ATOMISTIC REPRESENTATION ALIGNMENT FOR 3D MOLECULE GENERATION IN PROTEIN POCKETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Structure-based generative models have emerged as a powerful tool for creating novel 3D molecules within protein binding sites. However, these models often struggle to capture the fine-grained physical constraints of molecular systems. While typically trained on datasets of crystal structures and docked complexes, this spatial data provides only geometric supervision, lacking explicit information about the underlying energy landscape and atomic forces. In this work, we propose enhancing pocket-conditioned generative models by leveraging the rich internal representations of a machine-learned interatomic potential (MLIP). Because MLIPs are pre-trained on diverse quantum mechanical calculations, their latent spaces provide a highly structured, atom-aware target for our generator to learn from. By applying an equivariant projector to match the features of a denoiser layer to the MLIP, we expose the generator to a dense, physically grounded signal during training. We observe a significant improvement in the overall structural validity of the generated molecules. Evaluated on a challenging test set of out-of-distribution targets, we demonstrate that representation alignment transfers across architectures: when applied to two independent pocket-conditioned generators, it yields molecules with greater substructure overlap with native ligands and improves protein–ligand interaction fingerprint recovery in both, indicating more faithful reproduction of native binding modes. Furthermore, after a pocket-constrained relaxation in an independent computational force field, the fraction of generated molecules that are simultaneously valid, clash-free, low-strain, and well-docked increases. Together, these results indicate that aligning to an MLIP yields models whose outputs better recover native interaction patterns and converge to more physically favorable geometries.

1 INTRODUCTION

Generative models provide a flexible framework for exploring chemical space by learning to approximate the joint distributions of atoms in three-dimensional space. These systems generally follow three distinct paths: unconditional models (Hoogeboom et al., 2022; Xu et al., 2023a), which generate chemically valid molecules in isolation; property-guided models (Zeng et al., 2026; Xu et al., 2023b), which steer generation toward specific characteristics; and target-conditioned models (Guan et al., 2023; Luo et al., 2021; Peng et al., 2022; Schneuing et al., 2022; 2025), which explicitly incorporate the geometry and chemistry of a protein binding pocket. This target-conditioned approach offers profound advantages for drug discovery: rather than virtually screening millions of pre-existing compounds, conditional models directly sample custom-tailored candidate ligands within the target's microenvironment.

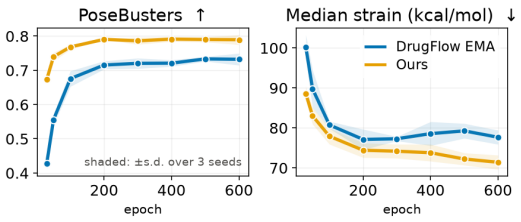

Figure 1: Atomistic representation alignment improves the 3D generated molecules PoseBusters score and strain.

However, capturing the fine-grained chemical realism of molecular systems remains a persistent challenge for these models. This difficulty stems primarily from their reliance on purely geometric supervision. Typically trained to reconstruct atomic coordinates from datasets of crystal structures

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and docked complexes (Francoeur et al., 2020), these models lack explicit information about the underlying energy landscapes or interatomic relationships that govern molecular stability. As a result, much like predictive docking and scoring methods trained on the same spatial data, pocket-conditioned generators often produce molecules that exhibit severe internal strain or steric clashes (Buttenschoen et al., 2024). Furthermore, this coordinate-only training signal restricts robust generalization to completely novel, unseen protein targets (Durairaj et al., 2024; Kopko et al., 2025).

We hypothesize that these structural and generalization bottlenecks arise because purely geometric data leaves crucial energetic context implicit. To explore whether a richer, supplementary training signal can alleviate these issues, we supervise the internal representations of a pocket-conditioned generator while leaving its standard coordinate-denoising objective untouched. We utilize a frozen teacher: a machine-learned interatomic potential (MLIP) pretrained on large-scale quantum-mechanical data (Wood et al., 2025). Because MLIPs are exposed to highly diverse stable and off-equilibrium configurations during pretraining, their internal features encode a rich, chemically aware context regarding favorable atomic arrangements. We transfer this representational richness via representation alignment (Yu et al., 2025): during training, an intermediate layer of the generator is projected into the teacher's feature space and matched to the MLIP's per-atom embeddings. By encouraging the generator to build representations from which this atomistic context can be derived, we provide a dense supplementary training signal. Because the MLIP remains frozen and the projectors are discarded after training, this alignment adds zero computational cost to the generation process itself.

Our contributions are summarized as follows:

1. **Atomistic Representation Alignment:** We propose an equivariant representation alignment scheme that enhances pocket-conditioned generative models by leveraging the rich internal representations of a frozen, pretrained machine-learned interatomic potential (MLIP). By matching the latent space of the generator to that of the MLIP, we provide a dense, atom-aware training signal that improves generation without adding inference overhead.
2. **Improved Physical Plausibility and Native Similarity:** We demonstrate that this alignment acts as a highly transferable inductive bias across independent architectures and datasets. Aligned models produce more physically plausible molecules with higher baseline validity and lower internal strain. Furthermore, on challenging out-of-distribution targets, they more faithfully recover native protein–ligand interaction fingerprints and generate molecules with greater substructure similarity to native binders.
3. **Higher Yield of Viable Candidates Post-Minimization:** By applying a pocket-constrained relaxation, we demonstrate that many geometric imperfections in raw generated poses are easily fixable artifacts. Crucially, the core advantages of representation alignment survive this refinement step, ultimately yielding a significantly higher absolute fraction of candidate molecules that are simultaneously valid, clash-free, low-strain, and well-docked.

Ultimately, our results establish representation alignment with pre-trained MLIPs as a pragmatic mechanism for transferring rich atomistic context into generative models, directly improving both the structural realism and the overall yield of viable ligand candidates.

2 Related Work

Generative Models for Molecules. Deep generative models sample novel molecular structures in unconditional (Hoogeboom et al., 2022; Xu et al., 2023a), property-guided (Zeng et al., 2026; Xu et al., 2023b), and target-conditioned settings. The latter (Guan et al., 2023; Luo et al., 2021; Peng et al., 2022; Schneuing et al., 2022; 2025) condition on the 3D geometry and chemistry of a binding pocket to generate high-affinity ligands directly *in silico*, an efficient alternative to high-throughput virtual screening. These models are predominantly E(3)-equivariant GNNs, differing mainly in training framework: early methods build molecules autoregressively (Luo et al., 2021; Peng et al., 2022), while recent diffusion (Guan et al., 2023; Schneuing et al., 2022) and flow-matching (Schneuing et al., 2025) approaches denoise the full structure iteratively. A related line injects auxiliary pretrained networks into this process: IPDiff (Huang et al., 2024b) shifts the diffusion trajectories using an affinity-supervised interaction prior, and IRDiff (Huang et al., 2024a) retrieves high-affinity reference ligands and augments denoising by cross-attention to them. Both draw their signal from curated PDB

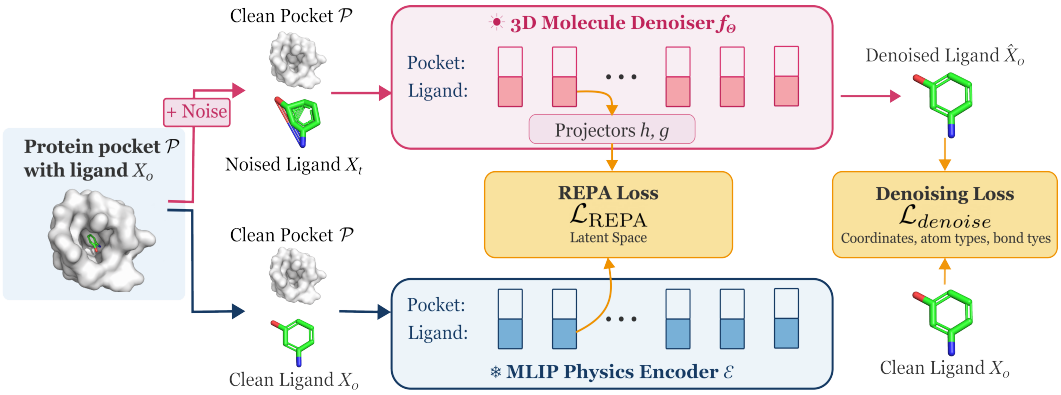

Figure 2: **Architectural overview of a 3D molecule denoiser trained with physics-guided representation alignment.** The generative process trains the 3D molecule denoiser f_θ, conditioned on the target pocket $\mathcal{P}$, to recover the clean ligand X_0 from its noised state X_t. The denoiser is supervised by the denoising loss ($\mathcal{L}_{\text{denoise}}$), which operates in atom-coordinate space and teaches the model to follow the denoising trajectory toward the clean ligand. In addition, atom-level features from an intermediate layer of the denoiser are extracted, transformed through equivariant projectors (h, g), and aligned with target representations from an intermediate layer of a frozen physics encoder $\mathcal{E}$ using the representation alignment (REPA) loss ($\mathcal{L}_{\text{REPA}}$). This alignment is done only for features corresponding to the ligand, and it regularizes the denoising trajectory with implicit physical constraints.

complexes and act on the sampling trajectory, whereas we align intermediate representations to a quantum-mechanically trained potential and leave sampling unchanged.

Generalization and Benchmarking. Models trained on PDB-derived datasets (Wang et al., 2005; Francoeur et al., 2020)—scoring, docking, cofolding, and generative alike—generalize poorly to novel targets and often violate basic physics. Recent benchmarks quantify this: PoseBusters (Buttenschoen et al., 2024) exposes chemically implausible poses, PLINDER (Durairaj et al., 2024) tests structural generalization, and Runs-N-Poses (Škrinjar et al., 2025) probes OOD behaviour of co-folding models. SPINDR (Cremer et al., 2026) extends PLINDER to generative models, using strict pocket-level clustering to minimize leakage and provide a test set of entirely novel targets.

Machine Learning Interatomic Potentials. MLIPs predict potential energies and atomic forces, approximating DFT at near *ab initio* accuracy for a fraction of the cost and enabling simulation far beyond its reach (Jacobs et al., 2025; Wang et al., 2025b). Modern MLIPs are typically equivariant GNNs (Batatia et al., 2022; Batzner et al., 2022; Gasteiger et al., 2021) trained on large datasets pairing diverse stable and strained geometries with energies and forces, from which they internalize the rules of atomic interaction. Highly diverse datasets (Levine et al., 2026; Barros-Luque et al., 2026; Gharakhanyan et al., 2026; Eastman et al., 2023) have driven a shift to universal potentials such as MACE-MP-0 (Batatia et al., 2025) and UMA (Wood et al., 2025), trained jointly across inorganic crystals, catalyst surfaces, drug-like molecules, and peptides. These now support zero-shot quantum-accurate simulation of solvated enzymes without QM/MM boundaries (Gao et al., 2026), and automated pipelines generate bespoke target-specific potentials (Lahouari et al., 2026).

Representation Alignment (REPA). REPA (Yu et al., 2025) showed that aligning intermediate layers of a diffusion model to a frozen vision foundation model (Oquab et al., 2024) accelerates training and improves sample quality. In molecular generation, REED (Wang et al., 2025a) aligns to pretrained molecular embeddings (Zhou et al., 2023), whereas MACE-REPA (Pinede et al., 2025) aligns to a pretrained MACE potential (Batatia et al., 2022). MLIPs are appealing targets (Didi, 2025) because they encode physically grounded geometry and energetics learned from abundant off-equilibrium data; Li & Walsh (2026) show that independently trained foundation MLIPs converge to a shared, architecture-independent latent geometry driven by physical supervision, indicating these representations capture transferable physics rather than model-specific artifacts. Critically, both prior molecular REPA works target different tasks: REED generates molecules unconditionally, while MACE-REPA performs Boltzmann emulation, drawing i.i.d. equilibrium conformers of a *fixed* system as a faster replacement for molecular dynamics. We instead train a pocket-conditioned generative model with alignment to a frozen MLIP, where the model must generalize to novel protein targets rather than resample a known one.

3 POCKET-CONDITIONED MOLECULE GENERATION WITH REPRESENTATION ALIGNMENT

We address *pocket-conditioned 3D molecule generation*: given a protein pocket $\mathcal{P}$, generate small-molecule ligands that bind to it. A clean (ground-truth) ligand is represented as a set of N_ℓ heavy atoms with 3D positions $\mathbf{X}_0 \in \mathbb{R}^{N_\ell \times 3}$, discrete atom types $\mathbf{a}_0 \in \mathcal{A}^{N_\ell}$, and bond types $\mathbf{b}_0 \in \mathcal{B}^{N_\ell \times N_\ell}$ over the five categories $\mathcal{B} = \{\varnothing, \text{single}, \text{double}, \text{triple}, \text{aromatic}\}$, one for each unordered pair of ligand atoms. The goal of the generative model is to learn the conditional distribution $p(\mathbf{X}_0, \mathbf{a}_0, \mathbf{b}_0 \mid \mathcal{P})$, from which we can sample novel ligands that are simultaneously chemically valid, geometrically realistic, and sterically complementary to the target pocket.

Limitations of geometry-only supervision. Despite their empirical success in pocket-conditioned molecule generation, denoising-based generative models are fundamentally limited by their reliance on purely geometric supervision. Because the training objective reverses a coordinate noising process, it treats all training co-crystal structures as equally valid positive samples, entirely ignoring binding affinity and omitting crucial negative signals from weak binders, decoys, or non-binders. This spatial-only approach is further constrained by the severe scarcity of high-quality experimental data. While the Protein Data Bank contains hundreds of thousands of structures, curated protein–ligand collections relevant for generation, such as PDBbind (Wang et al., 2005) and Binding MOAD (Hu et al., 2005), contain only tens of thousands of complexes. These datasets are heavily biased toward crystallizable, well-studied targets, fundamentally limiting the chemical and pocket diversity available to the generative model.

3.1 REPRESENTATION ALIGNMENT WITH A PRETRAINED ATOMISTIC ENCODER

We address the limitations of geometry-only supervision by aligning the generative model's latent representations with a physics-aware prior. Our framework (Fig. 2) couples a **pocket-conditioned denoiser** f_θ with a **frozen atomistic encoder** $\mathcal{E}$, here a pretrained machine learning interatomic potential (MLIP) whose features implicitly encode physical properties such as interatomic forces. By aligning the denoiser's intermediate features with those of $\mathcal{E}$ through a lightweight projection, we provide the model with a continuous gradient of chemical feasibility that extends beyond simple coordinate-wise reconstruction.

This alignment compensates for the lack of energetic information in standard denoising by supervising the latent space with representations sensitive to local physical environments. By leveraging an encoder pretrained on massive atomistic datasets, the denoiser inherits a foundational understanding of molecular physics that bypasses the scarcity of experimental co-crystal structures. Crucially, since these interatomic potentials are trained on configurations including repulsive and unfavorable interactions, they provide a vital "negative" signal that teaches the model to avoid high-strain geometries. This inductive bias is distilled during training, allowing $\mathcal{E}$ to be discarded at test time with *no additional computational overhead during inference*.

3.2 PROPOSED LEARNING FORMULATION

To realize the physics-guided prior described above, our representation alignment scheme pulls the internal latent space of the generative model toward the structured energy landscape of the MLIP. As illustrated in Figure 3, we achieve this by extracting and matching intermediate feature representations for each ligand atom i from both the trainable 3D molecule denoiser f_θ and the frozen physics encoder $\mathcal{E}$. Both networks output two distinct types of per-atom features that reflect the geometric character of the molecular system. *Scalar* features are $\mathrm{SO}(3)$-invariant ($\ell = 0$ channels) and encode frame-independent chemical information such as atom type, hybridization, and local bonding context. *Vector* features are $\mathrm{SO}(3)$-equivariant ($\ell = 1$ channels) and capture directional information such as interatomic forces and relative atomic displacements. Crucially, under a global rotation $R \in \mathrm{SO}(3)$, these vector features transform equivariantly as $\mathbf{V} \mapsto R\mathbf{V}$. Preserving this transformation in the projection step is what makes the vector alignment loss (see $\mathcal{L}_{vec}$, Eq. 4) a rotation-invariant scalar, so the penalty on a generated pose depends only on its internal geometry and strain, not on its orientation in the global coordinate frame.

Concretely, as the denoiser f_θ processes the noisy ligand $(\mathbf{X}_t, \mathbf{a}_t, \mathbf{b}_t)$ conditioned on the pocket $(\mathbf{X}_p, \mathbf{a}_p)$, we intercept the hidden states at a chosen intermediate layer L. For a given ligand atom i,

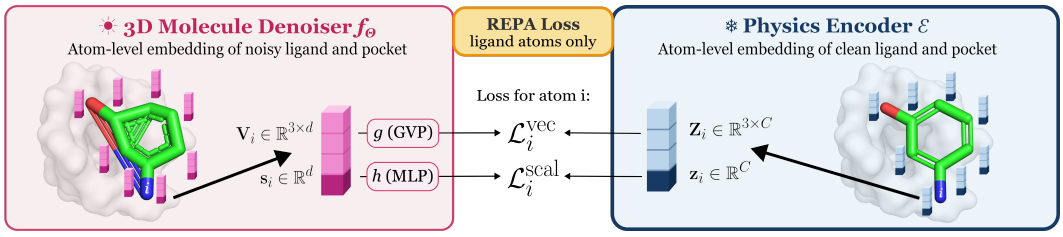

Figure 3: **Representation alignment loss.** For each ligand atom i, intermediate latent representations, contextualized from the entire protein-ligand complex, are extracted from both the trainable 3D molecule denoiser f_θ and a frozen physics encoder $\mathcal{E}$ (parameterized by a pretrained MLIP). Because latent dimensions differ, equivariant projectors (g and h) map the denoiser's scalar (s_i) and vector (V_i) features into the physics encoder's feature space (z_i and Z_i). We minimize a cosine similarity loss which regularizes the embedding space, which we find helps avoid high strain.

we extract these intermediate embeddings as:

$$\underbrace{\mathbf{s}_i \in \mathbb{R}^d}_{\text{scalar features}}, \qquad \underbrace{\mathbf{V}_i \in \mathbb{R}^{3\times d}}_{\text{3D vector features}}. \tag{1}$$

Similarly, by passing the clean ligand $(\mathbf{X}_0, \mathbf{a}_0, \mathbf{b}_0)$ conditioned on the pocket $(\mathbf{X}_p, \mathbf{a}_p)$ through the frozen MLIP encoder $\mathcal{E}$ up to layer K, we extract the corresponding target latent representations for atom i:

$$\underbrace{\mathbf{z}_i \in \mathbb{R}^C}_{\text{scalar target}}, \qquad \underbrace{\mathbf{Z}_i \in \mathbb{R}^{3\times C}}_{\text{3D vector targets}}, \tag{2}$$

where d and C denote the latent channel dimensions of the denoiser and the encoder, respectively.

Equivariant projectors. Our goal is to align $\mathbf{s}_i$ with $\mathbf{z}_i$ (scalar-to-scalar) and $\mathbf{V}_i$ with $\mathbf{Z}_i$ (vector-to-vector). Because the generative denoiser and the pre-trained encoder operate in different latent dimensions ($d \neq C$), we introduce two lightweight projectors that map the denoiser's features directly into the encoder's representation space:

$$h : \mathbb{R}^d \to \mathbb{R}^C, \qquad g : \mathbb{R}^{3\times d} \to \mathbb{R}^{3\times C}. \tag{3}$$

We implement the scalar projector h as a standard Multi-Layer Perceptron (MLP). For the vector channels, we must satisfy the equivariance requirement established above; therefore, we implement the projector g as an $\mathrm{SO}(3)$-equivariant Geometric Vector Perceptron (GVP) (Jing et al., 2021).

Representation alignment loss. With both representations in the same space, we can directly penalize their deviation. For each ligand atom i, we define the alignment loss as the negative cosine similarity between the projected denoiser features and the frozen MLIP targets, computed independently for the scalar and vector channels:

$$\mathcal{L}_i^{\text{scal}} = -\frac{\langle h(\mathbf{s}_i),\, \mathbf{z}_i \rangle}{\|h(\mathbf{s}_i)\|\, \|\mathbf{z}_i\|}, \qquad \mathcal{L}_i^{\text{vec}} = -\frac{\langle g(\mathbf{V}_i),\, \mathbf{Z}_i \rangle_F}{\|g(\mathbf{V}_i)\|_F\, \|\mathbf{Z}_i\|_F}, \tag{4}$$

where $\langle \cdot, \cdot \rangle$ denotes the standard Euclidean inner product and $\langle \cdot, \cdot \rangle_F, \|\cdot\|_F$ denote the Frobenius inner product and norm on $\mathbb{R}^{3\times C}$. Crucially, because g is equivariant and the Frobenius inner product cancels orthogonal rotation matrices ($R^T R = I$), the vector loss $\mathcal{L}_i^{vec}$ inherently evaluates to a rotation-invariant scalar. The loss for the full ligand is then the average over its N_ℓ heavy atoms.

Joint training objective. Finally, these physics-aware alignment terms are coupled with the standard geometry-based denoising objective ($\mathcal{L}_{denoise}$). The denoiser and the lightweight projectors (h, g) are trained jointly to minimize the composite loss:

$$\mathcal{L} = \mathcal{L}_{\text{denoise}} + \mathcal{L}_{\text{REPA}}, \qquad \mathcal{L}_{\text{REPA}} = \lambda_{\text{scal}} \frac{1}{N_\ell} \sum_{i=1}^{N_\ell} \mathcal{L}_i^{\text{scal}} + \lambda_{\text{vec}} \frac{1}{N_\ell} \sum_{i=1}^{N_\ell} \mathcal{L}_i^{\text{vec}}. \tag{5}$$

where λ_{scal} and $\lambda_{\text{vec}} \geq 0$ are hyperparameters governing the strength of the alignment. Because the MLIP encoder $\mathcal{E}$ remains completely frozen, and the projectors h and g are only required to calculate the alignment gradient during training, all three components can be safely discarded at sampling time and do not introduce any computational overhead during inference.

4 EXPERIMENTAL SETUP

Models and implementation. We implement representation alignment (REPA) on two pocket-conditioned 3D generative models built on flow matching (Lipman et al., 2022): DrugFlow (Schneuing et al., 2025) and FLOWR (Cremer et al., 2026). They propagate information in fundamentally different ways, which makes them a useful pair for testing whether alignment acts as a general inductive bias. DrugFlow uses an E(3)-equivariant heterogeneous graph network of geometric vector perceptrons, passing messages over a 10 Å cutoff graph with the pocket coarse-grained to one node per residue, so its receptive field is local and grows with depth. FLOWR builds on SEMLA (Irwin et al., 2025), combining equivariant self-attention over all ligand atoms with cross-attention into a separately encoded all-atom pocket, so every atom attends to every other at every layer. DrugFlow thus carries a stronger local geometric bias, while FLOWR reasons globally. As the frozen physics encoder we use UMA (Wood et al., 2025), an MLIP pretrained on large-scale quantum-mechanical data spanning broad chemistry and off-equilibrium configurations. Architectural details, alignment layers L and K, projectors, loss weights, and remaining hyperparameters are given in Section B.

Datasets. We train on each model's established dataset. For DrugFlow we first use the standard CrossDocked2020 (Francoeur et al., 2020) splits for direct comparison with prior work, but docking-augmented datasets are inherently limited: because denoising objectives require positive examples, cross-docked pools are filtered to poses within a small RMSD of a reference crystal structure. This enforces redundancy, since surviving poses are by construction close to known structures, and discards energetically favourable but structurally distinct poses alongside true failures, as RMSD is a purely geometric proxy. We therefore also train both models on SPINDR, a more challenging modern split of high-quality crystallographic complexes only.

Metrics. *PoseBusters* is the rule-based pass rate of Buttenschoen et al. (2024) on the generated pose. *Strain* is the MMFF94s local strain energy (Halgren, 1996) reported as the median over a pocket's molecules then averaged over pockets so each pocket contributes equally. *Vina* is the AutoDock Vina score of the pose in its target pocket. *High quality (HQ)*, our primary objective, is the conjunction of three criteria: PoseBusters-valid, strain $\leq 20\,\text{kcal mol}^{-1}$, and Vina $\leq -6\,\text{kcal mol}^{-1}$. PoseBusters and HQ are fractions of *all* sampled molecules; strain, Vina and strain-threshold rates are conditional on the PoseBusters column, i.e. computed over molecules that are PoseBusters-valid in the pose state being reported, which keeps degenerate or severely clashing geometries out of energetic averages. *PLIF recovery* is the fraction of the crystallographic ligand's protein–ligand interaction fingerprint reproduced by the generated molecule. *MCS* is the maximum common substructure with the crystallographic ligand, normalised by that ligand's size. Significance is a paired Wilcoxon signed-rank test over targets with seeds pooled, comparing each REPA model against its own baseline ($^{***}p < 10^{-3}$, $^{**}p < 10^{-2}$, $^{*}p < 0.05$); per-pocket aggregates carry 95% bootstrap intervals over pockets. Bold marks the better of each baseline/REPA pair except in Table 1, where it marks the best value among all generative models and underline the second. Native ligands are a reference ceiling throughout and are excluded from every ranking.

Baselines. This gives three evaluation setups: DrugFlow on CrossDocked, DrugFlow on SPINDR, and FLOWR on SPINDR, each compared against the original model without alignment. To compare against the strongest possible baseline, we augment DrugFlow with an exponential moving average (EMA) of the denoiser weights (Ho et al., 2020); since DrugFlow + EMA strictly outperforms vanilla DrugFlow, we adopt it as the primary baseline and apply alignment on top of it, so that any gain is attributable to the physics-guided inductive bias rather than to training stabilisation. Ablation of this is provided in Section C.1.

Table 1: **Evaluation on CrossDocked2020.** External baselines use single sample sets, while DrugFlow variants are averaged over 3 seeds.

Model	Size	PB ↑	Strain ↓	Vina ↓
Pocket2Mol	17.72	0.719	141.3	-5.49
DiffSBDD	24.63	0.357	261.6	-5.41
TargetDiff	22.97	0.491	217.8	-5.74
DrugFlow	21.86	0.732	<u>77.7</u>	**-6.05**
+ REPA	21.94	**0.789**	**71.4**	-5.85
Native ligands	22.75	0.950	71.3	-6.78

Table 2: **Evaluation on SPINDR.** Representation alignment added to DrugFlow and FLOWR models. Results are averaged over 3 seeds.

Method	Size	PB ↑	Strain ↓	Vina ↓
DrugFlow	24.63	0.745	52.9	**-6.33**
+ REPA	24.94	**0.808**	**51.3**	-6.20
FLOWR	26.52	0.890	**92.4**	-7.00
+ REPA	26.52	**0.896**	101.1	**-7.04**
Native ligands	26.68	1.000	31.1	-7.54

5 RESULTS

5.1 COMPARISON WITH PUBLISHED MODELS ON CROSSDOCKED2020

We first position REPA against published work on the standard CrossDocked2020 benchmark, then move to SPINDR — challenging dataset without docked complexes — for the analyses that follow. The CrossDocked test set contains 100 target pockets; we sample 100 ligands per target. Our primary goal is to evaluate how adding REPA improves the DrugFlow model. Both models are trained for 600 epochs and evaluated under the same protocol, with results averaged over three seeds. For broader comparison, we additionally include Pocket2Mol (Peng et al., 2022), DiffSBDD (Schneuing et al., 2022), and TargetDiff (Guan et al., 2023) using their released CrossDocked checkpoints and default sampling settings.

The results are shown in Table 1. REPA gives the highest PoseBusters pass rate of any method evaluated, improving over DrugFlow from 0.732 to 0.789 against a reference-ligand ceiling of 0.950. It also gives the lowest median local strain of any method evaluated, $77.7 \rightarrow 71.4\,\text{kcal mol}^{-1}$ ($p = 7.4 \times 10^{-7}$) — a value close to the $71.3\,\text{kcal mol}^{-1}$ of the reference ligands of the same pockets. We note that the CrossDocked reference ligands include synthetic, cross-docked poses, which is why their strain is high in absolute terms; the comparison shows that REPA reaches the benchmark's own reference level, not that it reaches crystallographic quality. Mean molecule size moves toward the reference as well, so the strain gain is not obtained by generating smaller molecules.

The Vina score moves the other way, from -6.05 to -5.85. We believe this reflects a physical trade-off between internal strain and pocket complementarity rather than weaker binding: the same trade-off appears in the crystallographic ligands themselves, whose Vina score worsens from -7.54 to -6.69 once relaxation relieves their strain (Section 5.3).

Figure 1 and Figure 6 show that REPA also accelerates training substantially: the aligned model matches the final PoseBusters and strain of DrugFlow after roughly six times fewer epochs, and the alignment objective adds no cost at sampling time, since the MLIP and the projectors are discarded after training. Vina is the exception: it plateaus early and does not improve further with training. Section C.2 further shows that the PoseBusters and strain improvements are not specific to alignment with UMA and hold for other MLIP encoders.

5.2 SPINDR: TRANSFER TO A SECOND DATASET AND A SECOND ARCHITECTURE

We next train and evaluate both models on SPINDR, a modern split built from high-quality crystallographic complexes only, with strict pocket-level clustering to control leakage between train and test. The difference in data quality is visible in the reference ligands themselves: the crystallographic ligands of the SPINDR test pockets carry a median local strain of $31.1\,\text{kcal mol}^{-1}$ against 71.3 for the cross-docked CrossDocked references, and score -7.54 on Vina against -6.78. Because SPINDR provides genuine crystallographic references and a stricter split, we use it as the primary benchmark for the analyses that follow.

The DrugFlow improvements survive the change of training data and target distribution (Table 2). Against DrugFlow, REPA raises the PoseBusters pass rate from 0.745 to 0.808 and lowers median strain from 52.9 to $51.3\,\text{kcal mol}^{-1}$, with the same Vina cost seen in Section 5.1 ($-6.33 \rightarrow -6.20$).

Testing on the SPINDR dataset also allows us to evaluate against a second strong baseline model. FLOWR is a useful test because it has very different baseline performance trends than DrugFlow: it places molecules in the site better (Vina -7.00 vs. -6.33) but resolves local geometry worse (strain 92.4 vs. 52.9). Interestingly, REPA does not have a significant positive effect on strain or PoseBusters validity of this model. And even further, the strain of the molecule increases: $92.4 \rightarrow 101.1\,\text{kcal mol}^{-1}$.

Overall, strain remains high relative to the native ligands (last row in (Table 2) for both DrugFlow and Flowr models. For this reason we focus on evaluating both architectures after pocket-constrained minimization, which we discuss in the next section.

5.3 GENERATING HIGH-QUALITY CANDIDATES

The ultimate goal of structure-based drug design is not achieving a high score on any single computational metric, but identifying actionable candidates worthy of experimental verification.

Table 3: **SPINDR: quality gates before and after pocket-constrained energy minimisation.** Each cell reads *as generated* → *after minimisation*. Within each backbone, **bold** marks the value significantly better between the baseline and its REPA variant (paired Wilcoxon signed-rank over pockets, $p < 0.05$); pairs with no bold do not differ significantly. Higher is better in every column. Native ligands of the same pockets are the reference ceiling.

Each cell: *as generated* → *after minimisation*

Method	PoseBusters ↑	Strain ≤ 20 ↑	Vina ≤ −6 ↑	High quality ↑
DrugFlow	0.745 → 0.708	0.114 → 0.584	**0.553** → **0.486**	0.037 → 0.214
+ REPA	**0.808** → **0.773**	0.119 → **0.598**	0.532 → 0.476	0.043 → **0.239**
FLOWR	0.890 → 0.831	**0.018** → 0.540	0.660 → 0.547	**0.010** → 0.279
+ REPA	**0.896** → 0.828	0.009 → **0.555**	**0.669** → 0.546	0.005 → **0.292**
Native ligands	1.000 → 1.000	0.257 → 0.676	0.752 → 0.626	0.180 → 0.459

A viable candidate must simultaneously be chemically sound, physically well-posed, low in internal strain, and complementary to the binding pocket. Currently, raw generated molecules approach native ligands in Vina scores (-6.20 to -7.04 vs. -7.54) but sit in highly strained conformations (52.9 and 92.4 vs. 31.1 kcal mol^{-1}). Because generative sampling is computationally cheap, we can filter out suboptimal outputs; consequently, a model's true utility is defined by its yield of high-quality molecules after standard preparation, rather than the average quality of its raw samples.

Inspired by traditional virtual screening workflows (Madhavi Sastry et al., 2013), we make this pipeline-centric evaluation explicit by scoring every pose twice: as generated, and after a restrained in-situ (pocket-constrained) minimisation. To gently optimize the pose and alleviate strain without allowing the molecule to drift into a fundamentally new binding mode, the pocket is held fixed and the ligand's heavy atoms are harmonically restrained to within 1 Å of their generated positions for 1000 MMFF94s steps. We apply this procedure to the SPINDR test set; because its reference complexes are experimental crystal structures and its data split is the stricter of the two, it better quantifies true generalization to real-world scenarios. Ultimately, we ask whether the improvements observed in Sections 5.1 and 5.2 survive this refinement step and translate into a higher volume of surviving, promising high-quality candidates.

Relaxing the native crystallographic ligands themselves establishes the baseline behaviour, and settles the Vina trade-off noted in Section 5.1. Their low-strain fraction rises sharply ($0.257 \rightarrow 0.676$) while their Vina worsens ($-7.54 \rightarrow -6.69$), because a ligand relaxing toward its own energy minimum moves off its tightest contacts with the pocket (Table 3). Because this tension between empirical docking scores and internal strain exists even in ground-truth native structures, a slight trade-off during relaxation is expected physical behavior rather than just a model artifact.

Relaxation results. Relaxation sharply reduces strain for every model, baseline and REPA alike: the low-strain fraction rises from 0.114 to 0.584 for DrugFlow and from 0.018 to 0.540 for FLOWR (Table 3). The native trade-off reappears — relaxation trims the fraction clearing the Vina threshold, $0.553 \rightarrow 0.486$ for DrugFlow — but the high-quality yield, which requires low strain and a good docking score at once, climbs steeply for every model ($0.037 \rightarrow 0.214$ and $0.010 \rightarrow 0.279$). The question is whether REPA's advantage survives this equalising step.

It does, for both backbones, on the metric that conjoins all four criteria. After relaxation the aligned models raise the high-quality yield from 0.214 to 0.239 for DrugFlow and from 0.279 to 0.292 for FLOWR (both $p < 10^{-3}$), against a native-ligand ceiling of 0.459 — closing 10% and 7% of the remaining gap to real bound ligands. Both aligned models also retain higher low-strain fraction Table 3. The PoseBusters advantage behaves differently on the two backbones: DrugFlow keeps a 6.5-point lead after relaxation (0.708 vs. 0.773), while FLOWR's pre-relaxation advantage does not survive (0.831 vs. 0.828). We also note that relaxation lowers the PoseBusters pass rate of every generated set while leaving the native ligands unchanged at 1.000. This drop occurs because the relaxation process can occasionally push atoms into slight steric clashes with the protein pocket, or cause severely malformed—yet initially valid—molecules to break apart into separate fragments during force-field minimization.

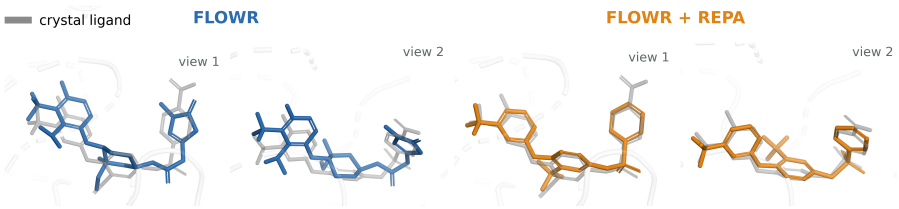

Figure 4: **Example of improved results.** The best MCS molecule each model produces for target 5yea, shown in two views of the same pose 60° apart, with the true crystallographic ligand behind it in grey. The aligned model (FLOWR + REPA) generates a structural scaffold that much more closely resembles the native ligand (MCS $0.545 \rightarrow 0.848$) while simultaneously exhibiting a highly favorable, low-strain geometry (relaxed strain $20.6 \rightarrow 7.8\,\mathrm{kcal\,mol^{-1}}$, Vina $-9.73\,\mathrm{kcal\,mol^{-1}}$). This demonstrates that alignment successfully guides the generator toward structurally realistic and highly promising candidate binders.

5.4 Interaction recovery and molecular identity

Every metric evaluated so far scores the spatial pose. A model could theoretically improve on those metrics without changing the generated chemistry simply by placing the same atoms better. The Maximum Common Substructure (MCS) metric cannot be improved this way. MCS measures the largest connected chemical fragment a generated molecule shares with the target's native ligand, expressed as a fraction of that native ligand's size (higher is better). Because it is computed strictly on the molecular graph, it evaluates the underlying chemical identity and is entirely invariant to the spatial pose. MCS rises for both backbones ($0.265 \rightarrow 0.275$ for DrugFlow and $0.307 \rightarrow 0.322$ for FLOWR, both $p < 10^{-3}$; Table 4). Alignment therefore actively steers the model toward generating native-like chemical identities, not merely better conformations of the same chemistry. Figure 4 shows a representative case of this improved scaffold recovery.

Table 4: **Interaction recovery and molecular identity on SPINDR.** PLIF Tanimoto additionally penalises interactions absent from the native ligand; MCS excludes exact reproductions.

Method	PLIF rec.	Tanimoto	MCS
DrugFlow	0.412	0.362	0.265
+ REPA	**0.423**	**0.373***	**0.275*****
FLOWR	0.452	0.399	0.307
+ REPA	**0.475*****	**0.417*****	**0.322*****
Native ligands	1.000†	1.000†	1.000†

Higher is better throughout. † perfect by construction.

Moreover, interaction recovery improves under both the absolute recovery rate—which measures the fraction of true native contacts successfully reproduced—and the PLIF Tanimoto similarity, which evaluates the exact overlap of interaction fingerprints and strictly penalises spurious contacts that the native ligand does not make. This benefit carries to the other tasks FLOWR supports: recovery improves in all four inpainting modes and high quality in three of them (Table 13).

Crucially, these gains do not compromise other vital molecular properties. Synthetic accessibility, an estimate of how easily the generated structures could be synthesized in a wet lab, remains unchanged for FLOWR and for DrugFlow (Table 12). Novelty (the fraction of generated molecules entirely unseen during training) remains unchanged and essentially saturated. Uniqueness and structural diversity each fall by only one to two points—a modest cost that, given the consistently high novelty, confirms the model is actively exploring valid chemical space rather than simply collapsing toward memorised training examples (Table 11).

6 Conclusion

We introduced an equivariant representation alignment scheme that transfers the internal representations of a frozen machine-learned interatomic potential into pocket-conditioned 3D molecule generation, at no cost at sampling time. Across two datasets and two architectures, aligned models produce more physically plausible molecules with higher validity and lower internal strain. Furthermore, alignment changes which molecules are generated and not only how they are posed, giving greater substructure similarity to native binders and improved protein-ligand interaction fingerprint recovery. Crucially, by applying a pocket-constrained minimization, we demonstrate that many geometric imperfections in raw generated poses are easily fixable. Even after this equalizing relaxation step, the fundamental physical advantages of our method survive: the aligned models yield a significantly higher fraction of high-quality candidates that are simultaneously valid, low-strain, and well-docked, closing 10% and 7% of the gap to native ligands.

AI USE STATEMENT

We used Claude and Claude Code (Anthropic) at several stages of this project. The idea of using an MLIP to improve pocket-conditioned generative models came from us and from the literature; we used Claude to sanity-check it, to suggest related work and adjacent directions worth considering, and we then read the papers ourselves. Claude Code implemented the method from our specification, and we reviewed the code and corrected it. During development, Claude Code was used to write training configurations and to launch training runs and evaluations; the configurations were checked manually before use. Claude also gave feedback on our experimental design, helped refine the hypotheses we present in the appendix, and was used in interpreting results; in each case the conclusions were checked against the data and accepted or rejected by us. Beyond this, Claude helped write plotting scripts and was used to improve the writing: we supplied the ideas and arguments and then iterated on the formulation with Claude, editing and rejecting its output as we went. We did not use AI tools to generate data or to clean or reformat datasets. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Our work builds on two publicly released models, DrugFlow (Schneuing et al., 2025) and FLOWR (Cremer et al., 2026), whose published results we reproduce before adding the alignment objective, and on UMA (Wood et al., 2025), which is also publicly available. Both benchmarks, CrossDocked2020 and SPINDR, are public. The alignment objective and the encoders, layers and hyperparameters used are described in Section 4, with further implementation details and ablations in the Sections C and D. All main results are averaged over three seeds. We will release our code upon acceptance.

REFERENCES

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

Luis Barros-Luque, Muhammed Shuaibi, Xiang Fu, Brandon M. Wood, Misko Dzamba, Meng Gao, Ammar Rizvi, Matt Uyttendaele, C. Lawrence Zitnick, and Zachary W. Ulissi. The Open Materials 2024 (OMat24) inorganic materials dataset and models. *Nature Computational Science*, 6(6):642–652, June 2026. ISSN 2662-8457. doi: 10.1038/s43588-026-00996-w. URL https://www.nature.com/articles/s43588-026-00996-w.

Ilyes Batatia, David P Kovacs, Gregor Simm, Christoph Ortner, and Gabor Csanyi. MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields. In *NeurIPS*, 2022.

Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M. Elena, Dávid P. Kovács, Janosh Riebesell, Xavier R. Advincula, Mark Asta, Matthew Avaylon, William J. Baldwin, Fabian Berger, Noam Bernstein, Arghya Bhowmik, Filippo Bigi, Samuel M. Blau, Vlad Cărare, Michele Ceriotti, Sanggyu Chong, James P. Darby, Sandip De, Flaviano Della Pia, Volker L. Deringer, Rokas Elijošius, Zakariya El-Machachi, Edvin Fako, Fabio Falcioni, Andrea C. Ferrari, John L. A. Gardner, Mikołaj J. Gawkowski, Annalena Genreith-Schriever, Janine George, Rhys E. A. Goodall, Jonas Grandel, Clare P. Grey, Petr Grigorev, Shuang Han, Will Handley, Hendrik H. Heenen, Kersti Hermansson, Cheuk Hin Ho, Stephan Hofmann, Christian Holm, Jad Jaafar, Konstantin S. Jakob, Hyunwook Jung, Venkat Kapil, Aaron D. Kaplan, Nima Karimitari, James R. Kermode, Panagiotis Kourtis, Namu Kroupa, Jolla Kullgren, Matthew C. Kuner, Domantas Kuryla, Guoda Liepuoniute, Chen Lin, Johannes T. Margraf, Ioan-Bogdan Magdău, Angelos Michaelides, J. Harry Moore, Aakash A. Naik, Samuel P. Niblett, Sam Walton Norwood, Niamh O'Neill, Christoph Ortner, Kristin A. Persson, Karsten Reuter, Andrew S. Rosen, Louise A. M. Rosset, Lars L. Schaaf, Christoph Schran, Benjamin X. Shi, Eric Sivonxay, Tamás K. Stenczel, Christopher Sutton, Viktor Svahn, Thomas D. Swinburne, Jules Tilly, Cas van der Oord, Santiago Vargas, Eszter Varga-Umbrich, Tejs Vegge, Martin Vondrák, Yangshuai Wang, William C. Witt, Thomas Wolf, Fabian Zills, and Gábor Csányi. A foundation model for atomistic materials chemistry. *The Journal of Chemical Physics*, 163(18):184110, November 2025. ISSN 0021-9606. doi: 10.1063/5.0297006. URL https://doi.org/10.1063/5.0297006.

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13(1):2453, May 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-29939-5. URL https://www.nature.com/articles/s41467-022-29939-5.

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*, 2024.

Julian Cremer, Ross Irwin, Alessandro Tibo, Jon Paul Janet, Simon Olsson, and Djork-Arné Clevert. FLOWR: flow matching for structure-aware de novo, interaction- and fragment-based ligand generation. *Nature Computational Science*, 6(6):565–574, June 2026. ISSN 2662-8457. doi: 10.1038/s43588-026-00998-8. URL https://www.nature.com/articles/s43588-026-00998-8.

Kieran Didi. The unification of representation learning and generative modelling, 2025. URL https://kdidi.netlify.app/blog/ml/2025-12-31-r4g/.

Janani Durairaj, Yusuf Adeshina, Zhonglin Cao, Xuejin Zhang, Vladas Oleinikovas, Thomas Duignan, Zachary McClure, Xavier Robin, Emanuele Rossi, Guoqing Zhou, et al. Plinder: The protein-ligand interactions dataset and evaluation resource. In *ICML'24 Workshop ML for Life and Material Science: From Theory to Industry Applications*, 2024.

Peter Eastman, Pavan Kumar Behara, David L. Dotson, Raimondas Galvelis, John E. Herr, Josh T. Horton, Yuezhi Mao, John D. Chodera, Benjamin P. Pritchard, Yuanqing Wang, Gianni

De Fabritiis, and Thomas E. Markland. SPICE, A Dataset of Drug-like Molecules and Peptides for Training Machine Learning Potentials. *Scientific Data*, 10(1):11, January 2023. ISSN 2052-4463. doi: 10.1038/s41597-022-01882-6. URL https://www.nature.com/articles/s41597-022-01882-6.

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

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

Meng Gao, Armin Shayesteh Zadeh, Aniruddha Seal, Siva Dasetty, Siddarth K. Achar, Misko Dzamba, Benjamin K. Miller, Leif D. Jacobson, C. Lawrence Zitnick, Brandon M. Wood, Zachary W. Ulissi, Daniel S. Levine, and Andrew L. Ferguson. Quantum-accurate atomistic modeling of enzyme catalysis using a machine learned potential, September 2026. URL http://arxiv.org/abs/2609.09293. arXiv:2609.09293 [physics.chem-ph].

Johannes Gasteiger, Florian Becker, and Stephan Günnemann. GemNet: Universal Directional Graph Neural Networks for Molecules. In *NeurIPS*, November 2021. URL https://openreview.net/forum?id=HS_sOaxS9K-.

Vahe Gharakhanyan, Luis Barroso-Luque, Yi Yang, Muhammed Shuaibi, Kyle Michel, Daniel S. Levine, Misko Dzamba, Xiang Fu, Meng Gao, Xingyu Liu, Haoran Ni, Keian Noori, Brandon M. Wood, Matt Uyttendaele, Arman Boromand, C. Lawrence Zitnick, Noa Marom, Zachary W. Ulissi, and Anuroop Sriram. Open Molecular Crystals 2025 (OMC25) dataset and models. *Scientific Data*, 13(1):354, February 2026. ISSN 2052-4463. doi: 10.1038/s41597-026-06628-2. URL https://www.nature.com/articles/s41597-026-06628-2.

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. *arXiv preprint arXiv:2303.03543*, 2023.

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. doi: https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6<490::AID-JCC1>3.0.CO;2-P. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291096-987X%28199604%2917%3A5/6%3C490%3A%3AAID-JCC1%3E3.0.CO%3B2-P.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3D. In *ICML*. PMLR, 2022.

Liegi Hu, Mark L Benson, Richard D Smith, Michael G Lerner, and Heather A Carlson. Binding moad (mother of all databases). *Proteins: Structure, Function, and Bioinformatics*, 60(3):333–340, 2005.

Zhilin Huang, Ling Yang, Xiangxin Zhou, Chujun Qin, Yijie Yu, Xiawu Zheng, Zikun Zhou, Wentao Zhang, Yu Wang, and Wenming Yang. Interaction-based Retrieval-augmented Diffusion Models for Protein-specific 3D Molecule Generation. In *ICML*, 2024a. URL https://openreview.net/forum?id=eejhD9FCP3.

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 *ICLR*, 2024b. URL https://openreview.net/forum?id=qH9nrMNTIW.

Ross Irwin, Alessandro Tibo, Jon Paul Janet, and Simon Olsson. SemlaFlow – Efficient 3D Molecular Generation with Latent Attention and Equivariant Flow Matching. In *The 28th International Conference on Artificial Intelligence and Statistics*, February 2025. URL https://openreview.net/forum?id=bee2G6pEh0.

Ryan Jacobs, Dane Morgan, Siamak Attarian, Jun Meng, Chen Shen, Zhenghao Wu, Clare Yijia Xie, Julia H. Yang, Nongnuch Artrith, Ben Blaiszik, Gerbrand Ceder, Kamal Choudhary, Gabor Csanyi, Ekin Dogus Cubuk, Bowen Deng, Ralf Drautz, Xiang Fu, Jonathan Godwin, Vasant Honavar, Olexandr Isayev, Anders Johansson, Boris Kozinsky, Stefano Martiniani, Shyue Ping Ong, Igor Poltavsky, K. J. Schmidt, So Takamoto, Aidan P. Thompson, Julia Westermayr, and Brandon M. Wood. A practical guide to machine learning interatomic potentials – Status and future. *Current Opinion in Solid State and Materials Science*, 35:101214, 2025. ISSN 1359-0286. doi: https://doi.org/10.1016/j.cossms.2025.101214. URL https://www.sciencedirect.com/science/article/pii/S1359028625000014.

Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael JL Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. In *ICLR*, 2021.

Jakub Kopko, David Graber, Saltuk Mustafa Eyrilmez, Stanislav Mazurenko, David Bednar, Jiri Sedlar, and Josef Sivic. Generalization Beyond Benchmarks: Evaluating Learnable Protein-Ligand Scoring Functions on Unseen Targets, 2025. URL https://arxiv.org/abs/2512.05386.

David Peter Kovacs, J. Harry Moore, Nicholas J. Browning, Ilyes Batatia, Joshua T. Horton, Yixuan Pu, Venkat Kapil, William C. Witt, Ioan-Bogdan Magdau, Daniel J. Cole, and Gabor Csanyi. Mace-off: Short-range transferable machine learning force fields for organic molecules. *Journal of the American Chemical Society*, 147(21):17598–17611, 05 2025. ISSN 0002-7863. doi: 10.1021/jacs.4c07099. URL https://doi.org/10.1021/jacs.4c07099.

Adam Lahouari, Jutta Rogal, and Mark E. Tuckerman. Automated Machine Learning Pipeline: Large Language Models-Assisted Automated Data set Generation for Training Machine-Learned Interatomic Potentials. *Journal of Chemical Theory and Computation*, 22(1):305–317, January 2026. ISSN 1549-9618. doi: 10.1021/acs.jctc.5c01610. URL https://doi.org/10.1021/acs.jctc.5c01610.

Daniel S. Levine, Muhammed Shuaibi, Evan Walter Clark Spotte-Smith, Michael G. Taylor, Muhammad R. Hasyim, Kyle Michel, Ilyes Batatia, Gábor Csányi, Misko Dzamba, Peter Eastman, Nathan C. Frey, Xiang Fu, Vahe Gharakhanyan, Aditi S. Krishnapriyan, Joshua A. Rackers, Sanjeev Raja, Ammar Rizvi, Andrew S. Rosen, Zachary Ulissi, Santiago Vargas, C. Lawrence Zitnick, Samuel M. Blau, and Brandon M. Wood. The Open Molecules 2025 (OMol25) Dataset, Evaluations, and Models, March 2026. URL http://arxiv.org/abs/2505.08762. arXiv:2505.08762 [physics.chem-ph].

Zhenzhu Li and Aron Walsh. Platonic representation of foundation machine learning interatomic potentials. *Nature Machine Intelligence*, 8(5):830–840, May 2026. ISSN 2522-5839. doi: 10.1038/s42256-026-01235-7. URL https://www.nature.com/articles/s42256-026-01235-7.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

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.

G. Madhavi Sastry, Matvey Adzhigirey, Tyler Day, Ramakrishna Annabhimoju, and Woody Sherman. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *Journal of Computer-Aided Molecular Design*, 27(3):221–234, March 2013. ISSN 1573-4951. doi: 10.1007/s10822-013-9644-8. URL https://doi.org/10.1007/s10822-013-9644-8.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas

Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Xingang Peng, Shitong Luo, Jiaqi Guan, Qi Xie, Jian Peng, and Jianzhu Ma. Pocket2Mol: Efficient molecular sampling based on 3D protein pockets. In *ICML*, 2022.

Lucas Pinede, Soojung Yang, Juno Nam, and Rafael Gomez-Bombarelli. Unifying force prediction and molecular conformation generation through representation alignment. In *ICML Generative AI and Biology (GenBio) Workshop*, 2025. URL https://openreview.net/forum?id=yzkHGHvC74.

Arne Schneuing, Yuanqi Du, Charles Harris, Arian Jamasb, Ilia Igashov, Weitao Du, Tom Blundell, Pietro Lió, Carla Gomes, Max Welling, Michael Bronstein, and Bruno Correia. Structure-based drug design with equivariant diffusion models. *arXiv preprint arXiv:2210.13695*, 2022.

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

Peter Škrinjar, Jérôme Eberhardt, Gerardo Tauriello, Torsten Schwede, and Janani Durairaj. Have protein-ligand cofolding methods moved beyond memorisation? *bioRxiv*, 2025. doi: 10.1101/2025.02.03.636309. URL https://www.biorxiv.org/content/early/2025/08/04/2025.02.03.636309.

Chenyu Wang, Cai Zhou, Sharut Gupta, Zongyu Lin, Stefanie Jegelka, Stephen Bates, and Tommi Jaakkola. Learning diffusion models with flexible representation guidance. In *NeurIPS*, 2025a.

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.

Yuanqing Wang, Kenichiro Takaba, Michael S. Chen, Marcus Wieder, Yuzhi Xu, Tong Zhu, John Z. H. Zhang, Arnav Nagle, Kuang Yu, Xinyan Wang, Daniel J. Cole, Joshua A. Rackers, Kyunghyun Cho, Joe G. Greener, Peter Eastman, Stefano Martiniani, and Mark E. Tuckerman. On the design space between molecular mechanics and machine learning force fields. *Applied Physics Reviews*, 12(2):021304, April 2025b. ISSN 1931-9401. doi: 10.1063/5.0237876. URL https://doi.org/10.1063/5.0237876. _eprint: https://pubs.aip.org/aip/apr/article-pdf/doi/10.1063/5.0237876/20468851/021304_1_5.0237876.pdf.

Brandon M. Wood, Misko Dzamba, Xiang Fu, Meng Gao, Muhammed Shuaibi, Luis Barroso-Luque, Kareem Abdelmaqsoud, Vahe Gharakhanyan, John R. Kitchin, Daniel S. Levine, Kyle Michel, Anuroop Sriram, Taco Cohen, Abhishek Das, Ammar Rizvi, Sushree Jagriti Sahoo, Zachary W. Ulissi, and C. Lawrence Zitnick. UMA: A Family of Universal Models for Atoms, 2025.

Minkai Xu, Alexander Powers, Ron Dror, Stefano Ermon, and Jure Leskovec. Geometric latent diffusion models for 3d molecule generation. In *ICML*, 2023a.

Minkai Xu, Alexander S. Powers, Ron O. Dror, Stefano Ermon, and Jure Leskovec. Geometric latent diffusion models for 3d molecule generation. In *ICML*, pp. 38592–38610, 2023b. URL https://proceedings.mlr.press/v202/xu23n.html.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *ICLR*, 2025. URL https://openreview.net/forum?id=DJSZGGZYVi.

Cheng Zeng, Jirui Jin, Connor Ambrose, George Karypis, Mark Transtrum, Ellad B. Tadmor, Richard G. Hennig, Adrian Roitberg, Stefano Martiniani, and Mingjie Liu. PropMolFlow: property-guided molecule generation with geometry-complete flow matching. *Nature Computational Science*, 6(3):233–242, March 2026. ISSN 2662-8457. doi: 10.1038/s43588-025-00946-y. URL https://www.nature.com/articles/s43588-025-00946-y.

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 *ICLR*, 2023.

APPENDIX

A Limitations and Broader Impacts **16**

B Implementation details **16**

C Ablations **17**

C.1 DrugFlow with Exponential Moving Average 17

C.2 Encoder type 18

C.3 Layer selection 19

C.4 Projector architecture 19

C.5 Strength of alignment 19

C.6 Training curves 20

D Additional results on the SPINDR dataset **21**

D.1 Strain decomposition 21

D.2 Chemical properties of the generated molecules 22

D.3 Conditional generation tasks 24

A LIMITATIONS AND BROADER IMPACTS

Limitations. Several limitations remain. First, our alignment loss is applied only to ligand atoms and ignores the pocket; incorporating pocket-side alignment is a natural extension.

Second, the strain gain comes with a small but consistent cost in docking score. We argue in Section 5.3 that this reflects a genuine physical trade-off, since the same tension appears when the native crystallographic ligands are relaxed, but it is a real cost nonetheless: a pipeline that ranks purely on docking score would not see our method as an improvement.

Third, all models are trained to a fixed epoch budget and we do not perform per-metric early stopping. The metrics converge at different rates — PoseBusters and strain improve throughout training while Vina plateaus early (Figure 6) — so the reported balance between them depends on where training stops and a different stopping criterion would possibly shift that balance to a more favorable one.

Finally, our protonation scheme needed for strain computation and UMA embedding calculations is based on a simple RDKit-based implementation and more elaborate techniques could be used in the future iterations of the method.

Broader Societal Impacts. This work accelerates the computational pipeline for structure-based drug design, potentially reducing the time and cost required to discover novel therapeutics. While generative chemistry models inherently carry dual-use risks (e.g., the theoretical generation of toxic compounds), our target-conditioned approach is specifically designed to fit complex biological pockets rather than blindly exploring general chemical space, inherently limiting its utility for malicious off-target generation.

B IMPLEMENTATION DETAILS

3D Molecule Denoisers. We instantiate the denoiser f_θ with two architectures. Both are flow-matching models for pocket-conditioned 3D ligand generation, producing continuous coordinates via conditional flow matching and discrete atom and bond types via Markov bridge models; they differ in how ligand atoms exchange information and in how the pocket is represented.

DrugFlow (Schneuing et al., 2025) uses a geometric heterogeneous graph neural network with separate weights for ligand and pocket nodes and SE(3)-equivariant GVP-based message passing (Jing et al., 2021), exchanging messages within a distance cutoff and representing the pocket as one node per residue. Of the 5 message-passing layers, we extract the intermediate per-atom embeddings $\mathbf{s}_i, \mathbf{V}_i$ (Eq. 1) from the alignment layer $L = d_3$, justified by the layer sweep in Appendix C.3.

FLOWR (Cremer et al., 2026) replaces local message passing with attention: equivariant self-attention lets every ligand atom attend to every other irrespective of distance. The pocket is encoded at all-atom resolution, keeping side chains explicit, and the ligand attends to it through cross-attention. We extract embeddings from layer $L = d8$.

Projectors. Since denoiser f_θ and encoder $\mathcal{E}$ operate in different latent spaces ($d \neq C$), the projectors (Eq. 3) map denoiser features into the encoder's representation space. The scalar projector h is a one-hidden-layer MLP (hidden dimension 512, SiLU); the vector projector g is a single GVP block with vector gating (Jing et al., 2021). Both are trained jointly with f_θ and discarded at inference. We report ablation of the projector architecture in Section C.4.

Frozen Physics Encoder (MLIP). We instantiate the frozen physics encoder $\mathcal{E}$ with UMA-S-1p1, the small variant of the Universal Models for Atoms (UMA) (Wood et al., 2025), a recent machine-learned interatomic potential pretrained on large-scale quantum-mechanical datasets covering a broad range of chemical environments and off-equilibrium configurations. Like DrugFlow, it is an equivariant message-passing network that operates on the graph of the pocket-ligand complex. We pass it the clean ligand $\mathbf{X}_0$ together with the pocket $\mathcal{P}$, so the targets encode the ligand in its binding context rather than in isolation. Because an MLIP models interatomic forces and is undefined for a molecule with missing hydrogens, the ligand is protonated before it enters the encoder; the targets are then read off the heavy atoms only, so that they correspond one-to-one with the denoiser's atoms. From its 4 message-passing layers we extract the per-atom targets $z_i, \mathbf{Z}_i$ (Eq. 2) from layer $K = 2$. The choice of UMA size and extraction layer is determined by the ablation in Appendix C.3. This encoder remains frozen throughout training.

Training. f_θ, h, and g are trained jointly from scratch to minimize the composite objective in Eq. 5 with $\lambda_{\text{scal}} = \lambda_{\text{vec}} = 1.0$. An Exponential Moving Average of the denoiser weights with decay 0.999 is maintained for inference (Ho et al., 2020). Otherwise we use the default training hyperparameters of DrugFlow. DrugFlow-based models are trained for 600 epochs.

Compute Resources. The training was performed on a single NVIDIA H200 GPU, with 32 CPU cores and 250 GB RAM. The DrugFlow backbone has 12.1 M trainable parameters; the REPA variant has a total of 12.6 M parameters. We trained with batch size 64 and two-step gradient accumulation.

C Ablations

For all ablations we train at small scale, running each DrugFlow variant for the first 100 epochs only, so that the comparisons are made under an identical and affordable budget. We ablate five design choices: the encoder type, to test whether the gains are specific to alignment with an MLIP or hold for other pretrained encoders (Section C.2); the UMA model size and the layer K from which targets are extracted (Section C.3); the projector architecture, comparing a linear map against an MLP on the scalar channel and a linear map against a GVP on the vector channel (Section C.4); the alignment strength λ (Section C.5).

C.1 DrugFlow with Exponential Moving Average

Throughout the paper we report results of the DrugFlow model using an exponential moving average (EMA) of the model parameters, maintained during training with decay $\beta = 0.999$. Tables 5 and 6 justify this choice: on both CrossDocked2020 and SPINDR, the EMA weights improve PoseBusters validity, strain energy and Vina score over the final checkpoint, at no additional training or inference cost. The same holds for the REPA variants, so all comparisons in the main text are made between EMA models.

Table 5: **Evaluation on CrossDocked2020.** DrugFlow variants are averaged over 3 seeds.

Model	Size	PB ↑	Strain ↓	Vina ↓
DrugFlow	22.05	0.577	117.3	-5.92
DrugFlow^{EMA}	21.86	0.732	77.7	**-6.05**
+ REPA	21.94	**0.789**	**71.4**	-5.85
Native ligands	22.75	0.950	71.3	-6.78

Table 6: **Evaluation on SPINDR.** Representation alignment added to DrugFlow.

Method	Size	PB ↑	Strain ↓	Vina ↓
DrugFlow	24.98	0.689	62.5	-6.13
DrugFlow^{EMA}	24.63	0.745	52.9	**-6.33**
+ REPA	24.94	**0.808**	**51.3**	-6.20
Native ligands	26.68	1.000	31.1	-7.54

C.2 Encoder type

This section asks how the choice of encoder to which DrugFlow is aligned affects the generator. We report the PoseBusters pass rate, and strain and Vina as per-target medians over PoseBusters-valid molecules, averaged across targets. The experiments below show that the gains hold for multile MLIP encoders and that the gains do not hold for other control encoders.

Ablation of MLIP. To determine whether the gains are specific to UMA, we repeated the experiment with two further state-of-the-art MLIPs based on the MACE architecture: MACE-OFF24 (Kovacs et al., 2025) (medium, 1.43M parameters), trained on SPICE (Eastman et al., 2023), and MACE-OMol (extra-large, 52.4M parameters), trained on OMol25, the dataset UMA also uses.

As shown in table Table 7 improvements over the baseline DrugFlow model are consistent regardless of which MLIP supplies the targets: all three variants raise the PoseBusters pass rate and lower strain. Which encoder is best depends on the metric — MACE-OMol gives the highest PoseBusters pass rate and the lowest median strain, UMA retains the best Vina score among the aligned models, and MACE-OFF yields the most high-quality molecules — but the spread between them is small next to the gap to the baselines. The improvements are therefore not limited to the UMA model.

Naive baselines. Two controls test whether the improvement could come from the auxiliary loss alone rather than from the MLIP teacher.

First, we ask whether it is enough for the generator to see any clean-ligand embedding during training, which would supply a stronger gradient and speed up learning irrespective of what the embedding represents. We replace the MLIP with an already trained DrugFlow model as the frozen encoder, so one DrugFlow provides the targets and a second is aligned to them. The PoseBusters and strain gains largely disappear (0.696 and 80.8 against 0.763 and 76.0 for UMA; Table 7).

The second asks whether the benefit could come from information common to all UMA embeddings, or again from the extra gradient. We keep the UMA targets but shuffle them within each ligand, so every atom is aligned to the embedding of a different atom of the same molecule. This does not recovers any improvement (0.693 and 80.5). The improvement therefore depends on each atom receiving its own physically meaningful target.

Table 7: **Effect of encoder choice on physical and binding metrics for CrossDocked.** Values after $\pm$ are the standard deviation *across targets*, i.e. the spread of the per-target values being averaged. Best per column in **bold**, second best underlined. The improvements in energetic plausibility remain robust regardless of the specific pretrained MLIP encoder used.

Model	Size	PoseBusters ↑	Strain ↓	Vina ↓	High Quality ↑
DrugFlow	21.63	0.509	123.9	-5.36	0.000
DrugFlow EMA	21.87	0.691	81.6	$\mathbf{-5.86}$	0.002
Control: Align to UMA shuffled	21.52	0.693	80.5	$\underline{-5.78}$	0.002
Control: Align to trained DrugFlow EMA	21.40	0.696	80.8	-5.73	0.002
Ours (UMA)	21.82	$\underline{0.763}$	76.0	-5.78	$\underline{0.005}$
Ours (MACE-OFF)	21.74	0.749	$\underline{75.9}$	-5.71	**0.006**
Ours (MACE-OMol)	21.67	**0.777**	**74.0**	-5.74	0.004

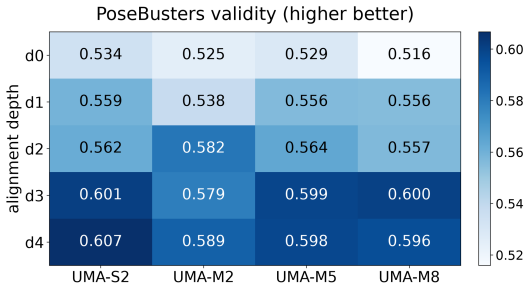

Figure 5: PoseBusters validity across the depth grid. Rows are the DrugFlow injection depth (d_0–d_4), columns the UMA encoder and extraction layer. Each cell is one model trained for 100 epochs and sampled on the validation set with 50 steps. Higher is better (darker). Injecting at d_0 is the weakest cell in every column.

C.3 Layer selection

In this section we ablate the two depth choices that define how UMA features enter the model: the extraction depth — which layer of the UMA encoder the embeddings are read from — and the injection depth — which layer of the DrugFlow denoiser they are aligned against.

Embeddings are obtained by a single forward pass of a frozen UMA model over the protein–ligand complex, taking the per-atom representation at the chosen encoder layer. UMA-S-1p1 has 4 layers, so we extract at layer 2 (e_2); the larger UMA-M has 10, for which we sweep three extraction points spanning early, middle and late representations, e_2, e_5 and e_8. These targets are aligned against the DrugFlow denoiser at one of its 5 layers, d_0 through d_4, giving a 5×4 grid of configurations. All other settings are held fixed: the alignment weight is $\lambda = 1$ on both the invariant and equivariant terms, and the projector is an MLP for scalars and a GVP for vectors.

We evaluate each configuration by PoseBusters validity (higher is better, measuring physical plausibility of generated poses). All models are trained for 100 epochs. Results are reported on the validation set with 50 sampling steps to enable faster iteration over the grid (Figure 5); 50 steps yields lower-quality generations than the 500-step schedule used at test time, so absolute values here are systematically worse than each configuration would achieve with full sampling, and we rely on the relative ranking rather than the absolute numbers.

The clearest takeaway is that injection depth dominates extraction depth. Injecting at d_0 is uniformly worse than deeper injection across all UMA backbones and extraction layers — it is the weakest cell in every column (0.516–0.534) while d_3 and d_4 reach 0.579–0.607. We attribute this to the limited receptive field of the message-passing GNN at the input layer: deeper injection points first allow the GNN to build up node representations that capture more of the surrounding structure, which UMA features can then modulate. By contrast, the choice of backbone and extraction layer matters far less: at fixed injection depth the four columns differ by at most 0.03. We selected UMA-S layer 2 with DrugFlow depth d_3, which is among the strongest configurations while using the smaller UMA-S backbone.

C.4 Projector architecture

To validate our equivariant design, we ablate the projector configurations Table 8. We compare (i) a linear projector on scalar representations (l_0), (ii) an MLP on l_0, and (iii) our full model (l_0 MLP + l_1 GVP). While the l_0 MLP improves over the linear baseline, our joint scalar-vector alignment strictly outperforms both. This confirms that aligning the $SO(3)$-equivariant vector representations is critical for capturing directional geometric physics.

C.5 Strength of alignment

We have performed additional ablations on the sensitivity of our model to the alignment weights, shown in Table 9. For simplicity, we set all alignment loss weights to the same value, i.e., $\lambda_{\text{REPA}} = \lambda_{\text{scal}} = \lambda_{\text{vec}}$. Under this setting, our method is robust across a broad range of values, with the best performance obtained at $\lambda_{\text{REPA}} = 1$. Importantly, tested values in range $\lambda_{\text{REPA}} \in [0.3, 3]$ all achieve

Table 8: REPA projector ablation. L0 aligns the rotation-invariant scalars only; L0+L1 additionally aligns the equivariant vector block. L0+L1 with MLP+GVP is the main configuration, identical to $\lambda = 1$. Best per column in **bold**, second best underlined.

Model	Size	PB ↑	Strain ↓	Vina ↓
DrugFlow EMA	21.87	0.691	81.6	**−5.86**
L0, linear	22.01	0.685	83.2	−5.82
L0, MLP	21.52	0.719	84.0	−5.81
L0+L1, MLP+GVP	21.82	**0.763**	**76.0**	−5.78

significantly better PoseBusters scores and strain than the DrugFlow EMA baseline, suggesting that our method is not overly sensitive to the precise choice of the alignment weight.

Table 9: REPA loss-weight ablation. CrossDocked2020 test set at 100 training epochs, 100 targets × 100 sampled ligands, seed 42, EMA weights. Strain and Vina are computed over PoseBusters-valid molecules only as the median within each target, then averaged over targets; size and PoseBusters are over all sampled molecules. Best per column in **bold**, second best underlined.

Model	Size	PoseBusters ↑	Strain ↓	Vina ↓
DrugFlow EMA	21.87	0.691	81.6	**−5.86**
$\lambda = 0.03$	21.43	0.732	80.9	−5.83
$\lambda = 0.1$	21.40	0.748	76.7	−5.80
$\lambda = 0.3$	21.56	0.757	79.1	−5.84
$\lambda = 1$	21.82	**0.763**	**76.0**	−5.78
$\lambda = 3$	21.62	0.750	77.6	−5.57
$\lambda = 10$	21.74	0.711	78.2	−5.48
$\lambda = 30$	21.67	0.705	80.7	−5.35

C.6 Training curves

To separate the effect of alignment from the effect of training, we track all four metrics over training for DrugFlow EMA with and without REPA, three seeds each, trained for 600 epochs.

Validity and the PoseBusters pass rate rise far more steeply for the aligned model (Ours), which reaches the baseline's final value within roughly the first hundred epochs and then holds a gap the baseline never closes. Strain drops sharply for both models within the first 50 epochs with the aligned model settling lower. Vina is the one metric where the aligned model is at a marginally worse value. We can observe the strain-versus-complementarity trade-off described in Section 5.1.

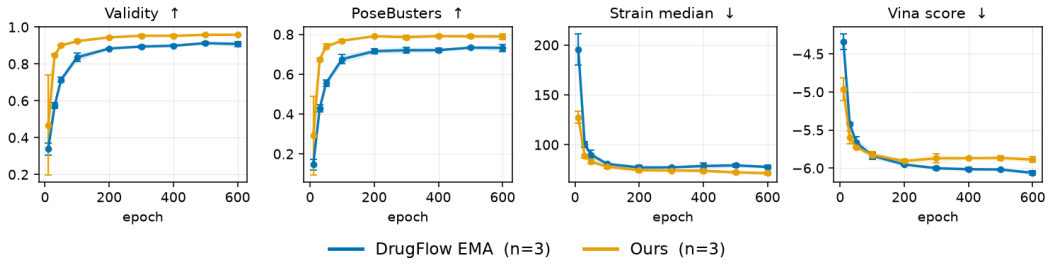

Figure 6: Training curves of DrugFlow EMA model with and witout representation alignment training. Model trained with representation alignment (Ours) trains much faster and achieves better performance in Validity and PoseBusters. Our strain is also better than the baseline at the cost of worse Vina score.

D ADDITIONAL RESULTS ON THE SPINDR DATASET

D.1 STRAIN DECOMPOSITION

In this work, we use the local strain energy as defined in the GenBench3D benchmark (Baillif et al., 2024). The local strain of a conformation $\mathbf{x}$ is the difference between its MMFF94s energy and the energy of the nearest local minimum, obtained by unconstrained minimization of $\mathbf{x}$ in vacuum (i.e. without the protein pocket):

$$E^{\text{local}}_{\text{strain}}(\mathbf{x}) = E_{\text{MMFF94s}}(\mathbf{x}) - E_{\text{MMFF94s}}\big(\min(\mathbf{x})\big), \tag{6}$$

where $\min(\mathbf{x})$ denotes the conformation reached from $\mathbf{x}$ after at most 1,000 iterations of MMFF94s energy minimization in RDKit, without restraints and without the protein. Since the minimizer is local and the number of iterations is capped, $\min(\mathbf{x})$ approximates the nearest local minimum rather than the global one. The quantity therefore measures the excess internal energy of the conformation relative to its own relaxed geometry. The generated molecules carry no explicit hydrogens; as in GenBench3D, hydrogens are added by RDKit at idealized positions before the energy is evaluated. For poses minimized in the pocket (Section 5.3), the hydrogens positioned by that minimization are kept rather than re-idealized. Strain is reported only for molecules that are connected and PoseBusters-valid in the pose state measured.

Decomposition by energy term. The MMFF94s energy is a sum of seven terms: bond stretching, angle bending, stretch–bend coupling, out-of-plane bending, torsion, van der Waals and electrostatic interactions, $E_{\text{MMFF94s}} = \sum_t E_t$. We attribute the strain to individual terms by evaluating each term at the pose and at its local minimum,

$$E^t_{\text{strain}}(\mathbf{x}) = E_t(\mathbf{x}) - E_t\big(\min(\mathbf{x})\big), \qquad \sum_t E^t_{\text{strain}}(\mathbf{x}) = E^{\text{local}}_{\text{strain}}(\mathbf{x}). \tag{7}$$

The contributions add up to the total strain. An individual contribution can be negative, because the minimization lowers the total energy and may raise one term while lowering others.

Stiff and soft strain. For the analysis we group the terms into two classes (Figure 7). We refer to bond stretching, angle bending, stretch–bend and out-of-plane terms as *stiff*: they depend on the positions of directly bonded neighbours, and each has a single ideal value. We refer to torsion, van der Waals and electrostatic terms as *soft*: the torsion term depends on rotation about a bond, and MMFF94s evaluates van der Waals and electrostatic interactions between atoms separated by three or more bonds, so all three depend on the overall conformation of the molecule. This grouping is our own convention, chosen to aid interpretation; it is not part of MMFF94s or of GenBench3D, and other groupings are possible. As shown below, the terms within a class do not always behave alike.

Working hypothesis. Our working hypothesis is that the two classes respond differently to the restrained in-pocket minimization of Section 5.3, in which the protein is held fixed, every ligand heavy atom may move at most 1 Å from its generated position, and hydrogens move freely. A distorted bond length or bond angle can be corrected by displacing a few atoms by a fraction of an ångström, well within this tolerance. Correcting a twisted torsion or an unfavourable electrostatic interaction may instead require rotating part of the molecule, which moves some atoms by more than 1 Å and can also be blocked by the surrounding pocket. We have not tested this mechanism directly; the observations below are consistent with it but do not establish it.

Observations. Table 10 reports the median strain per term before and after minimization, over all connected, PoseBusters-valid molecules. For the baselines, minimization lowers the median stiff strain per pocket from 14.8 to 2.6 kcal mol^{-1} for DrugFlow and from 39.3 to 3.3 for FLOWR (by 82% and 92%), and the soft strain from 35.8 to 17.9 and from 49.9 to 20.4 (by 50% and 59%). The soft class is not uniform. Van der Waals strain is largely removed (DrugFlow 2.7 to 0.7, FLOWR 13.1 to 1.1), as are the stiff terms, whereas torsional and electrostatic strain fall by only about 40–47%. After minimization, most of the remaining strain is therefore torsional and electrostatic. The native ligands of the same pockets, put through the identical minimization, show the same pattern (stiff 0.6, soft 10.0 kcal mol^{-1}).

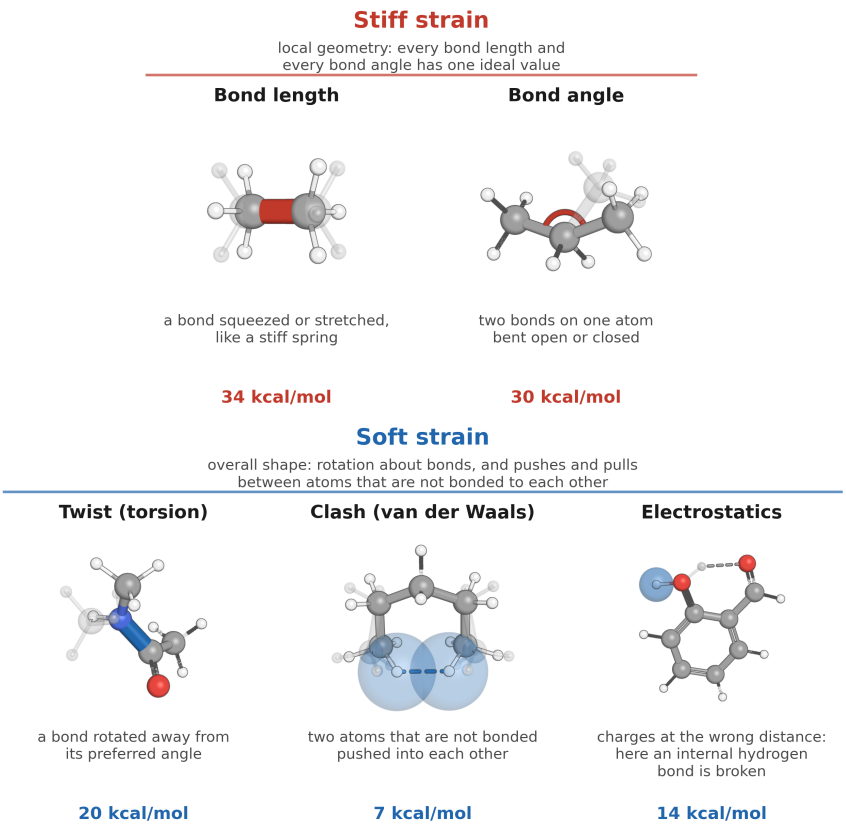

Figure 7: **Components of strain energy.** Each term is illustrated on a small molecule. Stiff terms (top) penalise deviations of bond lengths and bond angles; soft terms (bottom) penalise twisting about a bond, van der Waals clashes between non-bonded atoms, and unfavourable electrostatics, here a broken internal hydrogen bond. Coloured molecules are distorted, grey shows their relaxed shape; the values below the images are the MMFF94s energy of the shown term alone for the shape shown.

REPA raises the as-generated stiff strain in both models (DrugFlow 14.8 to 16.1, FLOWR 39.3 to 49.6 kcal mol^{-1}), but this difference does not persist after minimization: it is no longer significant for DrugFlow, and FLOWR with REPA is slightly lower (3.2 against 3.3). REPA lowers the as-generated soft strain in both models (35.8 to 33.0 and 49.9 to 48.5). After minimization, the soft strain remains significantly lower for FLOWR (19.8 against 20.4), with lower torsional and van der Waals strain. For DrugFlow the difference in total soft strain is not significant: the electrostatic strain remains lower with REPA (11.6 against 12.2), but the van der Waals strain is slightly higher (0.9 against 0.7).

D.2 Chemical properties of the generated molecules

Alignment to a physics encoder acts on geometry, so it is worth checking that it does not significantly degrade the properties on which a generated set is otherwise judged. We report uniqueness, diversity and novelty in Table 11 and synthetic accessibility in Table 12, in both cases per target over the PoseBusters-valid molecules and then averaged over targets.

Uniqueness and diversity fall slightly under REPA. The novelty against the training set is unchanged and sits at or near its ceiling, which confirms that the models are not collapsing toward memorized training molecules. They are, however, sampling a somewhat narrower region of novel chemistry. Given the magnitude, we regard this as a modest cost rather than a failure mode.

Synthetic accessibility remains close in both cases. The molecules REPA produces are therefore no harder to make than the baselines', which matters because a geometric prior could in principle push the model toward strained-but-relaxable ring systems or unusual substitution patterns that a synthesis route would struggle with.

Table 10: **Strain by MMFF94s energy term, SPINDR *de novo*.** Each cell reads *as generated* → *after minimisation*: the median strain of that term per pocket, averaged over the 222 pockets (kcal mol^{-1}, lower is better), over molecules that are connected and PoseBusters-valid in the pose state measured. *All* is the whole stiff (bond, angle, stretch–bend, out-of-plane) or soft (torsion, van der Waals, electrostatic) strain; medians are taken per column, so columns do not add up. Stretch–bend and out-of-plane have no column of their own (every median within ± 0.5). A term can be negative because the free minimisation lowers the total energy, not every term. As-generated poses carry RDKit-idealised hydrogens (the GenBench3D convention); minimised poses keep the hydrogens placed by the minimisation. The native ligands are one molecule per pocket, so their row is the median over pockets. Within each backbone, **bold** marks the value significantly lower between the baseline and its REPA variant (paired Wilcoxon signed-rank over pockets, seeds pooled, $p < 0.05$).

	Stiff strain ↓			Soft strain ↓			
	all	Bond	Angle	*all*	Torsion	vdW	Elec.
	as generated → *after minimisation*						
DrugFlow	**14.8** → 2.6	**5.8** → **−0.2**	**8.2** → 2.4	35.8 → 17.9	6.0 → 3.5	2.7 → **0.7**	23.0 → 12.2
+ REPA	16.1 → 2.8	6.4 → −0.1	9.1 → 2.5	**33.0** → 17.2	6.2 → 3.6	2.5 → 0.9	**21.4** → **11.6**
FLOWR	**39.3** → 3.3	**21.2** → −0.2	**17.3** → 3.0	49.9 → 20.4	7.2 → 4.2	13.1 → 1.1	24.9 → 13.7
+ REPA	49.6 → **3.2**	28.6 → −0.2	19.9 → **2.9**	**48.5** → **19.8**	**5.9** → **3.8**	13.0 → **1.0**	25.4 → 13.8
Native ligands	9.3 → 0.6	4.3 → −0.1	5.3 → 0.8	19.5 → 10.0	2.7 → 1.5	−0.9 → 0.5	10.8 → 6.5

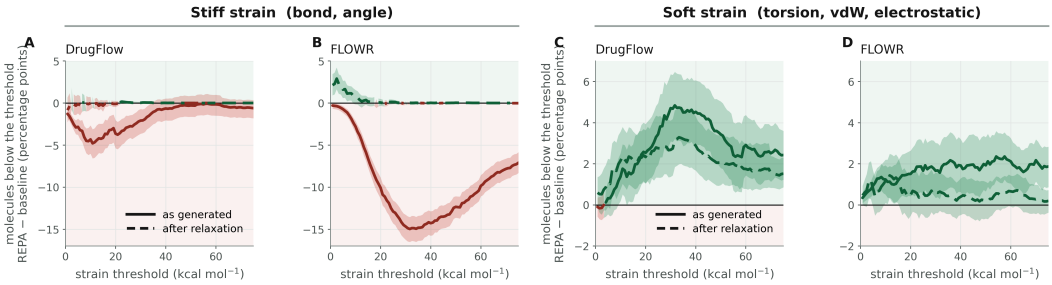

Figure 8: **Change in the share of molecules below a strain threshold, REPA minus baseline.** Each panel sweeps a threshold on stiff strain (A: DrugFlow, B: FLOWR) or soft strain (C: DrugFlow, D: FLOWR) and plots the difference, REPA minus baseline, in the share of molecules below it, as generated and after minimization; shaded bands are 95% confidence intervals. Values above zero mean that REPA has more molecules below the threshold. As generated, REPA has fewer molecules with low stiff strain (A, B), and minimization largely removes this difference. The soft strain is lower with REPA as generated (C, D), and part of this advantage remains after minimization.

Table 11: **Uniqueness, diversity and novelty.** 2D, over the PoseBusters-valid molecules of each target. REPA costs a small amount of uniqueness and diversity in both architectures; novelty against the training set is unchanged and essentially saturated.

Method	Uniqueness ↑	Diversity ↑	Novelty ↑
DrugFlow, 600 ep	**0.994 ± 0.002**	**0.885 ± 0.000**	1.000 ± 0.000
+ REPA (Ours)	0.987 ± 0.001	0.867 ± 0.001	1.000 ± 0.000
FLOWR	**0.971 ± 0.003**	**0.861 ± 0.002**	**0.988 ± 0.002**
+ REPA (Ours)	0.961 ± 0.000	0.854 ± 0.001	0.986 ± 0.000

Table 12: **Synthetic accessibility (Ertl & Schuffenhauer, 2009).** Per-target mean over PoseBusters-valid molecules. Lower is more accessible. REPA does not make molecules harder to make.

Method	SA ↓
DrugFlow, 600 ep	4.052 ± 0.052
+ REPA (Ours)	**4.028 ± 0.013**
FLOWR	**3.968 ± 0.027**
+ REPA (Ours)	3.993 ± 0.024

D.3 CONDITIONAL GENERATION TASKS

Practical structure-based design is rarely unconstrained. Parts of the molecule are typically fixed, such as a scaffold retained across a series, a set of interactions that must be preserved, or a functional group or linker to be filled in, and the model completes the rest. FLOWR supports these inpainting modes, therefore we can test whether alignment benefits only unconstrained sampling (*de novo*) or transfers to the conditional settings in which these models are used.

In Table 13 we report the ratio of molecules passing PoseBuster checks, interaction recovery and fraction of high quality molecules produced by our model. We show FLOWR *de novo* scores for reference. The results show that adding REPA improves the interaction recovery on four modes with statistical significance. High-quality yield improves in three of the four tasks them, with scaffold inpainting the exception where it is unchanged. PoseBusters is essentially saturated. MCS is omitted for these tasks, since part of the native molecule is given as input and the metric would measure the input rather than the model.

The same pattern as fror PLIF recovery holds under PLIF Tanimoto, which is the stricter of the two interaction measures because it penalises interactions the generated molecule makes that the crystallographic ligand does not. REPA improves it significantly in every mode and on both backbones *de novo* (Table 14). Recovering the right interactions without adding spurious ones is the property that matters when a proposed binder is judged against a known one, and alignment improves it across every generation task we tested.

Table 13: **SPINDR conditional (inpainting) generation, FLOWR.** Interactions, scaffold, functional groups or linkers held fixed. MCS is omitted because part of the native molecule is given as input. REPA improves interaction recovery in all four modes.

Method	PoseBusters ↑	PLIF rec. ↑	High quality ↑
de novo			
FLOWR	0.890 ± 0.002	0.452 ± 0.007	0.279 ± 0.003
+ REPA (Ours)	$\mathbf{0.896 \pm 0.003}^{***}$	$\mathbf{0.475 \pm 0.001}^{***}$	$\mathbf{0.292 \pm 0.004}^{***}$
interaction			
FLOWR	0.884 ± 0.003	0.766 ± 0.007	0.287 ± 0.003
+ REPA (Ours)	$\mathbf{0.892 \pm 0.004}^{***}$	$\mathbf{0.784 \pm 0.002}^{***}$	$\mathbf{0.307 \pm 0.005}^{***}$
scaffold			
FLOWR	$\mathbf{0.893 \pm 0.006}$	0.654 ± 0.002	$\mathbf{0.402 \pm 0.004}$
+ REPA (Ours)	0.890 ± 0.010	$\mathbf{0.671 \pm 0.007}^{***}$	0.400 ± 0.005
functional group			
FLOWR	0.876 ± 0.006	0.800 ± 0.006	0.286 ± 0.002
+ REPA (Ours)	$\mathbf{0.878 \pm 0.007}$	$\mathbf{0.824 \pm 0.006}^{***}$	$\mathbf{0.306 \pm 0.010}^{***}$
linker			
FLOWR	$\mathbf{0.874 \pm 0.009}$	0.803 ± 0.005	0.342 ± 0.001
+ REPA (Ours)	0.871 ± 0.009	$\mathbf{0.811 \pm 0.005}^{**}$	$\mathbf{0.354 \pm 0.007}^{**}$

Table 14: **PLIF Tanimoto similarity.** Recovery rate counts the crystallographic interactions that are reproduced; Tanimoto also penalises interactions the generated molecule makes and the crystal ligand does not. REPA improves both, in every generation mode.

Model	Mode	Baseline	+ REPA (Ours)
DrugFlow	*de novo*	0.362 ± 0.008	$\mathbf{0.373 \pm 0.005}^{*}$
FLOWR	*de novo*	0.399 ± 0.006	$\mathbf{0.417 \pm 0.002}^{***}$
FLOWR	interaction	0.688 ± 0.006	$\mathbf{0.703 \pm 0.003}^{***}$
FLOWR	scaffold	0.620 ± 0.002	$\mathbf{0.634 \pm 0.006}^{**}$
FLOWR	functional group	0.744 ± 0.004	$\mathbf{0.772 \pm 0.007}^{***}$
FLOWR	linker	0.771 ± 0.005	$\mathbf{0.778 \pm 0.006}^{*}$