site similarity. Following Luo et al. (2021); Schneuing et al. (2024); Guan et al. (2023; 2024); Pinheiro et al. (2023), we filter poses by RMSD and sequence identity, leaving 100,000 training pairs and 100 test pockets from held-out clusters. For each receptor, we retrieve the density map of the crystal structure it derives from by its PDB ID and align the receptor to that frame. The map therefore shows the native crystal ligand of that entry rather than the docked ligand pose, which is computational and has no density of its own.

Electron density maps. All maps are the likelihood-weighted $2mF_o$-DF_c maps that PDBe serves as pre-computed CCP4 grids, paired with the deposited coordinates:

```
https://www.ebi.ac.uk/pdbe/entry-files/{pdb_id}.ccp4
```

We use them as deposited, without re-refinement, re-phasing, or resolution and B-factor cutoffs.

A.2 PREPROCESSING PROCEDURE

Table 5: **Dataset preprocessing.** Counts are cumulative. PLINDER counts refer to ligand instances, whereas PDBBind counts refer to complexes.

PLINDER	
All annotated ligand instances	1357906
X-ray diffraction	1046290
Proper ligand; excluding ions, artifacts, and cofactors	300276
Valid and RDKit-loadable	289986
6–50 ligand heavy atoms	259509
Non-covalent	197369
Downstream validation/test PDB entries excluded	166448
No unresolved ligand heavy atoms	153270
Ligand RSCC ≥ 0.8	113976
Final grid-eligible manifest	112733
Encoder train / validation subset	112000 / 100
PDBBind / LP-PDBBind	
All PDBBind v2020 complexes	19443
Valid atomic grid and experimental map	12674
Ligand RSCC computable	11601
Ligand and pocket RSCCs ≥ 0.8	10926
Non-covalent	10424
Assigned to an LP-PDBBind split	10128
K_d or K_i label	5987
Final train / validation / test split	3850 / 817 / 1320

PLINDER filtering. We keep X-ray ligand instances whose density resolves the ligand. Each instance must have a proper, RDKit-loadable, non-covalent ligand with 6–50 heavy atoms, no unresolved ligand heavy atoms, and a ligand RSCC of at least 0.8 (Fusani et al., 2020). We impose no resolution or pocket-RSCC cutoff, since the ligand RSCC already gates map quality at the binding site. Of the 112733 instances that pass map retrieval and grid construction, we use 112000 for training and 100 for validation. Atoms outside the 11 atom-type channels are omitted during voxelization rather than rejecting the complex.

Leakage control. No downstream validation or test complex appears in pre-training. We remove the held-out PDB entries of CrossDocked2020 and PDBBind by exact PDB ID. We then verify that none of the LP-PDBBind validation (817) or test (1,320) complexes, nor the 92 CASF-2016 complexes, is in the pre-training set. 250 of the 3,850 LP-PDBBind training complexes do appear, which does not affect evaluation since pre-training uses no affinity labels.

Frame alignment. CrossDocked2020 receptors must be moved into the crystal frame of their maps. They are re-docked copies in an arbitrary frame, so the density and atoms would otherwise disagree by several Ångström. We superpose each receptor onto its deposited structure by a residue-matched Kabsch fit (Kabsch, 1976) and apply the same transform to the ligand. Complexes with too few matched atoms or a fit RMSD above 2 Å are dropped. PLINDER and PDBBind use deposited coordinates and need no alignment.

Grids. Each complex is a 64^3 grid at 0.25 Å spacing, a 16 Å cube centered on the ligand centroid, trilinearly resampled from the CCP4 map. The pocket consists of protein heavy atoms within 10 Å of any ligand heavy atom, cropped to this cube. For pre-training, we instead cache a 116^3 box (29 Å) per complex, stored in 8 bits over $[-2.2, 5.0]$. Its 6.5 Å margin per side lets rotation augmentation resample the map at load time rather than zero-fill a fixed crop.

Density normalization. We map raw density values onto a fixed, heavy-tail-compressed scale shared by all complexes:

$$\rho = \frac{\operatorname{asinh}(\rho_{\text{raw}}/0.5) - \mu_\rho}{\sigma_\rho}, \qquad \mu_\rho = -0.01544, \quad \sigma_\rho = 0.50848, \tag{3}$$

where μ_ρ and σ_ρ are estimated on PLINDER pockets. The gradient magnitude $\|\nabla\rho\|$ is computed on the normalized, resampled grid and standardized within each complex.

Affinity and generation datasets. Both downstream sets keep only complexes with a density map. For affinity regression, we keep voxelizable PDBBind complexes with ligand and pocket RSCCs of at least 0.8, a non-covalent ligand, an LP-PDBBind split assignment, and a K_d or K_i label. This yields 3,850/817/1,320 train/validation/test complexes, plus the 92 CASF-2016 complexes. For generation, maps exist for 78,612 CrossDocked2020 training pairs, split into 78,512 for training and 100 for validation, and for 79 of the 100 test pockets. Table 5 lists the cumulative filters.

B CoDE DETAILS

In this section, we explain the details of the CoDE encoder and how it is used.

- Encoder configuration and model hyperparameters: Appendix B.1
- Pre-training configuration: Appendix B.2
- Downstream task configuration: Appendix B.3
- Motivational analysis: Appendix B.4
- Ablation experiment configuration: Appendix B.5

B.1 ARCHITECTURE

Encoder. CoDE is a 99.47M-parameter Channel-ViT (Bao et al., 2023) that tokenizes each channel group separately. The 13 input channels form three groups of sizes $(7, 4, 2)$: ligand atoms, pocket atoms, and the density and gradient magnitude pair. Each group has its own 8^3 convolution, which embeds the 64^3 grid into 512 patch tokens of width 640. Every token receives a learnable positional embedding, shared across groups, and a learnable embedding of its group. The resulting 1536 tokens pass through 18 pre-LayerNorm (Ba et al., 2016) Transformer blocks (Vaswani et al., 2017) with 10 heads, an MLP ratio of 4, GELU activations (Hendrycks & Gimpel, 2016), and dropout of 0.1, followed by a final LayerNorm. Since all groups attend to one another, the encoder can relate density to atoms at every location.

Voxel projection and reconstruction head. The encoder includes a linear voxel projection. For each patch, the final tokens of its three groups are averaged and linearly projected to a 16-channel 8^3 feature cube. A separate two-layer reconstruction MLP head with hidden width 67 maps the 16-channel cube of each patch to the 13 input channels. The reconstruction MLP head is used only for pre-training and discarded afterwards.

Input construction. Each input has 13 channels: 7 ligand elements (C, O, N, S, F, Cl, P), 4 pocket elements (C, O, N, S), ρ, and $\|\nabla\rho\|$. Atoms are rendered as soft occupancies following Orlando et al. (2022), with R_i the van der Waals radius of atom i and $\sigma = 0.93$:

$$g_i(v) = \exp\left(-\frac{\|v - r_i\|^2}{\sigma^2 R_i^2}\right), \qquad a_c(v) = 1 - \prod_{i \in c}\big(1 - g_i(v)\big), \tag{4}$$

so each element channel c is a soft union of its atoms, bounded in $[0, 1]$. Pre-training uses van der Waals radii for both ligand and pocket, since a uniform ligand radius would reveal the ligand/pocket split through blob size. The generation model keeps VoxBind's uniform 0.5 Å ligand radius, so that our conditioning is a drop-in replacement.

Density and gradient channels. The density and gradient channels form their own group, so removing them leaves the atomic groups unchanged. ρ is the normalized map of Appendix A.2, and its gradient magnitude highlights atom boundaries and partially occupied regions. We compute the gradient in-model by central differences on the voxel grid,

$$\partial_x\rho(v) = \tfrac{1}{2}\big(\rho(v + \mathbf{e}_x) - \rho(v - \mathbf{e}_x)\big), \qquad \|\nabla\rho\|(v) = \sqrt{\partial_x\rho(v)^2 + \partial_y\rho(v)^2 + \partial_z\rho(v)^2}, \tag{5}$$

where $\mathbf{e}_x$ is a one-voxel step along x, ∂_y and ∂_z are defined likewise, and boundary voxels are replicated. We then standardize the channel within each complex,

$$\mathcal{Z}[f](v) = \frac{f(v) - \mu_f}{\sigma_f}, \tag{6}$$

where μ_f and σ_f are the mean and standard deviation of f over the grid, so the input channel is $\mathcal{Z}[\|\nabla\rho\|]$ and the voxel spacing cancels. Since the gradient is recomputed after each augmentation, it always matches the resampled density. Dropping one channel leaves a group of size one, and dropping both recovers CoE.

B.2 PRE-TRAINING CONFIGURATION

Masking. We mask 75% of the blocks, biased toward those occupied by atoms. Each of the 512 blocks is scored by its total atomic occupancy plus Gaussian noise with standard deviation $\tau = 1$ times the within-complex standard deviation of the scores. The 384 highest-scoring blocks are masked, and the same spatial mask is applied to all channels.

Loss weights. The channel weights balance sparse elements against the dense density channels. For each atomic channel, w_c is inversely proportional to its occupied-voxel frequency, estimated on 2000 training complexes with an occupancy threshold of 0.05 and capped at a largest-to-smallest ratio of 30. ρ and $\|\nabla\rho\|$ receive unnormalized weights of 0.1, and all 13 weights are then normalized to sum to 13. The voxel weight is $\omega_{c,v} = 10$ where an atomic channel exceeds 0.05 and 1 otherwise.

Optimization. We pre-train for 100 epochs with the settings of Table 6 and use the best valid checkpoint for the main affinity results. Atoms and density are rotated and translated jointly, and rotation resamples the cached box at the augmented pose, $\rho_{\text{aug}}(v) = \rho_{\text{box}}(R^\top(v - t))$, rather than shifting a fixed crop. All results use the exponential moving average (EMA) of the parameters.

Table 6: **CODE hyperparameters.** Pre-training and affinity-probing settings of the encoder behind the main affinity results.

Encoder	
Patch size	8^3 voxels (2 Å), 512 patches per group
Tokens	1536 (3 groups $\times$ 512)
Width / depth / heads	640 / 18 / 10
MLP ratio / activation	4 / GELU
Dropout	0.1
Positional embedding	learnable, shared across groups, plus a learnable group embedding
Voxel projection	group average, linear to 16 channels
Reconstruction MLP head	per patch, 2 layers, hidden 67
Parameters	99.47M
Pre-training	
Data	PLINDER, 112000 train / 100 validation
Mask ratio / strategy	0.75 / atom-biased, τ = 1
Mask block	8^3 voxels, shared across channels
Channel weight w_c	inverse occupied-voxel frequency, max ratio 30; 0.1 for ρ and $\|\nabla\rho\|$
Voxel weight $\omega_{c,v}$	10 where atomic occupancy $>$ 0.05, else 1
Optimizer	AdamW, (β_1, β_2) = (0.99, 0.999), ϵ = 0
Learning rate / schedule	10^{-4} / constant
Weight decay	0.05
Batch size	128 (4 per GPU $\times$ 8 accumulation steps $\times$ 4 GPUs)
Gradient clipping	1.0
Precision	bfloat16
EMA decay	0.999
Epochs / checkpoint	100 / best validation checkpoint (epoch 25)
Augmentation	random rotation, translation in $[-1, 1]$ Å per axis
Seed	42
Affinity probe	
Pooling	mean over 1536 final tokens (640-d)
Head	MLP 640 $\to$ 128 $\to$ 1, SiLU, dropout 0.1
Optimizer	Adam, learning rate 10^{-3}, weight decay 10^{-4}
Batch size / epochs	64 / at most 200, early stopping after 30
Model selection	best validation Spearman correlation
Seeds	0, 1, 2, 3, 4

B.3 Downstream task adaptation

Binding-affinity probing. We fit a small MLP on frozen, mean-pooled encoder tokens. The 1536 final tokens are averaged into a 640-dimensional complex embedding, and a two-layer MLP regresses affinity from it with mean-squared error (Table 6). We keep the checkpoint with the highest validation Spearman correlation and use no test-time rotation averaging.

Pocket-conditioned ligand generation. We keep the VoxBind (Pinheiro et al., 2024) walk-jump sampler and change only its pocket conditioning. The frozen encoder reads the reference complex with the experimental density. Its patch tokens are averaged over channel groups, and the encoder's linear output layer, trained with it during pre-training, maps each token to the features of the voxels in its patch, restoring the full grid as x_{dens}. This output is fused with VoxBind's ligand and pocket features x through a zero-initialized $1 \times 1 \times 1$ convolution, $x \leftarrow x + W[x \| x_{\text{dens}}]$. Only the denoiser and W receive gradients and are updated.

Generation optimization. The denoiser follows VoxBind: a 3D U-Net with base width 32, channel multipliers $[1, 2, 2, 4]$, two residual blocks per level, and attention at the two coarsest levels (111.6M parameters). We train it for 350 epochs from scratch with AdamW, a constant learning rate of 10^{-5}, weight decay 10^{-2}, noise level $\sigma = 0.9$, and a batch size of 32 per GPU on four H200 or RTX PRO 6000 Blackwell GPUs. Sampling uses walk-jump with $\sigma = 0.9$, friction $\gamma = 1$, step $\delta = \sigma/2$, and 400 warm-up and 100 sampling steps from a de novo initialization, with the pocket masked out of both noise and score. Grids are decoded to atoms by non-maximum-suppressed peak picking at threshold 0.25, followed by L-BFGS refinement against the predicted grid.

B.4 Motivational observation configuration

Complex contact B factor. The complex contact B factor summarizes how disordered the protein is where it touches the ligand. It is the mean B factor of protein heavy atoms within 5 Å of any ligand heavy atom,

$$B_{\text{complex}} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} B_i, \tag{7}$$

where $\mathcal{I}$ is this set of protein heavy atoms and B_i is the deposited B factor of atom i. It is the complex-level decoding target and defines the affinity bins below.

Decoding. In Figure 2a, we decode B factors from frozen representations of 732 CL3 test complexes without fine-tuning the encoders. At the complex level, the target is B_{complex}, predicted from one pooled embedding per complex. At the residue level, the target is the mean B factor of the evaluated atoms in each residue; at the atom level, it is the atom's B factor. These two levels use local tokens from the same 48,591 protein atoms covered by all compared token models. We train MLP decoders with five complex-disjoint folds and average predictions across three seeds. Residue and atom decoders are trained on B factors standardized within each complex, which preserves their within-complex ranks. We report Spearman correlation across complexes for the complex-level target and the mean of within-complex Spearman correlations for residue and atom targets. Error bars are 95% confidence intervals from 2,000 bootstrap resamples of complexes.

Binding affinity grouping. In Figure 2b, we compare the affinity gain within bins of B_{complex} in the CL3 setting of Table 1, where probes are trained and evaluated on the CL3-filtered LP-PDBBind split (1559/410/733). We split the 733 test complexes into $B_{\text{complex}} < 20$, [20,30), and ≥ 30 Å^2, which hold 330, 222, and 181 complexes. Within each bin, we compute the Spearman correlation of CoDE and CoE for each of five probe seeds and average the per-seed differences. The 95% intervals are percentile intervals from 8000 paired bootstrap resamples of complexes within each bin. The analysis is exploratory, without multiplicity correction.

B.5 ABLATION EXPERIMENTS CONFIGURATION

Shared recipe. All ablations change only the input or target channels of the CODE recipe. They share the architecture, PLINDER data, masking ratio of 0.75, atom-biased masking, channel weighting, optimizer, EMA, and seeds. Each row of Table 2 averages the five-seed CL3 probe means and standard deviations.

Channel selection. Each variant removes channels or moves them to the target only.

- CODE: 13 input channels with groups $(7, 4, 2)$
- w.o. D, G (COE): 11 atomic channels with groups $(7, 4)$
- COE + D recon.: 11 input channels, but all 13 channels are reconstructed, so density enters only as a target
- w.o. D input: atomic channels and $\|\nabla\rho\|$ of the real density, with groups $(7, 4, 1)$
- w.o. G input: atomic channels and ρ, with groups $(7, 4, 1)$

Noisy pseudo-density. The noise control keeps the channel count and value distribution of the density but removes its spatial structure. We build one histogram over all voxels of the 112000 pretraining crops and draw every voxel of ρ independently from it. The noise of each complex is fixed by a per-complex seed, so it stays the same across epochs and is never mixed between complexes. $\|\nabla\rho\|$ is recomputed from the noise with the same function and per-complex standardization as for the real density.

Atom-blob gradients. This control tests whether a gradient channel helps without experimental density. We replace the density and its gradient with the gradient magnitude of the atomic occupancies,

$$\|\nabla a\|(v) = \mathcal{Z}\Big[\sum_{c=1}^{11} \|\nabla a_c\|\Big](v), \tag{8}$$

where each $\|\nabla a_c\|$ uses the central differences of Equation (5) on atomic channel c, and $\mathcal{Z}$ is the standardization of Equation (6). The row thus carries coordinates plus $\|\nabla a\|$ with groups $(7, 4, 1)$ and no density.

C BASELINE DETAILS

C.1 BINDING AFFINITY REGRESSION

We compare against a broad set of baselines covering sequence-based, structure-based, and pre-trained protein–ligand complex representation models. For methods that provide pre-trained encoders, we follow the same probing protocol as our model and train a lightweight prediction head MLP on top of the frozen representations. For methods without publicly available pre-trained representations, we train the model on the LP-PDBbind training split using its original or recommended training objective as possible. The baselines considered in our experiments are as follows:

- **DeepDTA** (Öztürk et al., 2018) uses 1D convolutional networks on a ligand SMILES and the amino acid sequence of the target protein.
- **MolTrans** (Huang et al., 2021) uses Transformer encoders to model substructural tokens extracted from the ligand SMILES and the full protein sequence, without using 3D coordinates.
- **HonestAffinity** (Wei et al., 2026) uses frozen ESM-2 650M embeddings (Lin et al., 2023) for the full protein sequence, with a binary indicator specifying pocket residues. It does not use 3D coordinates as input.
- **Nesso-1** (Shenoy et al., 2026) takes the full protein sequence and a ligand represented by SMILES or molecular structure, following a co-folding style architecture stemming from **Boltz-2** (Passaro et al., 2025). It does not require a pre-defined pocket or experimentally determined protein coordinates as input.
- **HBGSA** (Kong et al., 2026) uses the 3D protein–ligand binding complex, including explicit coordinates of the protein–ligand hydrogen bonds. It models the spatial topology with a graph neural network and self-attention. Additionally, it uses the binding-site structure rather than the full protein sequence.
- **EGNN** (Duval et al., 2023) represents the protein pocket and ligands as atom features and 3D coordinates, and process them using an E(3)-equivariant graph neural network.
- **EGNN + TargetDiff** (Guan et al., 2023) uses the same protein-pocket and ligand atoms with 3D coordinates as above, while initializing the encoder with TargetDiff pre-trained for protein-conditioned 3D ligand generation.
- **GET** (Kong et al., 2023) represents the 3D protein–ligand complex as a hierarchical geometric graph and applies an E(3)-equivariant Transformer to model intra- and intermolecular interactions.
- **CheapNet** (Lim et al., 2025) uses the protein binding pocket and ligand with their 3D atomic structures and learns hierarchical atom- and cluster-level representations with cross-molecular interactions.
- **IPNet** (Huang et al., 2024) encodes the 3D protein pocket and ligand with SE(3)-equivariant networks and cross-attention layers, pre-trained with binding affinity supervision on PDBBind 2016 (Liu et al., 2015). Additionally, it is used as the interaction prior component of IPDiff.
- **AEV-PLIG** (Valsson et al., 2025) uses atoms from the protein binding site and the ligand, along with their 3D coordinates, represented as atomic environment vectors and a protein–ligand interaction graph.
- **GeoSSL** (Liu et al., 2022) pre-trains molecular representations from 3D atomic coordinates using SE(3)-invariant denoising distance matching. Unlike protein–ligand pre-training methods, its objective operates on individual molecular structures without explicit protein–ligand interactions.
- **DSMBind** (Jin et al., 2023) uses 3D protein–ligand complex and learns a binding-energy fucntion through SE(3) denoising score matching. Its protein representation is structure-based, rather than full amino-acid sequence-based.
- **BindNet** (Feng et al., 2024) uses Uni-Mol (Zhou et al., 2023) encoders for the 3D protein pocket and ligand, taking atom types and coordinates as input and pre-trains on cross-molecular interactions.

- **ProFSA** (Gao et al., 2024) encodes the 3D protein pocket with atom types and coordinates, and pre-trains it by aligning pocket representations with Uni-Mol embeddings of nearby protein fragments treated as pseudo-ligands.

C.2 POCKET-CONDITIONED LIGAND GENERATION

We compare our model against autoregressive, diffusion-based, and voxel-based pocket-conditioned molecular generation methods. We use the publicly released generated samples or official implementations with their recommended configurations. The baselines considered in our experiments are as follows:

- **AR** (Luo et al., 2021) sequentially samples ligand atoms from a learned spatial distribution conditioned on the protein pocket.
- **Pocket2Mol** (Peng et al., 2022) uses an E(3)-equivariant network to autoregressively generate atom positions, types, and bonds.
- **DiffSBDD** (Schneuing et al., 2024) generates ligand point clouds by jointly denoising atom positions and types conditioned on the pocket.
- **TargetDiff** (Guan et al., 2023) uses an equivariant diffusion model to jointly generate ligand coordinates and atom types through protein–ligand interactions.
- **DecompDiff** (Guan et al., 2024) decomposes ligands into arms and scaffolds, then generates them using separate diffusion priors.
- **VoxBind** (Pinheiro et al., 2024) represents molecules as atomic-density voxel grids, and generates ligands through conditional voxel denoising.
- **FuncBind** (Kirchmeyer et al., 2026) represents molecules as continuous neural fields for target-conditioned all-atom generation, trained on three modalities: small molecules, macrocyclic peptides(MCPs), and antibody complementarity-determining region (CDR) loops.

Note that our model is built on top of VoxBind, where the original pocket representation is augmented with the representation obtained from our density-pre-trained encoder. Unless otherwise stated, the remaining generative architecture and sampling procedure follow the original VoxBind implementation.

D EVALUATION DETAILS

D.1 BINDING AFFINITY REGRESSION

We evaluate binding affinity prediction by Pearson correlation, Spearman rank correlation, and root mean squared error (RMSE), following Townshend et al. (2020). Let y_i denote the ground-truth affinity in pK units, $\hat{y}_i$ the predicted affinity, and N the number of protein–ligand complexes.

- **Pearson correlation.** The linear correlation between predicted and measured affinities:

$$r = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}},$$

  where $\bar{y}$ and $\bar{\hat{y}}$ are the mean ground-truth and predicted affinities. Higher is better.

- **Spearman correlation.** The monotonic agreement between the rankings of predicted and ground-truth affinities:

$$\rho = \mathrm{corr}\left(\mathrm{rank}(y), \mathrm{rank}(\hat{y})\right) = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)},$$

  where the second form holds without ties and d_i is the rank difference of the i-th complex. Higher is better.

- **Root mean squared error (RMSE).** The absolute prediction error:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}.$$

  Lower is better.

For all probing experiments in the paper, we report the mean and standard deviation over five seeds $\{0, 1, 2, 3, 4\}$, unless otherwise stated. For CASF 2016, we compute 90% bias-corrected and accelerated (BCa) confidence intervals from $10,000$ bootstrap resamples of the test complexes, using random seed 42 (Efron, 1987). For methods trained with multiple seeds, the same bootstrap indices are applied to every seed, and the metric is averaged over seeds with each resample.

D.2 VINA DOCKING

We score every generated molecule with AutoDock Vina 1.2 (Trott & Olson, 2010; Eberhardt et al., 2021), following Guan et al. (2023); Pinheiro et al. (2024). AutoDock configuration is identical for every method as follows:

- Version: `1.2.2`
- Scoring function: `vina`
- Exhaustiveness: `32`
- Min RMSD: `1.0` Å (default)
- Seed: `0` (random seed per run)
- CPU: `1` per Vina instance, `20` instances in parallel
- Box size: ligand extent plus `5` Å per axis, $(\mathbf{x}_{\max} - \mathbf{x}_{\min}) + 5$Å
- Box center: center of ligand bounding box, $(\mathbf{x}_{\max} + \mathbf{x}_{\min})/2$
- Receptor: full CrossDocked2020 receptor, prepared with `PDB2PQR 3.6.1` (Dolinsky et al., 2007) with `AMBER` force field (Wang et al., 2000) and AutoDockTools (Morris et al., 2009)
- Ligand: loaded with RDKit `AddHs(addCoords=True)`, then processed with Meeko `0.1.dev3` `MoleculePreparation()` (Santos-Martins et al., 2025) with default configuration

Configuration. The box follows the molecule being scored, the generated pose for a sample, and the crystal pose for the reference ligand. The mean box is $12.1 \times 12.0 \times 11.8$ Å^3, and the per-axis edge spans $7.5 - 18.3$ Å. `PDB2PQR` assigns the standard protonation states at pH 7.0 without PROPKA titration. Afterwards, Gasteiger charges are added, non-polar hydrogens are removed, and waters and non-standard residues are removed without structural repair. Ligand preparation keeps the generated coordinates with no re-embedding, and Meeko keeps macrocycles and amide bonds rigid. Since a seed of `0` draws a fresh seed per run, Vina Dock carries small noise while Vina Score and Vina Min are deterministic.

Metrics. We generate 100 molecules per pocket and report the average and median over all generated molecules. Lower Vina scores indicate more favorable binding.

- **Vina Score**: Energy of the generated pose as is (`score()`), without search or relaxation
- **Vina Min**: Energy after local minimization on the generated pose (`optimize()`)
- **Vina Dock**: Energy of the pose from a global docking search within the box (`dock()`)
- **High affinity**: Per pocket computed, the fraction of generated molecules with $E^{(i)}_{\text{dock}} < E^{\text{ref}}_{\text{dock}}$, where $E^{\text{ref}_{\text{dock}}}$ is the Vina Dock energy of the reference ligand. We report its average and median over pockets.

D.3 Generation quality

We evaluate sample quality following the common protocols for structure-based drug design in Guan et al. (2023), with all properties computed by RDKit (Landrum et al., 2025). The configuration is as follows:

- Version: `RDKit 2026.03.3`
- Valid molecules: sanitized by RDKit, single connected component
- Aggregation: within each pocket, then over pockets

Invalid samples are discarded before every other metric, and n in Table 3 counts the remaining molecules only. VoxBind-based methods use the original VoxBind sampler, which keeps distinct valid molecules per pocket until it collects 100 or reaches 500 reconstruction attempts.

Metrics. Higher is better for every metric below.

- **Validity**: Fraction of drawn samples that are valid, $N_{\text{valid}}/N_{\text{raw}}$
- **Uniqueness**: Fraction of distinct canonical SMILES among the valid molecules of a pocket, $|\{\text{SMILES}(m_i)\}|/N_{\text{valid}}$
- **QED**: Quantitative estimate of drug-likeness (Bickerton et al., 2012), a score in $[0, 1]$ summarizing physicochemical properties
- **SA**: Synthetic accessibility score (Ertl & Schuffenhauer, 2009) in the normalized form, where a higher value means easier to synthesize
- **Diversity**: One minus the mean pairwise Tanimoto similarity among the valid molecules of a pocket,

$$\text{Diversity} = 1 - \frac{2}{N_{\text{valid}}(N_{\text{valid}} - 1)} \sum_{i<j} \text{Tanimoto}(f_i, f_j)$$

N_{raw} and N_{valid} denote the number of drawn and valid molecules of a pocket, m_i is the i-th valid molecule, and f_i is its fingerprint.

D.4 NOVELTY COMPUTATION

We measure how much of each method's output replays training chemistry, following Polykovskiy et al. (2020); Brown et al. (2019); Walters & Murcko (2020), with the following configuration:

- Comparison set: CrossDocked2020 training split, `100,000` ligands
- Canonicalization: isomeric SMILES (Weininger, 1988) with RDKit
- Scaffold: Bemis-Murcko (Bemis & Murcko, 1996) scaffold, used with the function `MurckoScaffold.GetScaffoldForMol`
- SNN fingerprint: ECFP4 (Rogers & Hahn, 2010), radius `2`, `2048` bits
- Aggregation: within each pocket, then over pockets

Figure 4c reports novelty and SNN against the training split. Figure 8 separately compares each generated molecule with the reference ligand of its corresponding test pocket.

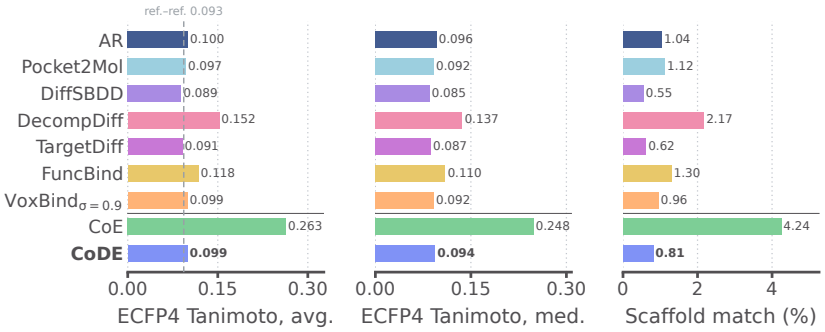


Figure 8: **Similarity to the test reference ligand.** Each generated molecule is compared with the crystal ligand of its pocket. The per-pocket mean (left) and median (middle) of the ECFP4 Tanimoto similarity are averaged over pockets. On the right is the percentage of molecules whose Bemis–Murcko scaffold equals the reference's. The dashed line marks 0.093, the mean similarity of the 79 crystal ligands to one another, which is the level of two unrelated ligands on this fingerprint.

Metrics. We use the following metrics to measure the distances to the training and the test set reference ligand similarity.

- **Molecule novelty**: Fraction of generated molecules whose canonical SMILES does not exist in the training split.
- **Scaffold novelty**: Same fraction computed with the Bemis-Murcko scaffolds
- **SNN**: Mean over the pocket of the largest Tanimoto similarity (Bajusz et al., 2015) of each generated molecule g to any training ligand t,

$$\mathrm{SNN} = \frac{1}{|G|} \sum_{g \in G} \max_{t \in T} \mathrm{Tanimoto}\big(f(g), f(t)\big),$$

  which catches the near-duplicates that exact matching misses.

G denotes the set of generated molecules of a pocket, T is the comparison set from the training data, and $f(\cdot)$ is the ECFP4 fingerprint.

D.5 POSECHECK

We use PoseCheck (Harris et al., 2023) to check the realistic structures of the generated molecules, with (i) strain energy, (ii) steric clashes, (iii) re-docking, and (iv) protein–ligand interactions of the generated poses. The configuration is as follows:

- Version: `1.3.1`
- Receptor: `10` Å pocket crop, protonated with reduce
- Clash tolerance: `0.5` Å (default)
- Strain conformers: `50` per molecule, embedded with seed `0`
- Strain timeout: `50` seconds per molecule
- Interactions: ProLIF `2.2.0` `Fingerprint()` with default configuration
- Re-docking: Vina settings stated in Appendix D.2
- Re-docking set: `13` molecules per pockets in generation order, over `78` pockets
- Re-docking RMSD: RDKit `rdMolAlign.CalcRMS` on heavy atoms, without superposition

Water and HETATM records are removed from the pocket before protonation. Molecules whose strain does not converge within the timeout are dropped. One excluded re-docking pocket has a chain gap that `PDB2PQR` cannot protonate, for all methods. We use PoseCheck `1.3.1` since prior releases differ in the strain energy definition, which uses the energy difference between the generated pose and the relaxed pose, rather than the relaxed pose to searched conformer poses. Our values are internally consistent, but not comparable with the numbers printed in Harris et al. (2023).

Metrics. Lower is better for clashes, strain, and re-docking displacement. For protein–ligand interactions, values similar to the reference are considered better.

- **Steric clashes**: Number of ligand atoms per molecule with ligand and protein atoms overlapping
- **Strain energy**: UFF energy of the relaxed pose relative to the best conformer,

$$E_{\text{strain}} = E\big(\text{UFF}_{\text{constrained}}(x)\big) - \min\Big(\min_{k=1}^{50} E\big(\text{UFF}(c_k)\big),\ E\big(\text{UFF}_{\text{free}}(x)\big)\Big),$$

  where x is relaxed under a 0.1 Å per atom constraint and c_k are the re-embedded conformers.
- **Protein–ligand interactions**: Number of ProLIF fingerprint bits per molecule for van der Waals contacts, hydrophobic contacts, and hydrogen bond donors and acceptors.
- **Re-docking RMSD**: Heavy atom RMSE between the generated pose and the top Vina pose. It is measured without superposition. Focuses on the displacement in the pocket.

d_{ij} denotes the distance between ligand atom i and the receptor atom j, and R^{vdw} denotes the van der Waals radius. Note that the crystal ligands re-docked by the same protocol result in a median of 2.69 Å, with 42.3% within 2Å.

D.6 MOLECULAR STRUCTURE

We compare the distributions of atom types, bond lengths, atom-pair distances, and ring statistics against reference ligands, following Guan et al. (2023); Pinheiro et al. (2024). For each descriptor, we histogram the generated and reference values on a fixed grid and compute the Jensen-Shannon distance (Lin, 1991) with SciPy (Virtanen et al., 2020). The configuration is as follows:

- Version: SciPy `1.15.2`, `2024.03.06`
- Reference: CrossDocked 2020 train and test ligands, `100,100` files
- Atom types: C, N, O, F, P, S, and Cl
- Atom-pair distances: all heavy atom pairs below 12 Å and C–C pairs below 2Å, 100 bins each
- Ring size: rings of each size from 3 to 9
- Rings per molecule: exact counts
- Aromaticity; aromatic share of heavy atoms in 20 bins and of rings in 10 bins

Only valid, single-component molecules are included. We use all CrossDocked 2020 ligands as the reference, as in Guan et al. (2023), since 100 test ligands are too few to estimate the bond-length histogram. 8,765 distinct molecules are used from 100,100 files in total.

D.7 POSEBUSTERS

We use PoseBusters (Buttenschoen et al., 2024) to check the physical validity of the generated poses without relaxation. We run its `dock` configuration, consisting of 22 tests and intended for a pose inside a receptor without a ground-truth ligand. The configuration is as follows:

- Version: PoseBusters `0.6.5`
- Mode: `dock`
- Receptor: `10` Å pocket crop, identical for every method
- Distance geometry: bound thresholds `0.25` for bond lengths, `0.25` for bond angles, and **0.3** for internal clashes, ignoring hydrogens
- Flatness: maximum deviation `0.25` Å for aromatic rings and trigonal-planar double bonds
- Non-flatness: minimum deviation `0.05` Å for non-aromatic six-member rings
- Internal energy: energy-ratio threshold `100` against `50` generated conformers
- Minimum distance: clash cutoff `0.75`, maximum distance `5.0` Å
- Volume overlap: clash cutoff `0.75`, van der Waals scale `0.8`, and `0.5` for inorganic co-factors and waters

`dock` mode tests 22 cases, where two are whether input files are loaded properly, and 20 are graded checks. The pocket crop contains no co-factors or waters, so four co-factor and two water checks pass by construction. The list of test cases is as follows.

- **Overall validity**: Percentage of molecules passing all 22 checks
- **Load (2)**: Valid loading of files for evaluation
- **Length (1)**: Bond lengths within distance-geometry bounds
- **Angle (1)**: Bond angle within distance-geometry bounds
- **Planarity (2)**: Flat aromatic rings and flat double bonds
- **Pucker (1)**: Non-planar non-aromatic six-member rings
- **Valence (4)**: RDKit sanitization, InChI conversion, a single connected component, and no radicals
- **Clash (9)**: Protein proximity, and minimum distance and volume overlap against protein, organic co-factors, inorganic co-factors, and waters
- **Self-clash (2)**: Internal steric clash and internal energy

D.8 RIGID-FRAGMENT CONSISTENCY

We measure whether rigid fragments keep their force-field geometry, following Guan et al. (2023). A rigid fragment, such as a benzene ring, an amide, or a fused bicycle, has one correct shape, so any deviation is a generation error. We relax each molecule with the Merck molecular force field (Halgren, 1996, MMFF) in RDKit, cut it at rotatable bonds, and align each fragment with the Kabsch algorithm (Kabsch, 1976). The configuration is as follows:

- Version: RDKit 2024.03.06
- Force Field: `MMFF`
- Fragmentation of every rotatable bond
- Alignment per fragment: Kabsch superposition on heavy atoms
- Size bins: fragments' heavy atoms
- Reference: crystal ligands of the test pockets, through the same pipeline.

Metrics. Lower is better, indicating that the rigid fragment did not move much.

- **Rigid-fragment RMSD**: Heavy-atom RMSD between a fragment's generated coordinates $\mathbf{x}_i$ and relaxed coordinates $\mathbf{y}_i$ after superposition over rotations R and translation $\mathbf{t}$,

$$\mathrm{RMSD} = \min_{R,\mathbf{t}} \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|R\mathbf{x}_i + \mathbf{t} - \mathbf{y}_i\|^2}\,,$$

  summarized by its median within each size bin.

n denotes the number of heavy atoms in the fragment.

E ADDITIONAL RESULTS

E.1 BINDING AFFINITY REGRESSION

We additionally report the CL1- and CL2-filtered test set and CASF 2016 results. † marks the affinity-leaked reference, included for context only. **Bold** marks the best and underlining marks the next-best; methods whose intervals overlap are marked together. N/A denotes unavailable metrics, and **Pre.** indicates methods with external-data pre-training.

Table 7: **LP-PDBBind binding-affinity prediction on the *CL1-filtered* and *CL2-filtered* test sets.** Results are averaged over five seeds, with zero-shot scorers reported with a single prediction

			CL1 filtered (N=1166)			CL2 filtered (N=1149)		
Input	**Method**	**Pre.**	**Pearson *r***	**Spearman ρ**	**RMSE ↓**	**Pearson *r***	**Spearman ρ**	**RMSE ↓**
Seq.	DeepDTA	✗	0.554 ± 0.008	0.508 ± 0.009	1.493 ± 0.008	0.541 ± 0.013	0.495 ± 0.006	1.491 ± 0.018
	MolTrans	✗	0.499 ± 0.036	0.467 ± 0.034	1.564 ± 0.051	0.513 ± 0.027	0.478 ± 0.026	1.528 ± 0.040
	HonestAffinity	✓	0.477 ± 0.025	0.453 ± 0.025	1.598 ± 0.020	0.483 ± 0.033	0.450 ± 0.034	1.580 ± 0.032
	Nesso-1†	✓	0.703	0.694	1.367	0.707	0.698	1.354
Struc.	HBGSA	✗	0.554 ± 0.029	0.537 ± 0.017	1.628 ± 0.169	0.567 ± 0.021	0.541 ± 0.015	1.953 ± 0.653
	EGNN	✗	0.610 ± 0.006	0.590 ± 0.005	1.413 ± 0.005	0.621 ± 0.009	0.597 ± 0.008	1.384 ± 0.013
	$\text{EGNN}_{\text{TargetDiff}}$	✗	0.609 ± 0.005	0.589 ± 0.004	1.412 ± 0.011	0.616 ± 0.005	0.591 ± 0.005	1.389 ± 0.006
	GET	✗	0.611 ± 0.003	0.591 ± 0.004	1.410 ± 0.005	0.619 ± 0.004	0.595 ± 0.002	1.389 ± 0.004
	CheapNet	✗	0.644 ± 0.007	0.609 ± 0.009	1.409 ± 0.025	0.645 ± 0.012	0.608 ± 0.013	1.378 ± 0.029
	IPNet	✗	0.603 ± 0.010	0.567 ± 0.010	1.424 ± 0.014	0.610 ± 0.022	0.578 ± 0.020	1.406 ± 0.033
	IPNet†	✓	0.594 ± 0.015	0.557 ± 0.013	1.455 ± 0.015	0.607 ± 0.012	0.563 ± 0.011	1.421 ± 0.014
	AEV-PLIG	✗	0.595 ± 0.013	0.562 ± 0.015	1.479 ± 0.022	0.589 ± 0.015	0.556 ± 0.013	1.466 ± 0.026
	DSMBind	✓	0.448	0.445	N/A	0.449	0.445	N/A
	BindNet	✓	0.533 ± 0.024	0.504 ± 0.022	1.577 ± 0.104	0.545 ± 0.025	0.516 ± 0.022	1.626 ± 0.094
	ProFSA	✓	0.645 ± 0.005	0.612 ± 0.005	1.473 ± 0.021	0.651 ± 0.007	0.615 ± 0.007	1.445 ± 0.053
	CoE	✗	0.583 ± 0.024	0.567 ± 0.012	1.498 ± 0.059	0.591 ± 0.025	0.573 ± 0.013	1.478 ± 0.059
		✓	0.635 ± 0.001	0.606 ± 0.000	**1.418** ± 0.048	0.646 ± 0.001	0.612 ± 0.000	1.396 ± 0.031
Struc., Density	CoDE (Ours)	✗	0.587 ± 0.018	0.571 ± 0.011	1.476 ± 0.042	0.594 ± 0.019	0.576 ± 0.012	1.458 ± 0.044
		✓	**0.671** ± 0.002	**0.652** ± 0.001	**1.380** ± 0.054	**0.671** ± 0.002	**0.650** ± 0.001	**1.336** ± 0.022

Table 8: **CASF-2016 held-out binding-affinity prediction.** 90% confidence intervals (CI) using BCa are reported.

			CASF-2016 clean held-out (N = 92)					
Input	**Method**	**Pre.**	**Pearson *r***	**90% CI**	**Spearman ρ**	**90% CI**	**RMSE ↓**	**90% CI**
Seq.	DeepDTA	✗	0.520 ± 0.054	[0.39, 0.63]	0.488 ± 0.037	[0.35, 0.61]	1.735 ± 0.058	[1.57, 1.93]
	MolTrans	✗	0.529 ± 0.026	[0.41, 0.63]	0.500 ± 0.035	[0.37, 0.61]	1.732 ± 0.034	[1.57, 1.93]
	HonestAffinity	✓	0.477 ± 0.071	[0.36, 0.58]	0.432 ± 0.074	[0.30, 0.55]	1.777 ± 0.085	[1.63, 1.95]
	Nesso-1†	✓	0.621	[0.47, 0.73]	0.631	[0.49, 0.74]	1.721	[1.50, 2.04]
Struc.	HBGSA	✗	0.598 ± 0.021	[0.48, 0.68]	0.586 ± 0.023	[0.45, 0.69]	1.800 ± 0.189	[1.63, 2.01]
	EGNN	✗	0.634 ± 0.021	[0.51, 0.73]	0.629 ± 0.017	[0.50, 0.73]	1.591 ± 0.024	[1.42, 1.79]
	$\text{EGNN}_{\text{TargetDiff}}$	✗	0.643 ± 0.019	[0.52, 0.74]	0.636 ± 0.016	[0.51, 0.73]	1.588 ± 0.027	[1.42, 1.79]
	GET	✗	0.632 ± 0.017	[0.51, 0.73]	0.615 ± 0.011	[0.48, 0.72]	1.584 ± 0.026	[1.41, 1.80]
	CheapNet	✗	0.681 ± 0.035	[0.57, 0.77]	**0.649** ± 0.037	[0.52, 0.75]	1.539 ± 0.067	[1.36, 1.76]
	IPNet	✗	0.568 ± 0.024	[0.45, 0.66]	0.548 ± 0.016	[0.41, 0.66]	1.714 ± 0.029	[1.54, 1.94]
	IPNet	✓	0.672 ± 0.021	[0.57, 0.75]	0.623 ± 0.025	[0.50, 0.72]	1.540 ± 0.019	[1.38, 1.75]
	AEV-PLIG	✗	0.592 ± 0.021	[0.49, 0.69]	0.535 ± 0.028	[0.41, 0.64]	1.737 ± 0.054	[1.58, 1.93]
	DSMBind	✓	0.658	[0.55, 0.74]	0.653	[0.52, 0.75]	N/A	–
	BindNet	✓	0.573 ± 0.030	[0.43, 0.68]	0.557 ± 0.022	[0.40, 0.68]	1.696 ± 0.061	[1.54, 1.90]
	ProFSA	✓	**0.712** ± 0.011	[0.61, 0.78]	0.672 ± 0.009	[0.55, 0.76]	1.693 ± 0.018	[1.52, 1.91]
	CoE	✗	0.610 ± 0.048	[0.51, 0.69]	0.581 ± 0.044	[0.47, 0.68]	1.712 ± 0.130	[1.57, 1.89]
		✓	0.676 ± 0.008	[0.57, 0.76]	0.623 ± 0.008	[0.49, 0.73]	1.528 ± 0.016	[1.37, 1.69]
Struc., Density	CoDE (Ours)	✗	0.617 ± 0.031	[0.51, 0.71]	0.580 ± 0.026	[0.46, 0.68]	1.637 ± 0.092	[1.50, 1.82]
		✓	**0.720** ± 0.025	[0.62, 0.79]	**0.683** ± 0.024	[0.56, 0.77]	**1.435** ± 0.039	[1.29, 1.61]

E.2 VINA DOCKING

We present plots with all the baselines for the Vina docking regarding the target pocket and number of heavy atoms, extended from Figures 3 and 4b. CoDE still shows better performance in a majority of pockets. Also, all methods score worse when there are fewer heavy atoms, while CoDE scores better in many regions with many heavy atoms.

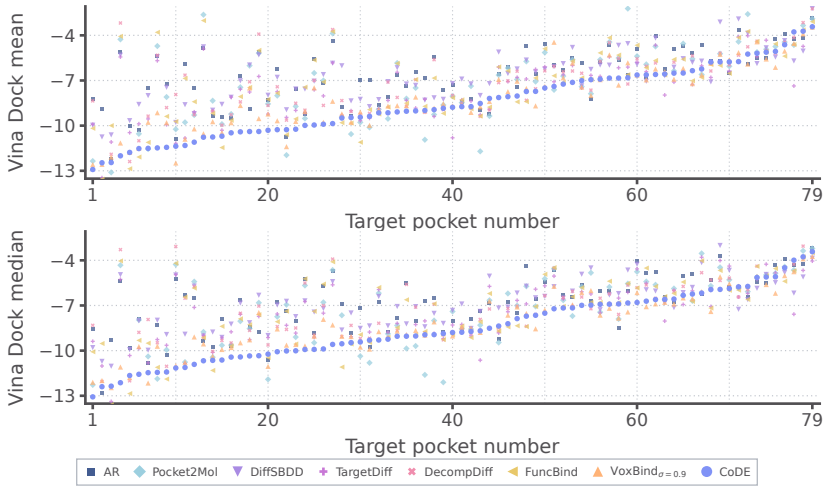


Figure 9: **Mean and median Vina docking scores per pocket.** of all generated molecules, with 79 test pockets having experimental density maps, sorted by the CoDE Vina docking score.

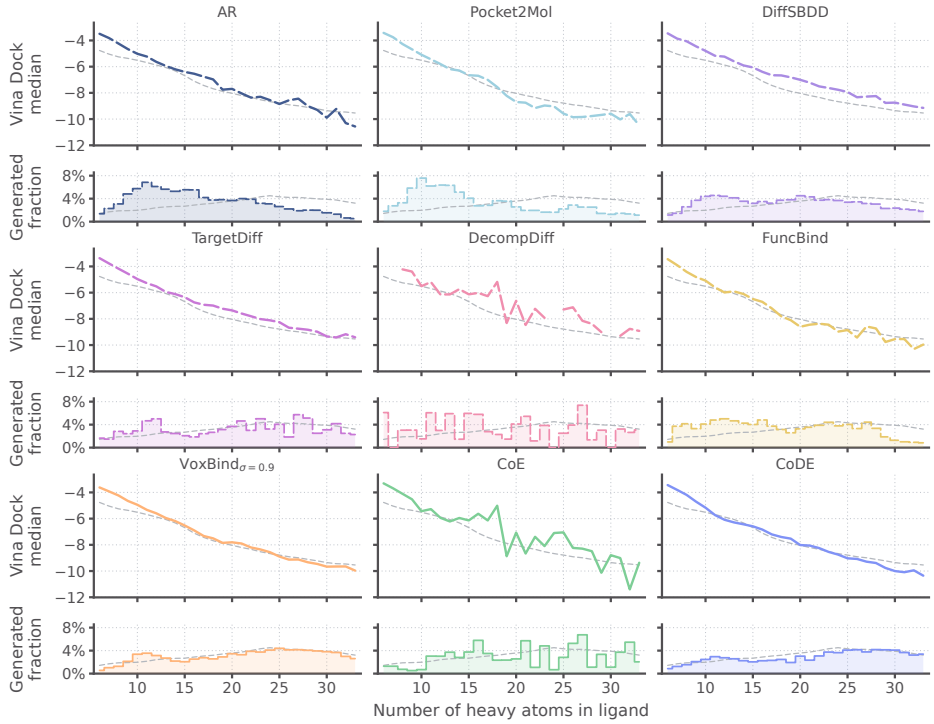


Figure 10: **Median Vina docking scores per number of heavy atoms.** of all generated molecules, with 79 test pockets having experimental density maps, sorted by the CoDE Vina docking score.

Also, we divide Figure 9 by CoDE for better visualization.



Figure 11: **Mean and median Vina docking scores per pocket compared against each baseline** with 79 test pockets having experimental density maps, sorted by the CoDE Vina docking score.

E.3 QUALITATIVE VISUALIZATION

Figure 12 extends Figure 4a with four additional pockets (5TJN, 1RS9, 4KCQ, and 1COY). For each pocket and method, we visualize the generated ligand with the lowest Vina docking score, annotated at the bottom right. CoDE yields the best scores, and VoxBind follows it; thus, the two voxel-based models occupy the binding site more fully than other Autoregressive and diffusion baselines. This aligns with the trend in Table 3 and Figure 3, where density conditioning steers ligands towards poses that fit the pocket. Note that these are all the best-scoring samples per method; they illustrate rather than quantify the gains.

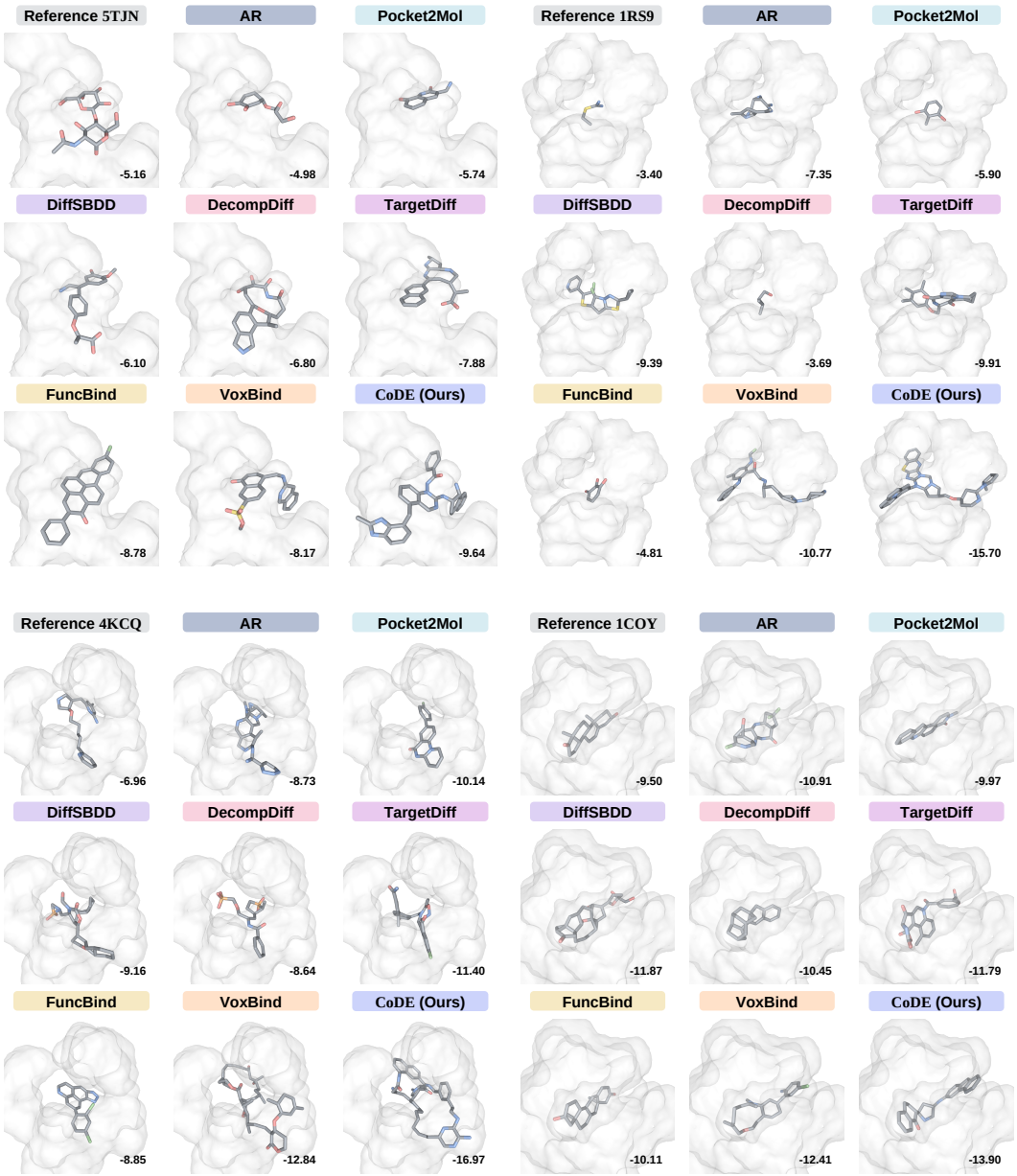
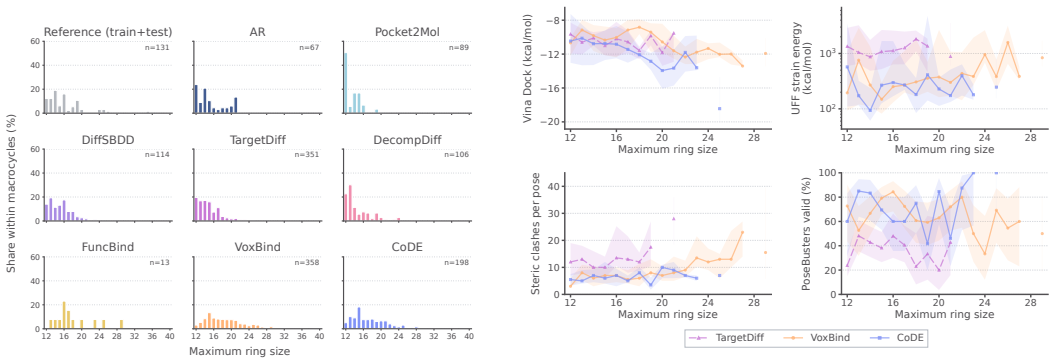


Figure 12: **Qualitative examples of the generated ligands.** The lowest Vina docking-score ligand generated is shown, with its docking score at the bottom right. Pockets 5TJN, 1RS9, 4KCQ, and 1COY are shown from left to right, from top to bottom.

E.4 MACROCYCLES IN LIGAND GENERATION

We ask whether the favorable docking scores of generated macrocycles come with favorable geometry. We define a macrocycle as a molecule with a ring of at least 12 atoms in RDKit's SymmSSSR, and characterize it by the size of that largest ring, which counts ring atoms rather than all heavy atoms in the molecule. Figure 13 compares TargetDiff, VoxBind, and CoDE on the same 79 test pockets. At every shared ring size, CoDE attains lower median strain, fewer clashes, and a higher PoseBusters pass rate than TargetDiff, and its median Vina Dock score is the lowest of the three for ring sizes of 18–21 atoms; against VoxBind the comparison varies with ring size.

These curves are descriptive rather than controlled. The three methods generate macrocycles in different pocket subsets (30, 40, and 60 pockets, with 24 shared by CoDE and VoxBind), and conditioning on ring size matches neither heavy-atom count nor pocket composition; a comparison matched on both leaves too few strata to support the pooled advantage over VoxBind. We also note that PoseCheck and PoseBusters assess the generated poses whereas Vina Dock scores follow re-docking, so passing PoseBusters does not validate the re-docked pose.



(a) Ring size distribution

(b) Quality by maximum ring size

Figure 13: **Generated macrocycles by maximum ring size.** **(a)** Distribution of the largest SymmSSSR ring size per method. **(b)** Median Vina Dock score, UFF strain energy (logarithmic axis), steric clash count, and PoseBusters pass rate at each exact ring size, without binning or smoothing. Shading shows the 25th–75th percentiles and error bars 95% Wilson intervals. Sizes with fewer than five evaluations for a method and metric are omitted, breaking the lines.

E.5 POSECHECK

We present additional results for PoseCheck on (i) strain energy against the number of rotatable bonds, and (ii) protein–ligand interactions. In Figure 14, FuncBind shows exceptionally high strain energy when the number of rotatable bonds increases. AR, DiffSBDD, and TargetDiff also show relatively high strain energy when the number of rotatable bonds increases, compared to Pocket2Mol, DecompDiff, and VoxBind. In Figure 15, we present the protein–ligand interaction regarding contacts and H-bonds. CODE does not degrade the performance compared to VoxBind in terms of both contacts and H-bonds.

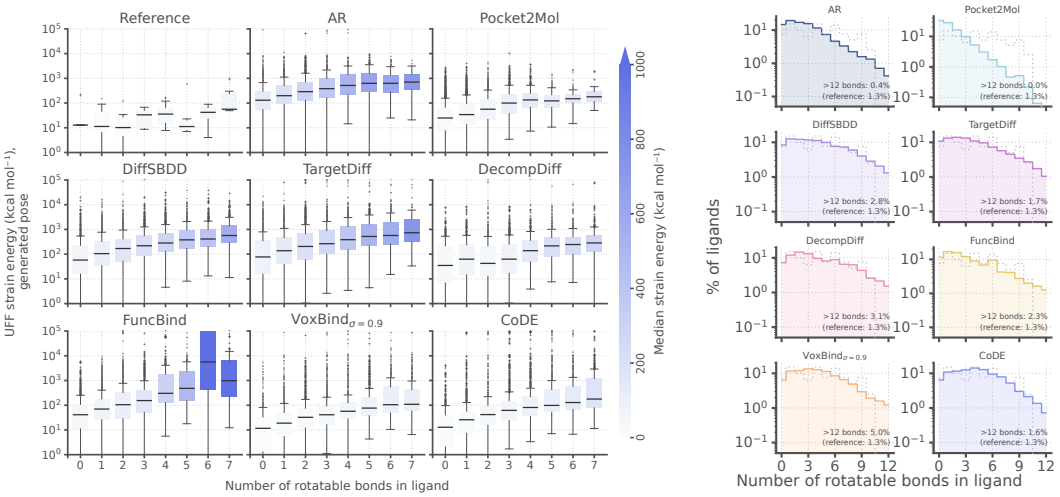


Figure 14: **Strain energy of rotatable bonds.** Strain energy is shown on a log scale, with the median value shown as a horizontal line and the box colored. The box represents the 25th-75th percentile, with outliers additionally plotted.

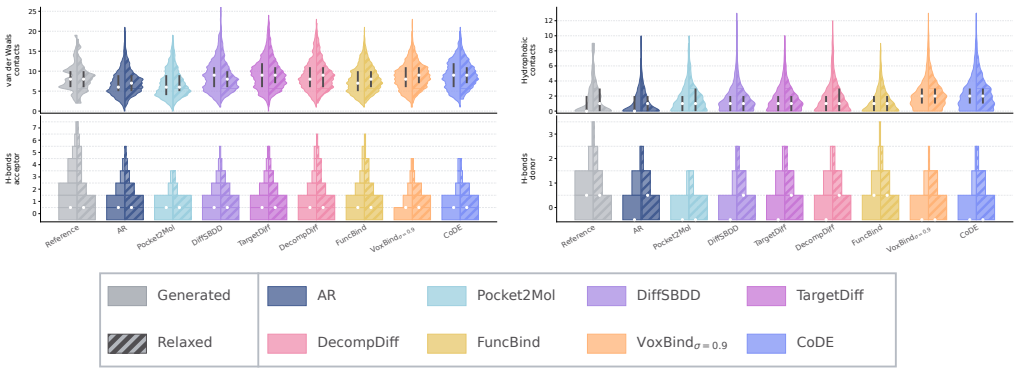


Figure 15: **Protein–ligand interaction.** Left side of each method represents metrics measured from the generated pose, while the right side of each method with slashes is measured from relaxed poses.

Additionally, we plot a qualitative example of re-docking in Figure 16. For each method, we show the ligand whose pose changes least under re-docking, i.e., the ligand with the lowest re-docking RMSD among the ligands re-docked per pocket, where the RMSD is computed over heavy atoms in place, without superposition. As this pocket is one where CoDE attains the lowest re-docking RMSD, the example illustrates rather than quantifies; the comparison over all pockets is given in Figure 5d.

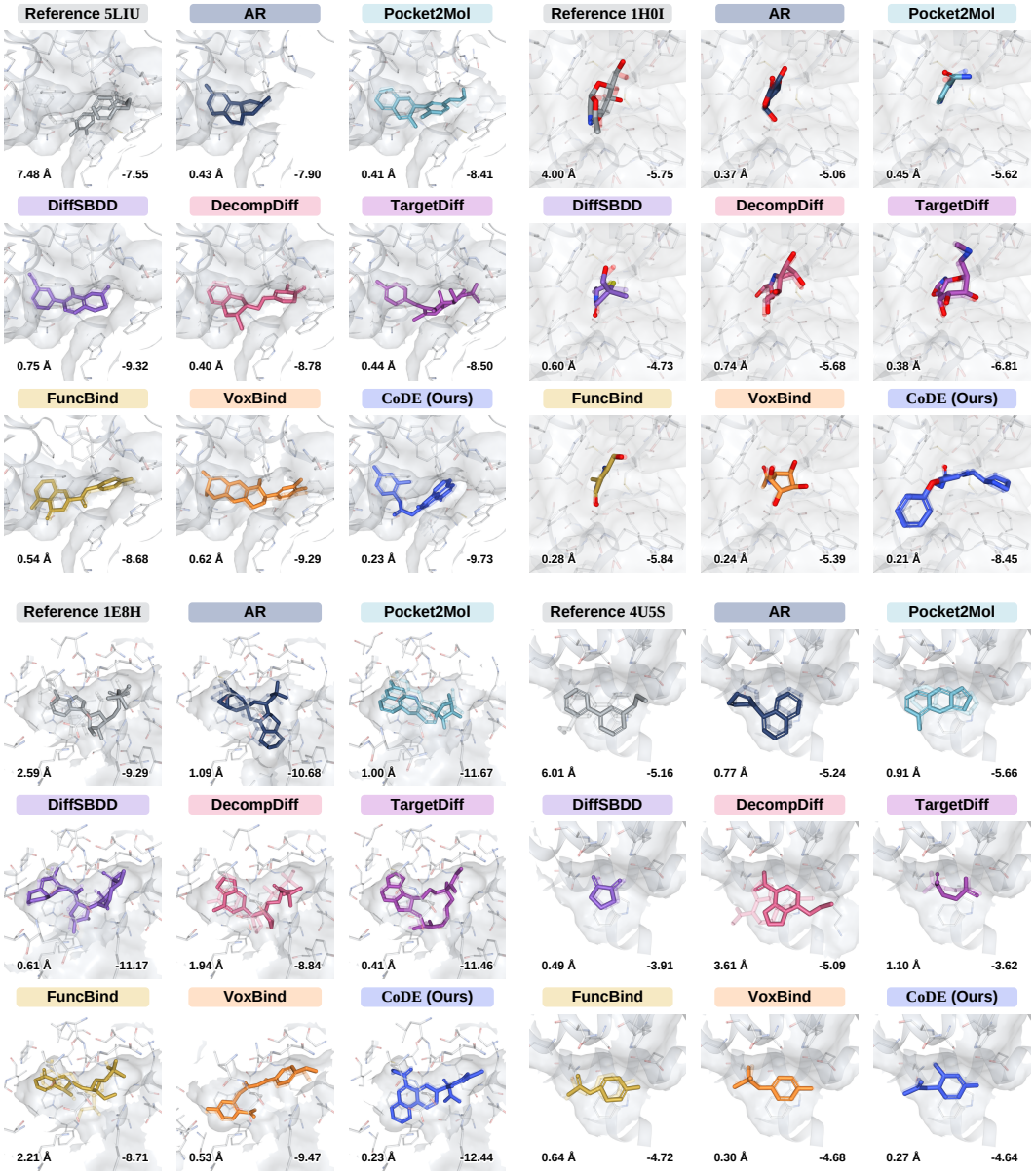


Figure 16: **Qualitative examples of the re-docked ligands.** Pockets `5LIU`, `1H0I`, `1E8H`, and `4U5S` are shown from left to right, from top to bottom. Carbon atoms are colored with the method color. Re-docking RMSD and docked Vina score are each visualized in the bottom left and bottom right.

E.6 MOLECULAR STRUCTURE

We report the detailed Jensen-Shannon distance (Lin, 1991, JSD) for molecular structures, such as atom types, bond distances, atom pair distances, and ring statistics.

- Heavy atom types: Figure 17
- Atom pair distance under 12Å: Figure 18
- Bond distance under 2Å: Figure 19
- Overall ring statistics: Figure 20
- Ring size: Figure 21
- Number of rings: Figure 22
- Aromatic rings: Figure 23

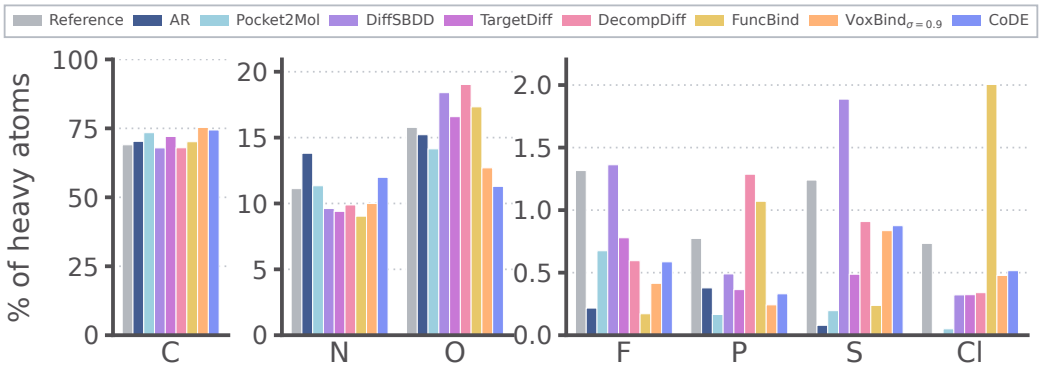


Figure 17: **Heavy-atom element shares.** Percentage of a set's heavy atoms at each element, split across three linearly scaled panels so that rare elements are not flattened against carbon: C, then N and O, then F, P, S, and Cl.

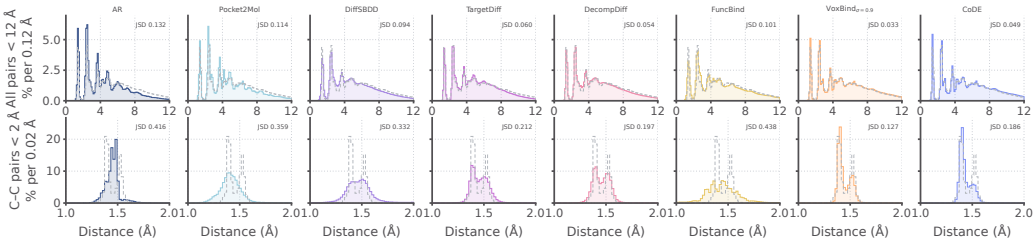


Figure 18: **Interatomic distance distributions.** All heavy-atom pairs within a molecule closer than 12Å is shown above, and carbon–carbon pairs closer than 2Å below. Distances are computed from coordinates and elements alone, without bond perception. The dashed gray outline is the reference ligand, the number in the corner shows the JSD, and the vertical scale is shared.

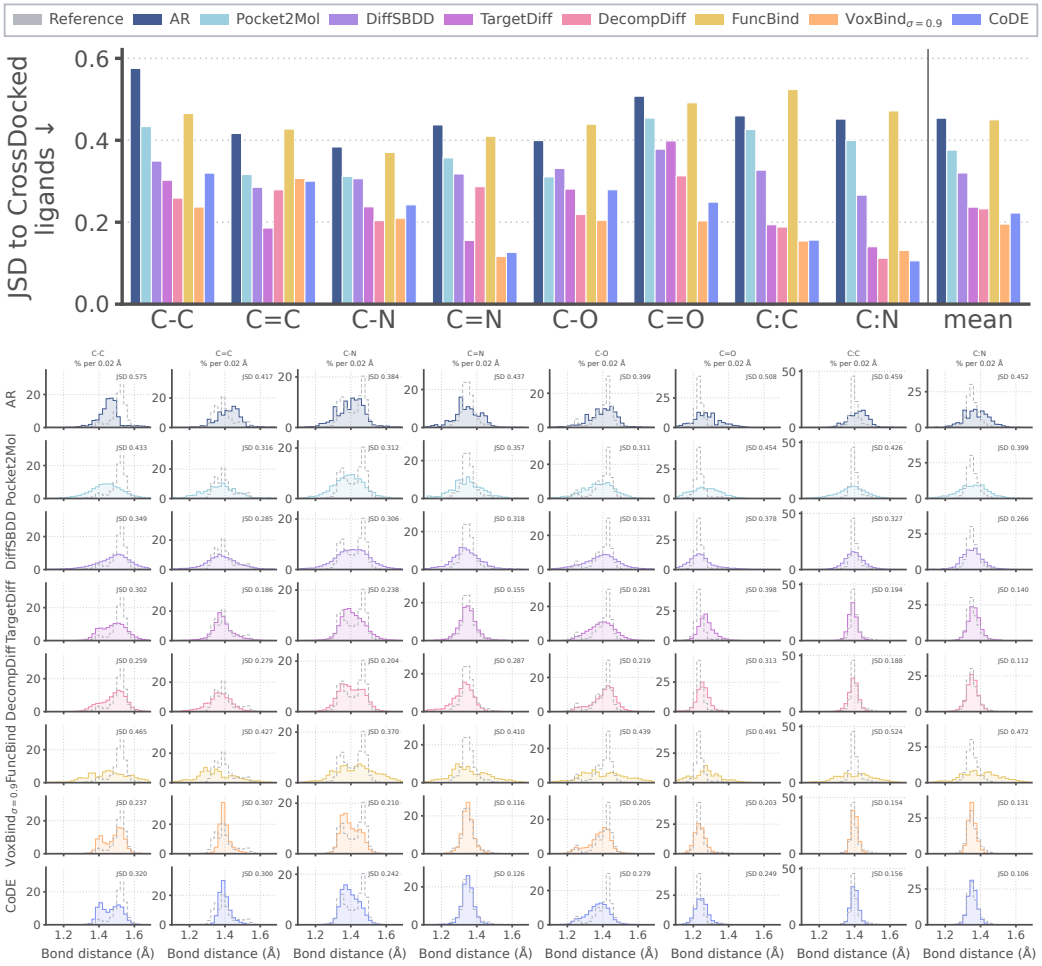


Figure 19: **Bond-length Jensen–Shannon distance per bond type (top) and detailed distribution (bottom)**, for the eight bond types, the lower the value, the closer to the reference distribution. The dashed gray outline is the reference distribution, and the number in the corner is that panel's JSD. Distances are scored in 0.005Å bins and drawn four bins at a time, and the vertical scale is shared.

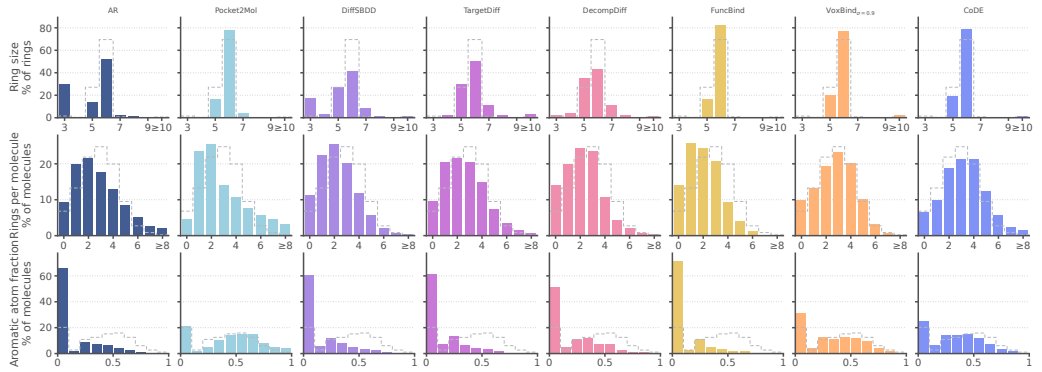


Figure 20: **Ring statistics, one column per method.** Ring sizes, rings per molecule, and the aromatic fraction of a molecule's atoms and of its rings, with the reference ligands repeated in gray.

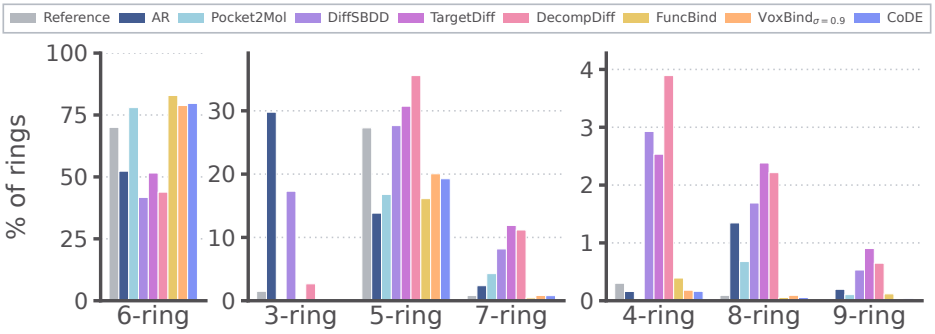


Figure 21: **Ring-size distribution.** The three panels are scaled independently: six-membered rings, then three-, five-, and seven-membered rings, then four-, eight-, and nine-membered rings.

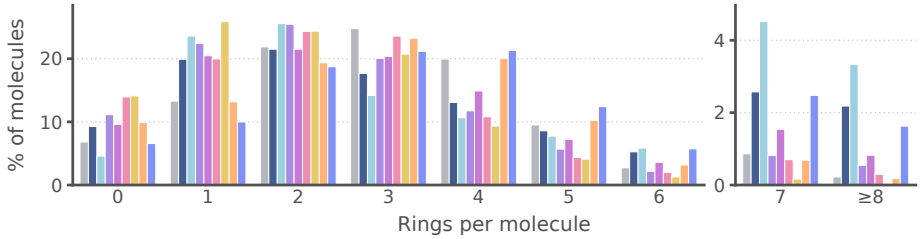


Figure 22: **Rings per molecule.** Percentage of a set's molecules containing each number of rings, with the final bar pooling molecules with eight or more rings.

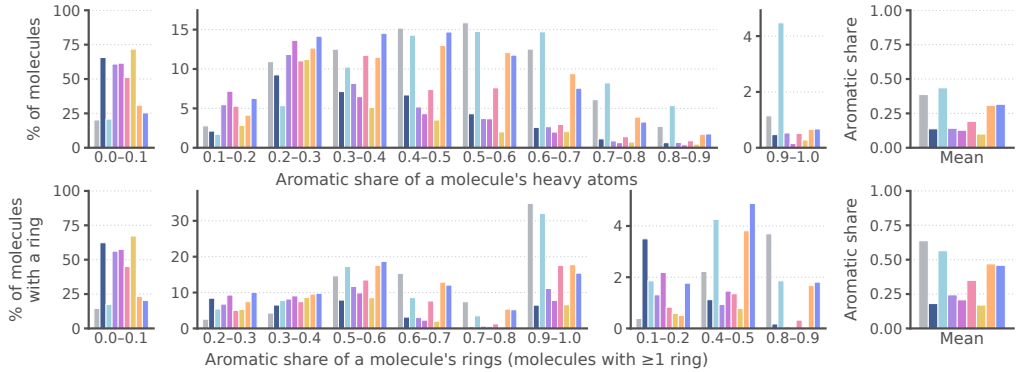


Figure 23: **Aromatic content.** Top: the fraction of a molecule's heavy atoms that are aromatic. Bottom: the fraction of a molecule's aromatic rings, meaning every bond of the ring is aromatic, computed over molecules with at least one ring. Both panels show the percentage of a set's molecules in each fraction bin, with the reference ligands drawn alongside the methods.

E.7 RIGID FRAGMENT CONSISTENCY

We present extended results from Figure 6, regarding the number of rigid fragments and the distribution of generated molecules for each method.

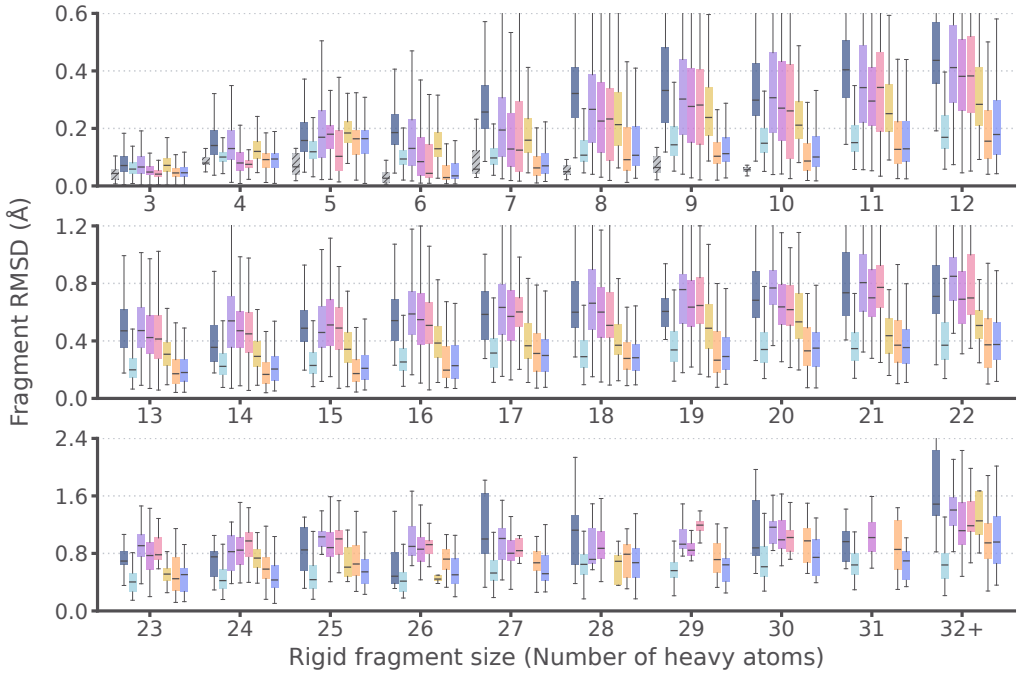


Figure 24: **Rigid fragment consistency expanded to more number of fragment size.**

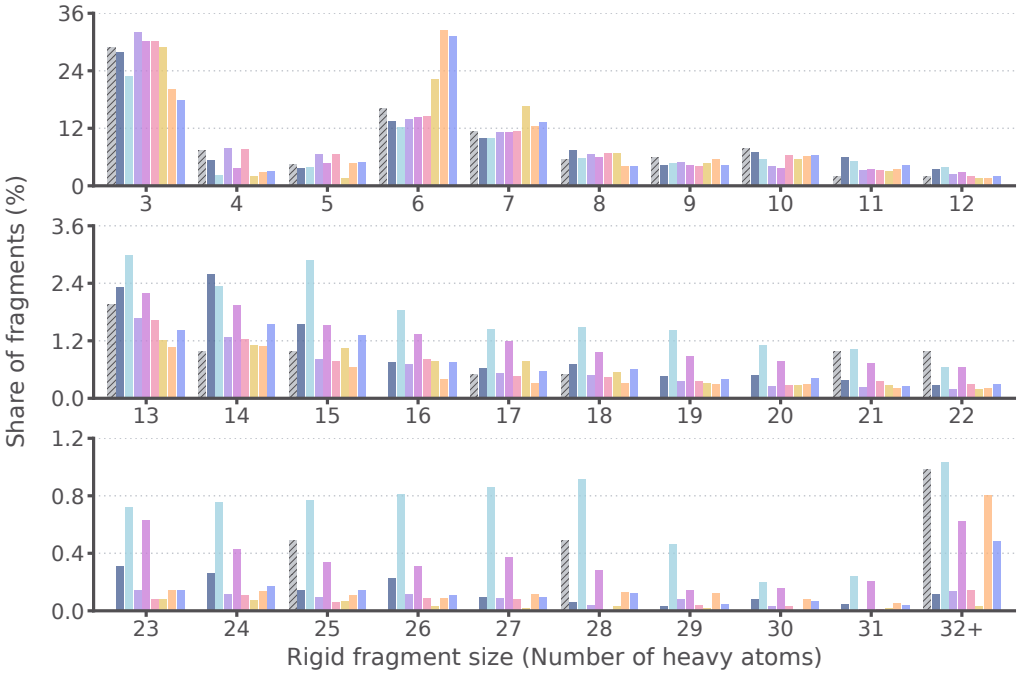


Figure 25: **Rigid fragment size distribution from generated molecules.**

E.8 POSEBUSTERS

We present PoseBusters failure cases across 22 tasks in `dock` mode, with an upset plot showing correlations between categories.

Category	Specific check	Reference	AR	Pocket2Mol	DiffSBDD	TargetDiff	DecompDiff	FuncBind	VoxBind$_{\sigma=0.9}$	CoDE
Overall valid		96.2	69.8	73.6	56.2	60.7	81.0	50.0	69.2	67.4
Bond geometry	Bond lengths	100.0	98.0	89.6	81.0	98.8	97.8	76.9	97.0	96.4
	Bond angles	100.0	86.5	97.9	78.7	77.4	96.1	85.4	94.4	94.0
Planarity	Aromatic ring flatness	100.0	91.5	99.5	99.8	99.9	100.0	100.0	99.9	100.0
	Double bond flatness	100.0	96.8	98.5	99.4	99.9	98.1	100.0	100.0	100.0
Ring pucker	Non-aromatic ring non-flatness	98.7	99.9	99.0	98.8	91.0	98.1	77.5	78.8	77.5
Valence & connectivity	Sanitization	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	All atoms connected	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	Inchi convertible	100.0	100.0	100.0	100.0	100.0	99.9	100.0	100.0	100.0
	No radicals	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Protein clash & overlap	Minimum distance to protein	97.4	96.1	83.2	79.0	89.9	90.1	95.9	98.7	98.3
	Volume overlap with protein	100.0	99.6	99.4	94.3	99.6	99.6	99.6	99.8	99.9
	Protein-ligand maximum distance	100.0	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0
	Minimum distance to organic cofactors	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	Minimum distance to inorganic cofactors	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	Minimum distance to waters	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	Volume overlap with organic cofactors	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	Volume overlap with inorganic cofactors	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	Volume overlap with waters	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Self-clash & energy	Internal steric clash	100.0	89.9	99.2	93.5	93.8	96.7	93.0	95.9	94.7
	Internal energy	100.0	99.1	99.8	98.5	98.2	99.3	94.8	97.4	97.4
Not scored	Mol pred loaded	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.9	99.8
	Mol cond loaded	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

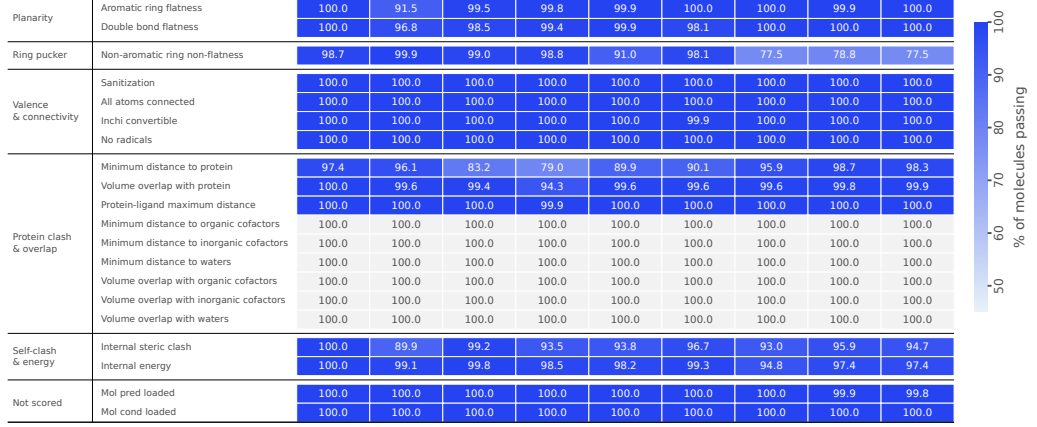


Figure 26: **PoseBusters validity per category.**

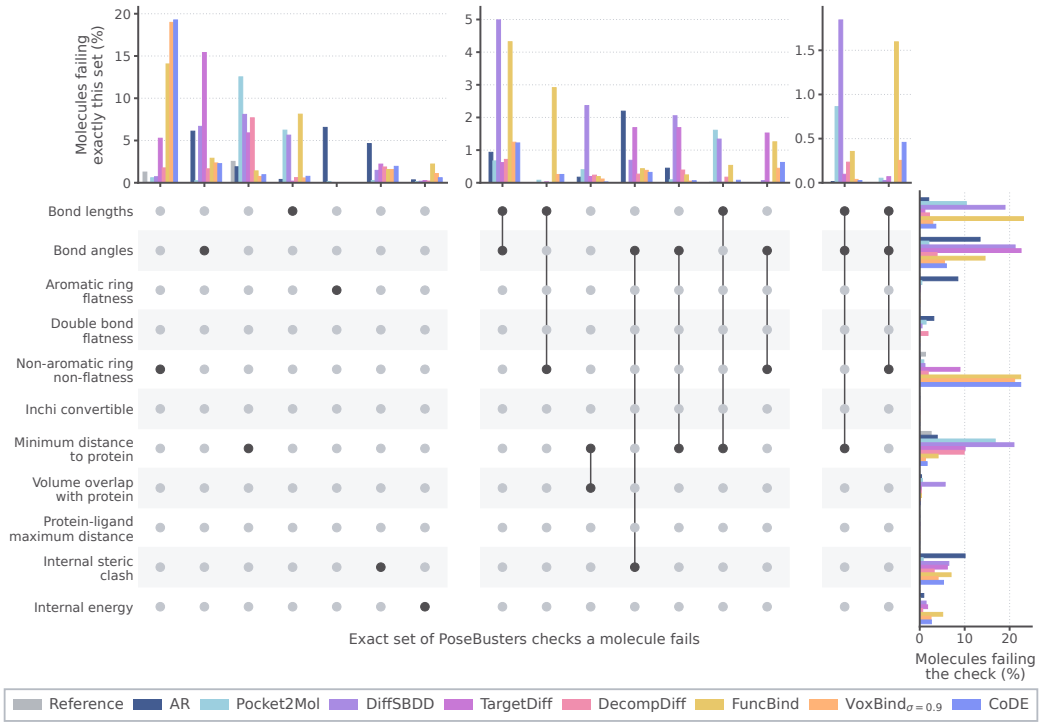


Figure 27: **PoseBusters validity upset plot.** Relation between PoseBusters categories.

Ring pucker failure analysis. The *Pucker* column of Figure 7 is the single PoseBusters test `non-aromatic_ring_non-flatness`. It fails a molecule when a non-aromatic six-membered ring has no atom at least 0.05 Å from the ring's best-fit plane; five-membered rings are not tested. VoxBind, CoDE, TargetDiff, and FuncBind generate atoms without bonds, and all four assign bonds

afterwards with the same OpenBabel-based construction (O'Boyle et al., 2011), which does not use aromaticity. To assess the contribution of molecular reconstruction, we classify failing molecules into four diagnostic categories. These categories describe geometric and reconstruction properties; they do not establish a unique underlying cause for each failure.

- **Flat non-aromatic ring.** A genuinely planar saturated or partially saturated ring, which is the failure the test is designed to catch.
- **Signed-distance rule.** A ring atom lies at least 0.05 Å from the plane, so the ring is puckered; PoseBusters nevertheless fails it because it compares the largest *signed* distance with the threshold, and the orientation of the plane normal is arbitrary.
- **Benzene-like, no valid aromatic assignment.** The failing ring has benzene-like geometry (ring bonds between 1.20 and 1.60 Å, mean internal angle between $116°$ and $124°$, every atom within 0.1 Å of the plane), but re-assigning it as aromatic cannot be kekulized with the rest of the molecule.
- **Valid aromatic reassignment.** The failing ring meets the benzene-like geometric criteria above, and an alternative aromatic bond assignment yields a valid, kekulizable molecule after adjusting attached hydrogens. Such cases are consistent with a mismatch between the generated geometry and the reconstructed molecular graph. However, the availability of an aromatic alternative does not establish that the original ring was intended to be aromatic.

For CoDE, 87.3% of the failing molecules fall into the valid-aromatic-reassignment category, while 11.6% have benzene-like geometry without a valid aromatic assignment under our diagnostic procedure. The flat non-aromatic category accounts for 0.9% of its failing molecules, or 0.20% of all generated molecules, compared with 0.23% of all generated molecules for VoxBind. These results suggest that molecular reconstruction contributes substantially to the observed ring non-flatness failures. The categories are diagnostic, however, and do not establish the physical validity of the original molecular graphs.

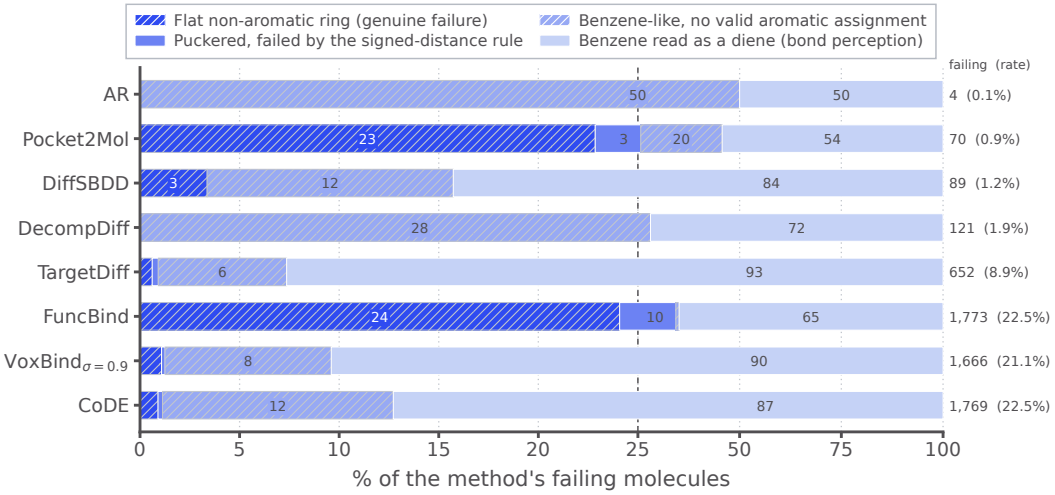


Figure 28: **Diagnostic categories of PoseBusters ring non-flatness failures.** Each bar splits the molecules of one method that fail `non-aromatic_ring_non-flatness` into the four categories, as a percentage of its failing molecules; segments narrower than about 5% of the axis are left unlabeled. The horizontal axis is stretched below 25% and compressed above it (dashed line), each part linear, so that the small causes remain legible. On the right are the number of failing molecules and the failure rate over all generated molecules; the AR bar rests on only four molecules.

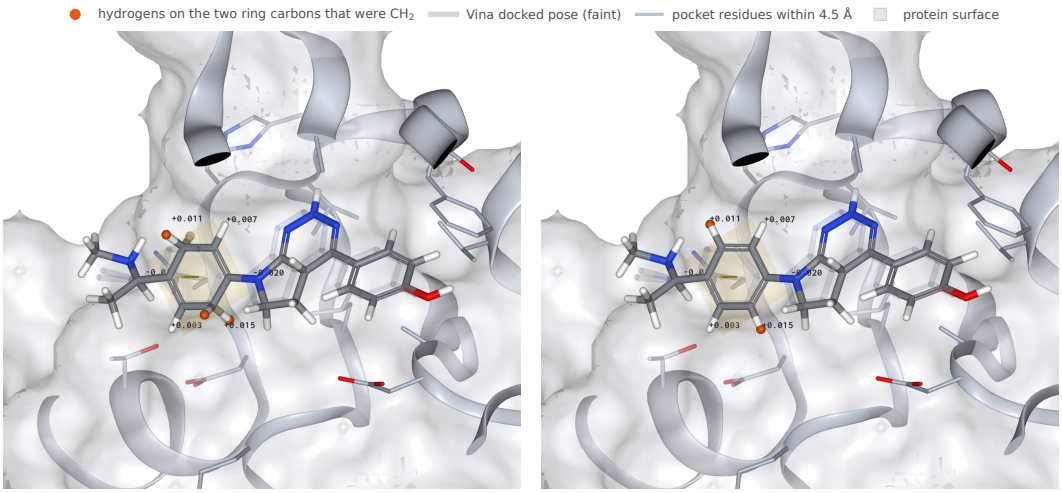
Figure 29 shows that alternative bond-order and hydrogen assignments can change the ring non-flatness outcome for the same heavy-atom geometry. It reduces the ring non-flatness failure rate of CoDE (Table 9), with modest changes in the aggregated molecular properties. Also, in the evaluated docking subset, the Vina score changes are small (Table 10). We interpret these results as sensitivity to molecular reconstruction, rather than as validation of the original molecular graphs.

Table 9: **Evaluation metrics before and after alternative aromatic reassignment.** A failing six-membered ring with benzene-like geometry is reassigned as aromatic by modifying the ring bond orders and the attached hydrogen counts, while preserving the heavy-atom connectivity and coordinates. Rings bearing an exocyclic double bond are left unchanged in the procedure. *Ring fail* is the share of molecules failing `non-aromatic_ring_non-flatness`, re-derived ring by ring (rings whose outcome depends on the sign of the plane normal count as failing). *All rings* is the rate if rings with an exocyclic double bond are re-assigned as well. QED, SA, and logP are means over all molecules; diversity is averaged over pockets.

	Ring fail (%)			**Changed**				
Method	Before	After	All rings	(%)	**QED**	**SA**	**logP**	**Diversity**
AR	0.1	0.1	0.0	0.0	0.504 → 0.504	0.629 → 0.629	0.37 → 0.37	0.709 → 0.709
Pocket2Mol	1.1	1.0	0.5	0.1	0.566 → 0.566	0.743 → 0.743	1.64 → 1.64	0.724 → 0.724
DiffSBDD	1.4	1.4	0.2	0.0	0.495 → 0.495	0.628 → 0.628	1.07 → 1.07	0.796 → 0.796
DecompDiff	1.9	1.8	0.5	0.2	0.494 → 0.494	0.666 → 0.666	1.12 → 1.12	0.711 → 0.711
TargetDiff	9.5	3.7	0.7	5.9	0.493 → 0.497	0.608 → 0.615	1.36 → 1.36	0.747 → 0.748
FuncBind	27.0	12.3	9.2	16.6	0.503 → 0.511	0.656 → 0.678	1.35 → 1.37	0.702 → 0.712
VoxBind$_{\sigma=0.9}$	21.3	8.7	2.1	14.0	0.540 → 0.544	0.704 → 0.720	2.65 → 2.66	0.763 → 0.764
VoxBind$_{\sigma=0.9}$ + CoDE	23.0	10.6	2.9	13.9	0.523 → 0.526	0.678 → 0.692	2.62 → 2.63	0.710 → 0.711

Table 10: **Vina before and after the re-assignment**: We randomly selected 50 molecules for each of two reassignment procedures: one that preserves exocyclic bond orders and one that also modifies them. Each molecule was docked before and after reassignment.Δ is after minus before in kcal/mol, mean (median); positive is worse. *Same pose*: the two docked poses lie within 0.5 Å RMSD. *Gen.→dock*: median RMSD between the generated and the docked pose, before → after.

Fix	n	**ΔScore**	**ΔMin**	**ΔDock**	**Same pose**	**Gen.→dock** (Å)
Ring bond-order / H assignment (used)	49	+0.008 (+0.000)	−0.068 (+0.000)	−0.007 (+0.000)	40/49	3.37 → 3.41
Including exocyclic bond-order changes	49	+0.187 (+0.241)	+0.242 (+0.252)	+0.188 (+0.231)	10/49	2.27 → 2.42



(a) **Original reconstruction: 1,4-cyclohexadiene.** Two ring carbons are CH_2; the ring's largest distance to its plane is 0.020 Å, so PoseBusters fails it as non-aromatic and flat. Vina score −8.54, dock −8.73 kcal/mol.

(b) **Alternative aromatic reconstruction: benzene.** Same heavy atoms at the same coordinates; the ring is aromatic and not tested. Vina score −8.54, dock −8.73 kcal/mol.

Figure 29: **Alternative molecular reconstruction of the same generated heavy atom geometry.** A molecule generated by CoDE for `CDK6`. The yellow square is the best-fit plane of the failing ring, with each ring atom's signed distance to it in Å; orange marks the hydrogens on the two ring carbons that were CH_2. The faint sticks are the Vina-docked pose, 0.70 Å RMSD from the generated pose and identical for both assignments. The two reconstructions differ in ring bond orders and hydrogen counts, while their heavy-atom coordinates are identical. For this example, the Vina-docked poses and the reported Vina scores are unchanged.