Boltz is a Strong Baseline for Atom-level Representation Learning

Anonymous Authors[1]

Abstract

Foundation models in molecular learning have advanced along two parallel tracks: protein models, which typically utilize evolutionary information to learn amino acid-level representations for folding, and small-molecule models, which focus on learning atom-level representations for property prediction tasks such as ADMET. Notably, cutting-edge protein-centric models such as Boltz now operate at atom-level granularity for protein-ligand co-folding, yet their atom-level expressiveness for small-molecule tasks remains unexplored. A key open question is whether these protein co-folding models capture transferable chemical physics or rely on protein evolutionary signals, which would limit their utility for small-molecule tasks. In this work, we investigate the quality of Boltz atom-level representations across diverse small-molecule benchmarks. Our results show that Boltz is competitive with specialized baselines on ADMET property prediction tasks and effective for molecular generation and optimization. These findings suggest that the representational capacity of cutting-edge protein-centric models has been underexplored and position Boltz as a strong baseline for atom-level representation learning for small molecules.

1. Introduction

Large-scale representation learning has shown strong transferability across a wide range of downstream tasks, particularly in language (Grattafiori et al., 2024; Achiam et al., 2023) and vision domains (Dosovitskiy et al., 2021; Oquab et al., 2024). Foundation models with large-scale pretraining have become essential for achieving state-of-the-art performance in downstream applications, serving as powerful initializations (Devlin et al., 2019) or feature extractors (Radford et al., 2021) for predictive tasks. Beyond predictive tasks, their representations have also been leveraged to enhance generative modeling, including distillation into denoising models (Yu et al., 2025) and use as latent embeddings for latent diffusion models (Zheng et al., 2025).

Similarly, researchers in the molecular domain have also explored foundation models that learn transferable representations across diverse downstream tasks. These approaches can be categorized by the granularity of representations they model. For small molecules, models such as MiniMol (Klaser et al., 2024) and MolGPS (Sypetkowski et al., 2024) operate at the atom level and primarily target molecular property predictions, e.g., ADMET (Huang et al., 2021). In parallel, protein-centric models, including AlphaFold (Jumper et al., 2021) and the ESM family (Lin et al., 2022; Hayes et al., 2025), learn amino-acid-level representations leveraged for a wide range of protein-based applications (Cheng et al., 2023; Ouyang-Zhang et al., 2023).

Notably, cutting-edge protein-centric models have reached atom-level granularity. AlphaFold3 (Abramson et al., 2024) and its open-weight counterpart Boltz (Wohlwend et al., 2024; Passaro et al., 2025) learn atom-level representations to predict protein-ligand complexes with full atomic detail. This raises a natural question:

> *Can cutting-edge protein-centric models, particularly Boltz, provide strong atom-level representations for standalone small-molecule tasks?*

If this holds, it suggests that separate pretraining pipelines for small molecules may not be strictly necessary, pointing toward a more unified molecular foundation model.

On the one hand, the co-folding training objective used in protein-centric models such as Boltz seems promising for learning strong atom-level representations. Boltz is trained on collections of protein-ligand complexes and learns atomistic representations of ligands using their bound conformations, which may encode atom-level mechanisms such as hydrogen bonding and electrostatic interactions. While prior work shows that pretraining with 3D molecular structures yields transferable representations (Zaidi et al., 2023; Zhou et al., 2023), Boltz leverages richer supervision from interaction context in protein-ligand co-structures.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

On the other hand, skepticism remains regarding Boltz's broader utility. Some studies argue that protein models like Boltz rely heavily on evolutionary patterns derived from MSAs rather than a genuine atom-level understanding (Stahl et al., 2023). In addition, Masters et al. (2025) has shown that predictions of AlphaFold3 depend on memorization rather than fully understanding the atom-level mechanism.

Contribution. In this work, we empirically investigate the potential of Boltz as a source of representations for small-molecule tasks, *bridging cutting-edge protein-centric models and atom-level representation learning for small molecules*. Additionally, for the first time, we show that foundation model representations can be leveraged to guide and accelerate molecular generation and optimization. As illustrated in Figure 1, we extract atom-level representations from Boltz for a given small molecule and evaluate their expressiveness across a broad range of small-molecule downstream tasks via probing and distillation.

We show that Boltz provides strong atom-level representations out of the box, beyond its original protein–ligand co-folding objective. Through systematic experiments, we identify three key findings:

- **Competitive supervised learning (Section 3)**: On the extensive TDC ADMET benchmark, Boltz representations match or outperform representations from specialized small-molecule representations like the ones from MiniMol and MolGPS.
- **Accelerated generative modeling (Section 4):** Leveraging the representation alignment (Yu et al., 2025), we show that distilling Boltz representations into molecular generative models improves molecular generation quality.
- **Structure-guided ligand discovery (Section 5):** We show that Boltz representations enhance optimization in structure-based ligand design, providing a stronger learning signal that improves sample efficiency.

These findings suggest that the representational capacity of cutting-edge protein models has been underexplored by the community, and position Boltz as a strong baseline for small molecular foundation models. We also note the complementarity of Boltz representations with existing foundation models, and the improvement in downstream performance from the use of protein context (Section 6).

Beyond empirical evaluations of Boltz representations, we also provide three actionable insights for the molecular machine learning community:

- **Co-folding as pretraining.** We show that predicting 3D protein-ligand interactions provides effective supervision for atom-level molecular representation learning. This also complements existing property-specific pretraining pipelines by yielding distinct representation spaces.

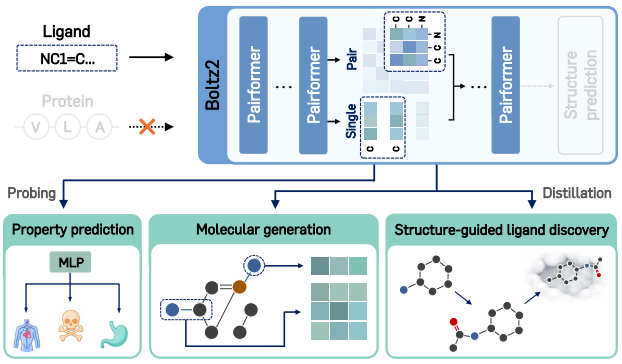


Figure 1. **Boltz as atom-level small molecular foundation models.** We repurpose Boltz, originally trained for protein-ligand co-folding, as a small-molecule representation model by leveraging atom-level ligand representations.

- **Representation-guided molecular optimization.** We extend representation alignment to online reinforcement learning for molecular discovery, showing that dense representation-level auxiliary supervision improves sample efficiency beyond scalar rewards alone.
- **Complementary representation ensembling.** We show that combining representations with low alignment, e.g., Boltz2 ensembled with MolE (Méndez-Lucio et al., 2024), outperforms higher alignment combinations, e.g., Boltz2 ensembled with MiniMol (Klaser et al., 2024), providing a useful strategy for representation fusion.

2. Background

2.1. Atom-level Representation Learning for Small-molecule Domain

Atom-level representation learning has been widely explored to support diverse molecular property prediction tasks, such as ADMET profiles, under minimal task-specific supervision. These approaches adopt large-scale atom-level pretraining to learn transferable representations that capture atomic interactions within molecules. Building on this paradigm, existing methods investigate a range of pretraining objectives, including masked atom or token prediction (Méndez-Lucio et al., 2024), supervision from large collections of bio-assay labels (Klaser et al., 2024), quantum-chemical property regression (Kim et al., 2024), and three-dimensional structure prediction (Zhou et al., 2023). Other approaches further incorporate structured chemical knowledge or scalability, e.g., KPGT (Li et al., 2023) and MolGPS (Sypetkowski et al., 2024).

2.2. Atom-level Representations in Cutting-edge Protein Co-folding Models

Cutting-edge protein co-folding models operate at atomistic granularity and explicitly incorporate ligands, enabling the

prediction of protein-ligand complex structures. Representative models such as Boltz (Wohlwend et al., 2024; Passaro et al., 2025) and AlphaFold3 (Abramson et al., 2024) are trained to predict 3D structures of protein-ligand complexes at atomic resolution for both proteins and ligands.

During co-folding structure prediction, these models learn joint representations of ligand atoms and protein residues under supervision from protein-ligand complex structures. Here, ligand atom representations are trained from their 3D conformations relative to surrounding protein atoms, implicitly encoding geometric and chemical traits of the bound state. Protein residues are also represented through aggregated atom-level features.

By learning representations at atomistic resolution within protein-ligand complexes, co-folding models naturally bridge protein-centric and small-molecule–centric representation learning. Although their training objectives are defined by protein-ligand structure prediction, the ligand representations capture atom-level interaction patterns, such as hydrogen bonding and electrostatic interactions, that have the potential to support standalone small-molecule tasks.

Boltz2 representation. We further explain Boltz2 (Passaro et al., 2025), which we primarily evaluate in this paper as an atom-level foundation model for small molecules. The core component is a 64-layer Pairformer trunk that produces pair and single representations over protein residues and ligand atoms for predicting 3D protein-ligand complex structures. Pair representations with 128 dimensions encode residue-residue, residue-atom, and atom-atom interactions, while single representations with 384 dimensions capture per-entity features updated via the pair representations.

We study the use of pretrained Boltz representations across a range of downstream small-molecule tasks. For standalone small-molecule tasks (Sections 3 and 4), we omit protein inputs and consider ligands only to obtain isolated atom-level representations of small molecules and evaluate their transferability. For structure-guided ligand discovery (Section 5), we retain protein inputs and use Boltz2 in its native protein-ligand setting, following SynFlowNet-Boltz pipeline (Passaro et al., 2025).

3. ADMET Property Prediction

To assess Boltz2 as an atom-level representation model for small molecules, we first consider ADMET property prediction, a standard benchmark in the literature (Li et al., 2023; Klaser et al., 2024; Méndez-Lucio et al., 2024; Kim et al., 2024; Sypetkowski et al., 2024). This benchmark covers absorption, distribution, metabolism, excretion, and toxicity prediction tasks, which depend on both global molecular structure and local atomic interactions.

3.1. Experimental Setup

Datasets. We conduct experiments on the Therapeutics Data Commons (TDC) ADMET benchmark datasets (Huang et al., 2021), which consist of 22 ADMET property prediction tasks for small molecules represented as SMILES strings (Weininger, 1988). We follow the standard benchmark protocol, including data splits, evaluation metrics, and five random seeds. We provide detailed dataset statistics and benchmark settings in Appendix A.1.

Implementation details. We apply probing to Boltz2 molecular representations for property prediction. Atom-wise pair representations are extracted from the Boltz2 Pairformer trunk, while single representations are discarded. We provide SMILES strings of molecules to the ligand modality of Boltz2, and protein sequence inputs are omitted.

To be specific, we concatenate 128-dimensional pair representations from the $\{16, 32, 48, 64\}$-th layers of the 64-layer Pairformer trunk, yielding a 512-dimensional representation.[1] We then concatenate pooled representations over diagonal entries, bonded atom-pair entries, and all entries in the pair representation. The resulting pooled representation is fed into a probing network for target label prediction. The probing network is implemented as a multi-layer perceptron, following prior works (Klaser et al., 2024; Sypetkowski et al., 2024). We provide implementation details and hyperparameter search ranges in Appendix A.2.

Baselines. We consider two deep learning-based methods, DeepAutoQSAR (Dixon et al., 2016) and DeepPurpose (Huang et al., 2020). Next, we consider six foundation models specialized for ADMET property prediction:

- **KPGT** (Li et al., 2023): pretrained with knowledge-guided objectives that enforce consistency with chemical structures, functional groups, and expert-defined rules.
- **UniQSAR** (Gao et al., 2023): ensembles various pretrained model with UniMol (Zhou et al., 2023) which is pretrained to predict 3D molecular structures.
- **MolE** (Méndez-Lucio et al., 2024): pretrained via masked language modeling on atom tokens with auxiliary losses to predict molecular properties and fingerprints.
- **MiniMol** (Klaser et al., 2024): pretrained to predict various labels spanning quantum chemistry, biological assays, and transcriptomic responses.
- **QIP** (Kim et al., 2024): pretrained to approximate electronic structure properties such as orbital interactions.
- **MolGPS** (Sypetkowski et al., 2024): ensembles three 1B-scale pretrained models incorporating additional modality labels from phenomic imaging.

[1] We evaluate the contribution of representations from each layer through ablation studies in Table 4.

Table 1. **Results on ADMET property prediction benchmark.** †Other denotes the highest score reported on the TDC ADMET benchmark leaderboards as of Jan 2026, excluding the baselines in the table. DP, DQ, and Mini. denote DeepPurpose, DeepAutoQSAR, and MiniMol. For the metabolism and excretion datasets, V., S., H., and M. denote Veith, Substrate, Hepatocyte, and Microsome, respectively. The results are averaged over five random seeds. **Bold** numbers indicate the best performance over all methods, while Underlined numbers indicate the best performance without ensemble. Overall, Boltz2 shows competitive performance compared to baselines. *This performance can be further improved by *incorporating relevant protein context* as input to Boltz2 (Table 6).

				w/ Pretraining					w/ Representation Ensembling		
Dataset	†**Other**	**DP.**	**DQ.**	**MolE**	**KPGT**	**Mini.**	**QIP**	**Boltz2**	**MolGPS**	**UniQSAR**	**Boltz2$^{\text{Mini.}}$**
Absorption											
Caco2 ↓	**0.26**	0.39	0.30	0.31	0.28	0.35	0.27	0.30	0.29	0.27	0.30
HIA ↑	**0.99**	0.97	0.98	0.96	0.98	**0.99**	**0.99**	**0.99**	0.98	**0.99**	**0.99**
Pgp ↑	0.94	0.92	0.92	0.92	0.94	0.94	0.93	0.93	**0.95**	0.93	0.93
Bioavailability ↑	0.75	0.67	0.68	0.65	0.75	0.69	0.73	0.75	0.70	0.73	**0.77**
Lipophilicity ↓	0.47	0.57	0.48	0.47	0.45	0.46	0.44	0.45	**0.39**	0.42	0.41
Solubility ↓	0.76	0.83	0.78	0.79	0.71	0.74	0.70	0.66	0.68	0.68	**0.64**
Distribution											
BBB ↑	0.92	0.89	0.88	0.90	0.91	0.92	0.90	0.93	**0.94**	0.93	**0.94**
PPBR ↓	7.44	9.99	8.04	8.07	7.68	7.70	7.36	7.65	**6.46**	7.53	7.59
VDss ↑	0.71	0.56	0.67	0.65	0.63	0.54	0.61	0.74	0.65	0.73	**0.75**
Metabolism											
CYP2C9 V. ↑	**0.86**	0.74	0.79	0.80	0.80	0.82	0.78	0.82*	0.84	0.80	**0.86**
CYP2D6 V. ↑	**0.79**	0.61	0.70	0.68	0.72	0.72	0.66	0.69*	0.75	0.74	0.72
CYP3A4 V. ↑	**0.92**	0.83	0.88	0.87	0.89	0.88	0.87	0.85*	0.90	0.89	0.88
CYP2C9 S. ↑	0.44	0.38	0.40	0.45	0.45	0.48	**0.52**	0.36	0.46	0.45	0.36
CYP2D6 S. ↑	**0.74**	0.68	0.70	0.70	**0.74**	0.73	0.67	0.52	0.71	0.72	0.51
CYP3A4 S. ↑	0.67	0.64	0.64	0.67	**0.73**	0.64	0.62	0.62	0.68	0.65	0.60
Excretion											
Half Life ↑	0.58	0.18	0.55	0.55	0.53	0.50	0.53	0.62	0.63	0.61	**0.65**
Clearance H. ↑	0.54	0.38	0.43	0.38	0.42	0.45	0.50	**0.62**	0.57	0.49	0.60
Clearance M. ↑	0.63	0.59	0.59	0.61	0.64	0.63	**0.66**	0.61	0.63	0.65	0.65
Toxicity											
LD50 ↓	0.55	0.67	0.59	0.82	0.55	0.59	0.56	**0.40**	0.56	0.55	**0.40**
hERG ↑	**0.88**	0.84	0.84	0.81	0.85	0.85	0.82	0.86	0.86	0.86	0.86
AMES ↑	0.87	0.82	0.86	0.88	0.87	0.85	0.86	**0.91**	0.86	0.88	**0.91**
DILI ↑	0.93	0.88	0.93	0.58	0.93	**0.96**	0.89	0.89	0.94	0.94	0.87
# 1st (# 1st w/o Ensembling)		0 (0)	0 (0)	0 (0)	2 (4)	2 (3)	3 (5)	4 (9)	4	1	9
# 2nd (# 2nd w/o Ensembling)		0 (0)	1 (2)	1 (2)	4 (7)	3 (5)	2 (4)	5 (4)	8	8	4

Among these baselines, MolGPS and UniQSAR employ ensemble-based approaches that combine representations from multiple pretrained models. For a direct comparison with such methods, we consider combining Boltz2 representations with those from MiniMol (Boltz2$^{\text{Mini.}}$).

3.2. Results

As reported in Table 1, one can see that Boltz2 representations overall exhibit competitive performance across the ADMET benchmark. Among non-ensemble methods, Boltz2 achieves the best performance on nine of the 22 tasks. When evaluated under ensemble settings, Boltz2$^{\text{Mini.}}$ likewise attains state-of-the-art results on nine tasks. We further summarize win, draw, loss statistics in Figure 2. Here, Boltz2 outperforms existing models on the majority of tasks, except for metabolism property prediction.

These results suggest that, despite being trained for protein-ligand co-folding prediction, Boltz2 representations transfer effectively to standalone small-molecule property prediction and perform competitively with existing models. In addition, the performance gains with representation ensembling of Boltz2$^{\text{Mini.}}$ indicate that Boltz2 provides complementary representations to existing molecular foundation models. We further analyze this in Figure 6 of Section 6.

Intriguingly, despite its strong overall performance, Boltz2 underperforms on metabolism-related tasks, particularly on substrate prediction. In alignment with the discussion by Sypetkowski et al. (2024), we hypothesize that this gap stems from minor data leakage in the bioassay pretraining datasets of the baselines. We also find that the Boltz2 performance can be further improved by incorporating relevant protein context, as shown in Table 6 of Section 6.

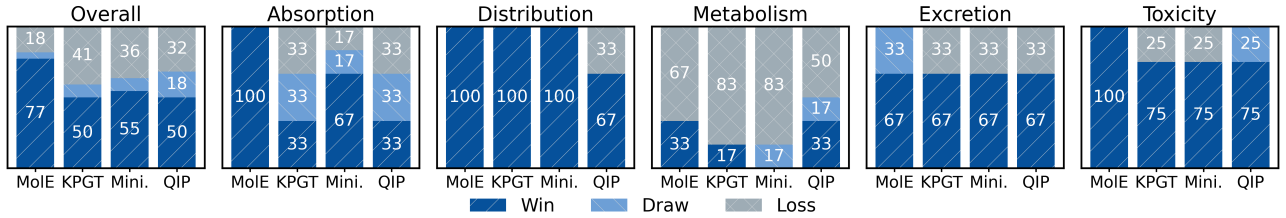


Figure 2. **Boltz2 vs. existing foundation models on ADMET benchmarks.** As illustrated, Boltz2 shows competitive performance compared to existing foundation models specialized for small molecules on four out of five domains.

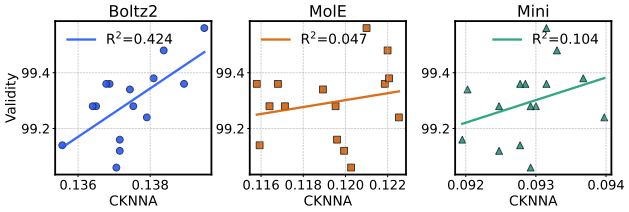


Figure 3. **Representation alignment with foundation models vs. generation quality.** Stronger alignment with Boltz2 representations correlates with higher molecular generation quality.

4. Molecular Generation

We next evaluate the quality of Boltz2 representations on small-molecule generation tasks. This experiment is motivated by recent work showing that representations from high-quality foundation models can improve the training of generative models via distillation (Yu et al., 2025). Following this approach, we distill Boltz2 representations into state-of-the-art molecular generative models and assess the representation quality via generation performance.

4.1. Observational Experiment

We first show that well-trained molecular generative models learn molecular representations that closely align with those produced by Boltz2. This alignment supports the use of Boltz2 representations as an additional training signal when training generative models from scratch (Yu et al., 2025).

Following prior work (Yu et al., 2025), we measure the representation alignment between Boltz2 and a state-of-the-art molecular generation model, GruM (Jo et al., 2024). Specifically, we adopt CKNNA (Huh et al., 2024), which quantifies the alignment between representations from molecular foundation models and generative models. Details of CKNNA are provided in Appendix B.1.

We report alignment scores and generation quality of generative models in Figure 3. One can see that stronger alignment between Boltz2 and the generative models correlates with improved molecular generation quality. Building on this observation, we apply distillation to explicitly enforce representation alignment with Boltz2 during the training of molecular generative models.

4.2. Experimental Setup

Datasets. We conduct molecular generation experiments on the widely used ZINC250k dataset (Jo et al., 2022; Vignac et al., 2023; Jang et al., 2024b; Kong et al., 2023; Jo et al., 2024; Seo et al., 2025). We train generative models on this dataset and evaluate their performance on 10,000 generated molecules following prior work.

Implementation details. We adopt a representation alignment-based distillation (Yu et al., 2025) to existing diffusion-based molecular generative models, specifically GruM (Jo et al., 2024). We parameterize the denoising model of GruM with a four-layer Pairformer that outputs single and pair representations as hidden representations for denoising atom and bond types, respectively.

We distill Boltz2 representations into GruM via representation alignment by training the denoising model to maximize the cosine similarity between its hidden representations of noisy molecules and the corresponding Boltz2 representations of clean molecules. We apply the alignment loss at the middle layer of the denoising model, aligning its single and pair representations with Boltz single and pair representations extracted as in Section 3. Additional implementation details are provided in Appendix B.2.

Baselines. We consider diffusion-based or autoregressive molecular generative models, including GDSS (Jo et al., 2022), DiGress (Vignac et al., 2023), GraphARM (Kong et al., 2023), HGGT (Jang et al., 2023), and the original GruM (Jo et al., 2024). In addition, we apply representation alignment to GruM using molecular representations from MolE, KPGT, MiniMol, and QIP to compare Boltz2-based alignment with alternative representation sources.

Metrics. Following prior works (Jo et al., 2024; Jang et al., 2023), we evaluate 10,000 generated molecules using eight metrics: chemical validity, Fréchet ChemNet Distance (FCD) (Preuer et al., 2018), the neighborhood subgraph pairwise distance kernel (NSPDK), novelty with respect to the training molecules, molecular uniqueness (Unique), and structural similarity metrics, including scaffold similarity (Scaffold), fragment similarity (Fragment), and similarity to the nearest neighbor (SNN).

Table 2. **Results on unconditional molecular generation.** GruM* denotes GruM parameterized with a Pairformer. The results are averaged over three random seeds. **Bold** numbers indicate the best performance. Applying representation alignment with Boltz2 enhances the training of molecular generative models to produce higher-quality molecules compared to the considered baselines.

Method	Validity ↑	FCD ↓	NSPDK ↓	Novelty ↑	Unique ↑	Scaffold ↑	Fragment ↑	SNN ↑
GDSS	97.01	14.65	0.0190	99.98	99.97	0.0467	0.8138	0.2789
Digress	93.99	3.48	0.0021	99.99	99.97	0.4163	0.9679	0.3457
GraphARM	88.23	16.26	0.0550	**100.00**	99.46	–	–	–
HGGT	92.87	2.26	0.0010	99.97	99.83	0.5298	0.9877	**0.4383**
GruM	98.65	2.26	0.0015	99.98	99.97	**0.5299**	0.9777	0.3650
GruM*	99.41	1.49	0.0007	99.99	**100.00**	0.4923	0.9852	0.3697
*Applying representation alignment-based distillation to GruM**								
Distill. w/ MolE	99.51	1.41	0.0006	**100.00**	**100.00**	0.4932	0.9864	0.3737
Distill. w/ KPGT	99.37	1.46	0.0006	99.99	**100.00**	0.4800	0.9860	0.3702
Distill. w/ Mini.	99.48	1.44	0.0007	99.99	**100.00**	0.4812	0.9856	0.3710
Distill. w/ QIP	99.37	1.43	0.0007	**100.00**	**100.00**	0.5242	0.9864	0.3718
Distill. w/ Boltz2	**99.65**	**1.31**	**0.0005**	**100.00**	**100.00**	0.5064	**0.9881**	0.3706

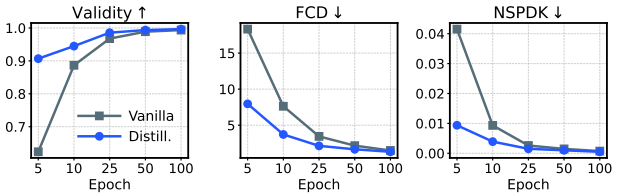


Figure 4. **Training acceleration using Boltz2.** Representation alignment with Boltz2 accelerates training of generative models.

Table 3. **Distillation into a large-scale model.** Consistent with Table 2, Boltz2 representation improves molecular generation performance on a large-scale backbone.

Method	Validity ↑	FCD ↓	NSPDK ↓
Pairformer$^{\text{large}}$	99.56	1.35	0.0006
Distill. w/ Boltz2	**99.76**	**1.28**	**0.0005**

4.3. Results

We report the performance in Figure 4 and Table 2. The results show that representation alignment with Boltz2 accelerates training and yields the largest gains across most evaluation metrics. In contrast, alignment with existing foundation models, which are primarily pretrained for property prediction, yields marginal improvements.

These results indicate that molecular representations vary in their effectiveness as supervision for generative modeling. Although Boltz2 is trained for protein-ligand co-folding, its atom-level representations provide stronger structural supervision for molecular generation than those of property-centric molecular foundation models.

Next, we further evaluate representation alignment-based distillation on a larger generative model with an eight-layer Pairformer architecture. As demonstrated in Table 3, Boltz2-based distillation consistently improves generation quality, indicating that the benefit of distillation is not limited to low-capacity generative models.

5. Structure-guided Ligand Discovery

We further evaluate Boltz2 representations in an online structure-guided ligand discovery setting (Passaro et al., 2025; Cretu et al., 2025). This task is formulated as an online reinforcement learning problem, where a molecular generative policy is iteratively updated using binding affinities of generated molecules to discover ligands that bind to a target protein. In this paper, we investigate whether Boltz2 representations can provide auxiliary supervision beyond scalar binding affinities to guide molecular generative policies and accelerate ligand discovery. It is noteworthy that our work is the first to improve online reinforcement learning using representation alignment-based distillation.

5.1. Experimental Setup

Target proteins. We use seven target proteins as benchmarks. First, we consider four benchmark targets from the Boltz2 paper, where Boltz2 shows strong binding affinity prediction performance (Passaro et al., 2025): TYK2, CDK2, JNK1, and P38. As these targets are all kinases, we additionally include three proteins from distinct classes: CA2 as a metalloenzyme, THROMBIN as a serine protease, and FABP4 as a lipid-binding protein, covering diverse protein domains and ligand interaction mechanisms.

Implementation details. We use the SynFlowNet-Boltz pipeline introduced in the Boltz2 paper (Passaro et al., 2025) for online ligand discovery. In this pipeline, the SynFlowNet policy, parameterized by a four-layer graph transformer (Yun et al., 2019; Cretu et al., 2025), sequentially selects actions to complete a ligand molecule at each iteration. Then, Boltz2 computes binding affinity scores for the generated molecules based on protein-ligand structure predictions. These scores are used as rewards to update the policy toward generating higher-score ligands.

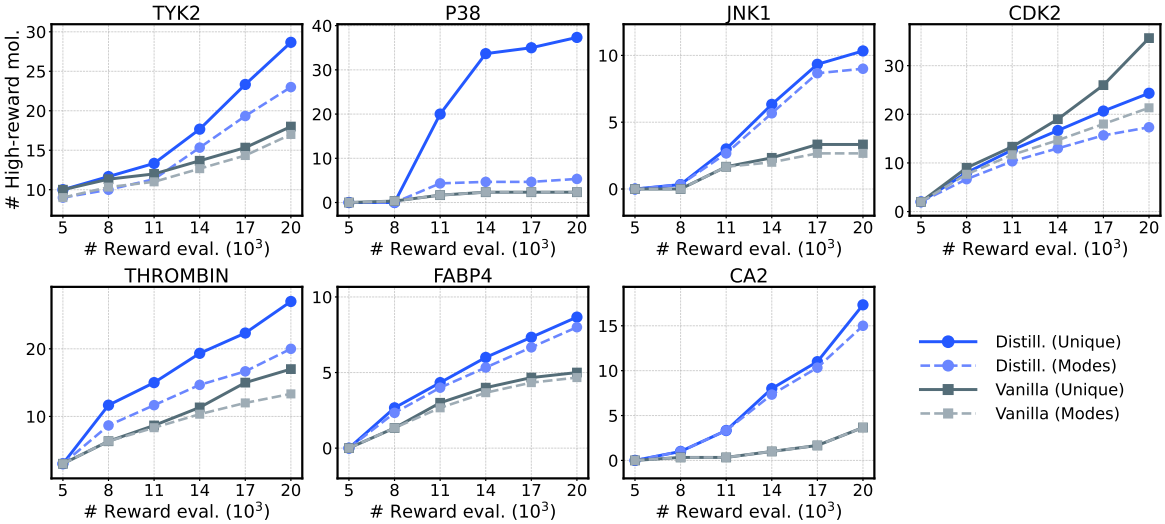


Figure 5. **Results on structure-guided ligand discovery.** The results are averaged over three random seeds. Representation alignment with Boltz2 improves the sample efficiency for discovering high-score molecules that bind to target structures.

In our experiments, we extend this setup by incorporating representation alignment-based distillation into the policy. Specifically, we maximize the cosine similarity between the policy's second-layer single representations and corresponding Boltz2 single representations. Note that representations are aligned on ligand molecules, using Boltz2 ligand representations that are obtained as a byproduct of the Boltz2 binding affinity computation. The overall implementation and hyperparameters follow prior settings (Passaro et al., 2025), with details provided in Appendix C.

Metrics. We evaluate the performance using two metrics. First, we measure the number of high-score molecules discovered as a function of the number of reward evaluations during online training, reflecting sample efficiency. A molecule is considered high-reward if its Boltz2 screening score (Passaro et al., 2025) exceeds 0.75. For CA2, we adopt a higher threshold of 1.2 to account for metal coordination effects. Second, we measure the number of modes, defined as the number of distinct high-scoring molecules with pairwise Tanimoto similarity below 0.6.

5.2. Results

We present the results of structure-guided ligand discovery in Figure 5. For six target proteins, one can see that incorporating representation alignment with Boltz2 consistently increases the number of discovered high-reward molecules compared to the vanilla pipeline under the same reward evaluation budget, i.e., the same number of Boltz2 binding affinity computations for newly generated molecules. Note that representation alignment with Boltz2 also promotes exploration, as evidenced by an increase in the number of discovered high-reward modes.

Table 4. **Layer-wise evaluation of Boltz2 representations. Bold** numbers indicate the best performance. Performance varies across layers depending on the task, while multi-layer concatenation generally yields promising results.

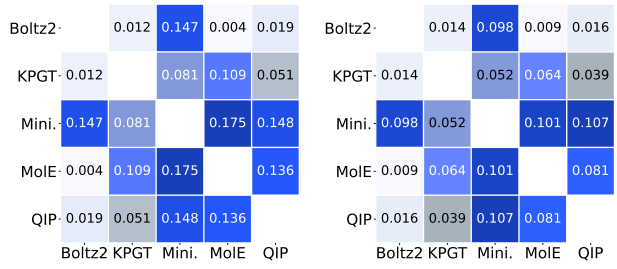
Method	16th	32th	48th	64th	Concat
Solubility ↓	0.67	0.67	0.66	**0.64**	0.66
BBB ↑	0.92	0.92	0.92	0.91	**0.93**
CYP2C9 V. ↑	0.81	**0.82**	**0.82**	0.81	**0.82**
Half Life ↑	0.60	**0.63**	0.61	0.62	0.62
LD50 ↓	0.41	0.41	0.40	**0.39**	0.40

Overall, these results show that atom-level representations from Boltz2 provide a stronger training signal than scalar rewards alone in online structure-guided discovery. By distilling interaction-aware representations computed during binding affinity prediction, the policy is guided toward high-rewarding regions, resulting in a faster discovery of high-affinity and diverse ligands. In Figure 8 of Appendix D.1, we additionally conduct ablation studies on variants of representation alignment in SynFlowNet-Boltz.

6. Ablation Studies on Boltz2 Representations

In this section, we conduct ablation studies to empirically analyze Boltz2 molecular representations in ADMET property prediction tasks.

Downstream performance varies across Boltz2 layers. We analyze the contribution of representations extracted from different layers of Pairformer trunk to downstream performance. Specifically, we extract pair representations from the $\{16, 32, 48, 64\}$-th layers and evaluate them using one task from each ADMET category. We use the same probing setup as in Section 3. As shown in Table 4, performance



(a) CKNNA ($K = 10$) *(b)* CKNNA ($K = 100$)

Figure 6. **Representation alignment between Boltz2 vs. existing molecular foundation models.** Boltz2 exhibits relatively weak CKNNA with existing molecular foundation models.

Table 5. **Combining Boltz2 representations with complementary representations.** Combining Boltz2 representations with complementary representations (MolE; see Figure 6) significantly improves performance on some benchmarks compared to combinations with more highly aligned representations (MiniMol).

Model	CYP2C9 V.	CYP2D6 V.	CYP3A4 V.
MolE	0.80	0.68	0.87
Mini.	0.82	0.72	0.88
Boltz2	0.82	0.69	0.85
Boltz2$^{\text{MolE}}$	**0.88**(+.06)	**0.77**(+.08)	**0.89**(+.04)
Boltz2$^{\text{Mini.}}$	0.86(+.04)	0.72(+.03)	0.88(+.03)

varies across layers depending on the task, indicating that task-relevant information is distributed across the depth of the Pairformer trunk. While no single layer consistently dominates, concatenating representations from multiple layers yields strong overall performance.

Boltz2 yields a distinct representation space that complements existing molecular foundation models. We analyze representation alignment between Boltz2 and existing molecular foundation models using the CKNNA metric described in Section 4. According to the results in Figure 6, Boltz2 shows relatively low average alignment with existing molecular foundation models. This suggests that Boltz2 learns a representation space for small molecules that is distinct from those of existing small-molecule foundation models, while consistently achieving strong performance across predictive and generative downstream tasks.

This observation raises the question of whether low representation alignment reflects complementary information that can be exploited through ensembling. While the performance improvements of Boltz2$^{\text{Mini.}}$ in Table 1 support this hypothesis, we further test this using an ensemble with lower alignment to Boltz2, namely Boltz2+MolE. As shown in Table 5, the Boltz2+MolE ensemble yields larger gains on several benchmarks than Boltz2+MiniMol, despite MiniMol exhibiting stronger performance than MolE. We provide full results of Boltz2+MolE in Table 8 of Appendix D.2.

Table 6. **Boltz2 performance with protein-context.** Incorporating protein context improves performance on enzyme inhibition tasks, but provides limited benefit when local features are critical.

Method	Without protein	With protein
Molecular (global) structure inhibits enzyme		
CYP2C9 V. ↑	0.82	0.83(+.01)
CYP2D6 V. ↑	0.69	0.72(+.03)
CYP3A4 V. ↑	0.85	0.87(+.02)
Enzyme substrates molecular (local) substructures		
CYP2C9 S. ↑	0.36	0.35(-.01)
CYP2D6 S. ↑	0.52	0.53(-.01)
CYP3A4 S. ↑	0.62	0.61(+.01)

Incorporating protein context as inputs can improve small molecular downstream performance. In Section 3, we omit protein inputs to use isolated molecular representations for property prediction. Here, we study how incorporating protein context affects downstream performance. In particular, we focus on metabolism tasks, which predict interactions between proteins and small molecules. We incorporate CYP2C9, CYP2D6, and CYP3A4 protein sequences as inputs and evaluate pooled pair representations over atom-atom and residue-atom indices.

The results are reported in Table 6. One can see that incorporating protein context improves performance on inhibition-related metabolism tasks. This indicates that Boltz2 representations can further benefit from additional task-relevant protein context. However, performance remains limited on substrate-related metabolism tasks. We hypothesize that this limitation stems from a failure to capture local chemical transformations of molecules, due to our global pooling strategy or inherent limitations of Boltz2 in understanding complete atom-level mechanisms (Masters et al., 2025).

7. Conclusion

In this work, we evaluate whether the protein-centric co-folding model Boltz encodes transferable atom-level representations for small molecules. We show that Boltz representations yield strong performance on molecular property prediction, small molecular generation, and structure-guided ligand discovery. These results position Boltz as a strong baseline for atom-level small-molecule representation models and highlight co-folding-based pretraining as an effective molecular representation strategy.

An interesting avenue for future work is to investigate the use of protein context as a "task-specific prompting strategy" to reinforce small-molecule representations across downstream tasks. In addition, we are the first to show that representation alignment-based distillation is effective for online reinforcement learning, which could be further explored in other general domains.

Impact Statement

Boltz2 representations have a potential to accelerate molecular discovery and improve property predictions, contributing to drug design advancements. However, there is a risk of misuse, including the intentional design of harmful compounds. We therefore emphasize the importance of responsible deployment, along with appropriate safeguards.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630 (8016):493–500, 2024.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L. H., Zielinski, M., Sargeant, T., Schneider, R. G., Senior, A. W., Jumper, J., Hassabis, D., Kohli, P., and Žiga Avsec. Accurate proteome-wide missense variant effect prediction with alphamissense. *Science*, 381(6664):eadg7492, 2023.

Cretu, M., Harris, C., Igashov, I., Schneuing, A., Segler, M., Correia, B., Roy, J., Bengio, E., and Lio, P. Synflownet: Design of diverse and novel molecules with synthesis constraints. In *International Conference on Learning Representations*, 2025.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2019.

Dixon, S. L., Duan, J., Smith, E., Von Bargen, C. D., Sherman, W., and Repasky, M. P. Autoqsar: an automated machine learning tool for best-practice quantitative structure–activity relationship modeling. *Future medicinal chemistry*, 8(15):1825–1839, 2016.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Gao, Z., Ji, X., Zhao, G., Wang, H., Zheng, H., Ke, G., and Zhang, L. Uni-qsar: an auto-ml tool for molecular property prediction. *arXiv preprint arXiv:2304.12239*, 2023.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Hayes, T., Rao, R., Akin, H., Sofroniew, N. J., Oktay, D., Lin, Z., Verkuil, R., Tran, V. Q., Deaton, J., Wiggert, M., et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

Huang, K., Fu, T., Glass, L. M., Zitnik, M., Xiao, C., and Sun, J. Deeppurpose: a deep learning library for drug–target interaction prediction. *Bioinformatics*, 36(22-23): 5545–5547, 2020.

Huang, K., Fu, T., Gao, W., Zhao, Y., Roohani, Y. H., Leskovec, J., Coley, C. W., Xiao, C., Sun, J., and Zitnik, M. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Huh, M., Cheung, B., Wang, T., and Isola, P. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Jang, H., Jang, Y., Kim, M., Park, J., and Ahn, S. Pessimistic backward policy for gflownets. In *Advances in Neural Information Processing Systems*, 2024a.

Jang, Y., Kim, D., and Ahn, S. Graph generation with k^2-trees. *arXiv preprint arXiv:2305.19125*, 2023.

Jang, Y., Lee, S., and Ahn, S. A simple and scalable representation for graph generation. In *International Conference on Learning Representations*, 2024b.

Jo, J., Lee, S., and Hwang, S. J. Score-based generative modeling of graphs via the system of stochastic differential equations. In *International Conference on Machine Learning*, 2022.

Jo, J., Kim, D., and Hwang, S. J. Graph generation with diffusion mixture. In *International Conference on Machine Learning*, 2024.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Kim, J., Chang, W., Ji, H., and Joung, I. Quantum-informed molecular representation learning enhancing admet property prediction. *Journal of Chemical Information and Modeling*, 64(13):5028–5040, 2024.

Klaser, K., Banaszewski, B., Maddrell-Mander, S., McLean, C., Müller, L., Parviz, A., Huang, S., and Fitzgibbon, A. W. Minimol: A parameter-efficient foundation model for molecular learning. In *ICML 2024 Workshop on Efficient and Accessible Foundation Models for Biological Discovery*, 2024.

Kong, L., Cui, J., Sun, H., Zhuang, Y., Prakash, B. A., and Zhang, C. Autoregressive diffusion model for graph generation. In *International Conference on Machine Learning*, 2023.

Li, H., Zhang, R., Min, Y., Ma, D., Zhao, D., and Zeng, J. A knowledge-guided pre-training framework for improving molecular representation learning. *Nature Communications*, 14(1):7568, 2023.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

Masters, M. R., Mahmoud, A. H., and Lill, M. A. Investigating whether deep learning models for co-folding learn the physics of protein-ligand interactions. *Nature Communications*, 16(1):8854, 2025.

Méndez-Lucio, O., Nicolaou, C. A., and Earnshaw, B. Mole: a foundation model for molecular graphs using disentangled attention. *Nature Communications*, 15(1):9431, 2024.

Okabe, K., Koshinaka, T., and Shinoda, K. Attentive statistics pooling for deep speaker embedding. *arXiv preprint arXiv:1803.10963*, 2018.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Ouyang-Zhang, J., Diaz, D., Klivans, A., and Krähenbühl, P. Predicting a protein's stability under a million mutations. In *Advances in Neural Information Processing Systems*, 2023.

Passaro, S., Corso, G., Wohlwend, J., Reveiz, M., Thaler, S., Somnath, V. R., Getz, N., Portnoi, T., Roy, J., Stark, H., Kwabi-Addo, D., Beaini, D., Jaakkola, T., and Barzilay, R. Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv*, 2025.

Preuer, K., Renz, P., Unterthiner, T., Hochreiter, S., and Klambauer, G. Fréchet chemnet distance: a metric for generative models for molecules in drug discovery. *Journal of chemical information and modeling*, 58(9):1736–1741, 2018.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Seo, H., Kim, T., Yu, S., and Ahn, S. Learning flexible forward trajectories for masked molecular diffusion. *arXiv preprint arXiv:2505.16790*, 2025.

Stahl, K., Graziadei, A., Dau, T., Brock, O., and Rappsilber, J. Protein structure prediction with in-cell photo-crosslinking mass spectrometry and deep learning. *Nature Biotechnology*, 41(12):1810–1819, 2023.

Sypetkowski, M., Wenkel, F., Poursafaei, F., Dickson, N., Suri, K., Fradkin, P., and Beaini, D. On the scalability of gnns for molecular graphs. In *Advances in Neural Information Processing Systems*, 2024.

Vignac, C., Krawczuk, I., Siraudin, A., Wang, B., Cevher, V., and Frossard, P. Digress: Discrete denoising diffusion for graph generation. In *International Conference on Learning Representations*, 2023.

Weininger, D. 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.

Wohlwend, J., Corso, G., Passaro, S., Getz, N., Reveiz, M., Leidal, K., Swiderski, W., Atkinson, L., Portnoi, T., Chinn, I., Silterra, J., Jaakkola, T., and Barzilay, R. Boltz-1: Democratizing biomolecular interaction modeling. *bioRxiv*, 2024.

Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., and Xie, S. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025.

Yun, S., Jeong, M., Kim, R., Kang, J., and Kim, H. J. Graph transformer networks. In *Advances in Neural Information Processing Systems*, 2019.

Zaidi, S., Schaarschmidt, M., Martens, J., Kim, H., Teh, Y. W., Sanchez-Gonzalez, A., Battaglia, P., Pascanu, R., and Godwin, J. Pre-training via denoising for molecular property prediction. In *International Conference on Learning Representations*, 2023.

Zheng, B., Ma, N., Tong, S., and Xie, S. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025.

Zhou, G., Gao, Z., Ding, Q., Zheng, H., Xu, H., Wei, Z., Zhang, L., and Ke, G. Uni-mol: A universal 3d molecular representation learning framework. In *International Conference on Learning Representations*, 2023.

Table 7. **Data statistics and evaluation metrics of the TDC ADMET benchmark.**

Dataset	Task	Metric	# Molecules
Caco2 (Caco-2 Permeability)	Regression	MAE	906
HIA (Human Intestinal Absorption)	Classification	AUROC	578
PgP (P-glycoprotein Inhibition)	Classification	AUROC	1,213
Bioavailability (Oral Bioavailability)	Classification	AUROC	640
Lipophilicity (LogD)	Regression	MAE	4,200
Solubility (Aqueous Solubility)	Regression	MAE	1,144
BBB (Blood–Brain Barrier Penetration)	Classification	AUROC	2,050
PPBR (Plasma Protein Binding Rate)	Regression	MAE	1,797
VDss (Volume of Distribution at Steady State)	Regression	Spearman	1,130
CYP2C9 V. (CYP2C9 Inhibition)	Classification	AUPRC	12,092
CYP2D6 V. (CYP2D6 Inhibition)	Classification	AUPRC	13,130
CYP3A4 V. (CYP3A4 Inhibition)	Classification	AUPRC	12,328
CYP2C9 S. (CYP2C9 Substrate)	Classification	AUPRC	666
CYP2D6 S. (CYP2D6 Substrate)	Classification	AUPRC	667
CYP3A4 S. (CYP3A4 Substrate)	Classification	AUROC	667
Half (Half-life)	Regression	Spearman	1,039
Clearance H. (Clearance Hepatocyte)	Regression	Spearman	1,102
Clearance M. (Clearance Microsome)	Regression	Spearman	1,102
LD50 (Acute Toxicity)	Regression	MAE	7,385
hERG (hERG Blockade)	Classification	AUROC	648
AMES (AMES Mutagenicity)	Classification	AUROC	6,512
DILI (Drug-Induced Liver Injury)	Classification	AUROC	475

A. Experimental Setup of ADMET Property Prediction

A.1. Benchmark Setup

We follow the standard TDC ADMET benchmark protocol (Huang et al., 2021) and use random seeds $\{1, 2, 3, 4, 5\}$. Data statistics and evaluation metrics are provided in Table 7.

A.2. Implementation Details

Pooling. We concatenate pair representations extracted from the 16th, 32nd, 48th, and 64th Pairformer layers, which correspond to the quarter, half, three-quarter, and final depths of the Pairformer stack, to obtain a 512-dimensional pair representation. Unlike existing small molecular foundation models, Boltz do not train a pooling layer to extract informative fixed-length representation for downstream tasks. Thus, we instead apply hybrid pooling to obtain a informative fixed-length representation. We concatenate pooled representations over diagonal entries, bonded atom-pair entries, and all entries in the pair representation. To be specific, we apply statistic pooling (Okabe et al., 2018), which concatenates mean and standard deviation vectors. The resulting representation has a dimensionality of 3072. Although the nominal dimensionality is high, strong correlations between features reduce the effective dimensionality.

Probing Network and Training Configuration. Our detailed probing and evaluation setups follow prior work (Klaser et al., 2024). We adopt a lightweight probing setup, where a task-specific MLP head is trained on top of frozen molecular representations. The hidden dimension and the number of layers of the MLP are selected from $\{512, 1024, 2048\}$ and $\{2, 3, 4\}$, respectively. Here, the hidden representation at each layer is concatenated with the input, following the model design in prior work (Klaser et al., 2024). The dropout rate is fixed to 0.0. The learning rate is selected from $\{1e{-}4, 3e{-}4, 5e{-}4\}$, and the weight decay is $1e{-}5$. Models are trained for 25 or 200 epochs. We use 32 batch size. Early stopping is applied with a patience of 25 epochs, and a cosine learning-rate scheduler is used. The hyper-parameters are selected based on validation performance. For each random seed, five-fold cross-validation is performed by mixing the training and validation sets. Following prior work (Klaser et al., 2024), the predictions on the test dataset are obtained by averaging the outputs of five models trained using five-fold cross-validation on the training and validation datasets.

B. Experimental Setup of Molecular Generation

We apply representation alignment (REPA) (Yu et al., 2025) to a denoising diffusion–based molecular generative model, specifically GruM (Jo et al., 2024). In this experiment, we use the official codebase of GruM.

B.1. Measuring CKNNA

We measure representation alignment using Centered Kernel Nearest-Neighbor Alignment (CKNNA) (Huh et al., 2024), which evaluates local alignment between two representation spaces based on shared nearest-neighbor structure. Given a set of molecules $\{m_i\}_{i=1}^N$ and two representation models f and g, yielding representations $\{f(m_i)\}_{i=1}^N$ and $\{g(m_i)\}_{i=1}^N$, CKNNA computes the alignment score as follows:

$$\begin{aligned}&\mathrm{ALIGN}\big(\{f(m_i)\}_{i=1}^N, \{g(m_i)\}_{i=1}^N\big) = \\ &\frac{1}{(N-1)^2}\sum_i\sum_j \alpha(i,j)\big(\langle f(m_i), f(m_j)\rangle - \mathbb{E}_l[\langle f(m_i), f(m_l)\rangle]\big)\big(\langle g(m_i), g(m_j)\rangle - \mathbb{E}_l[\langle g(m_i), g(m_l)\rangle]\big),\end{aligned}$$

where $\alpha(i,j)$ selects pairs of samples that lie within local neighborhoods:

$$\alpha(i,j;k) = \mathbb{1}\big[i \neq j \ \wedge \ f(m_j) \in \mathrm{KNN}(f(m_i);k) \ \wedge \ g(m_j) \in \mathrm{KNN}(g(m_i);k)\big], \tag{1}$$

and $\mathrm{KNN}(\cdot;k)$ denotes the set of k nearest neighbors under cosine similarity. CKNNA is then computed as

$$\mathrm{CKNNA} = \frac{\mathrm{ALIGN}\big(\{f(m_i)\}_{i=1}^N, \{g(m_i)\}_{i=1}^N\big)}{\sqrt{\mathrm{ALIGN}\big(\{f(m_i)\}_{i=1}^N, \{f(m_i)\}_{i=1}^N\big)\mathrm{ALIGN}\big(\{g(m_i)\}_{i=1}^N, \{g(m_i)\}_{i=1}^N\big)}}. \tag{2}$$

In our experiments, we use $N = 10{,}000$ molecules from ZINC250k dataset to compute CKNNA score, where the neighborhood size is $k = 10$. We use second-layer representations from molecular generative models and concatenated representations from the 16th, 32nd, 48th, and 64th layers of Boltz2 Pairformer trunk.

B.2. Implementation Details

Denoising model. The denoising network is parameterized by a four-layer Pairformer architecture with 4 attention heads, using hidden dimensions of 256 for single-node representations and 128 for pair representations. Single-node representations are used to denoise atom attributes of the molecular graph using a two-layer MLP with a hidden dimension of 256, while pair representations are used to denoise bond attributes using a two-layer MLP with a hidden dimension of 128. Before being passed into the Pairformer, pair representations are initialized by transforming edge features, concatenations of atom-pair features, and a time conditioning signal with separate two-layer MLPs and combining the resulting embeddings additively. Single representations are initialized from atom features and the time conditioning signal via two-layer MLPs.

Training configuration. The learning rate is set to $3\mathrm{e}{-4}$ with a cosine learning-rate scheduler. The batch size is 512, and models are trained for 100 epochs. All other hyper-parameters, including weight decay, adjacency matrix scaling, and noise scheduling, follow the default GruM configuration (Jo et al., 2024).

Representation alignment. We adopt a representation alignment-based distillation objective to train molecular generative models with auxiliary supervision from Boltz2 representations. Given a noisy molecule m_t and its corresponding clean molecule m, the training objective is defined as follows:

$$\mathcal{L}(m_t, m) = \mathcal{L}_{\text{GruM}}\big(s_\theta(m_t), m\big) - \lambda \cdot \cos\big(h_\theta(m_t), f(m)\big),$$

where $\mathcal{L}_{\text{GruM}}$ denotes the original denoising loss of GruM, $s_\theta(m_t)$ is the denoising prediction, $h_\theta(m_t)$ denotes the hidden representation of the generative model, and $f(m)$ denotes the frozen Boltz2 representation of the clean molecule. The cosine similarity is computed between the generative model representations and the Boltz2 representations, and λ denotes the coefficient for representation alignment. We set $\lambda = 4$ in our experiments. Note that representation alignment is applied at the second Pairformer layer of the denoising model for both single and pair representations, with the alignment target defined as the concatenated representations from the $\{16, 32, 48, 64\}$-th layers of the Boltz2 Pairformer trunk.

For distillation, we additionally introduce a lightweight distillation network parameterized as a two-layer MLP with a hidden dimension of 1536. The distillation network maps generative model representations to the Boltz2 representation space. Boltz2 representations $f(m)$ are precomputed for all molecules. We also flatten the single and pair representations when applying representation alignment, while masking out-of-range indices.

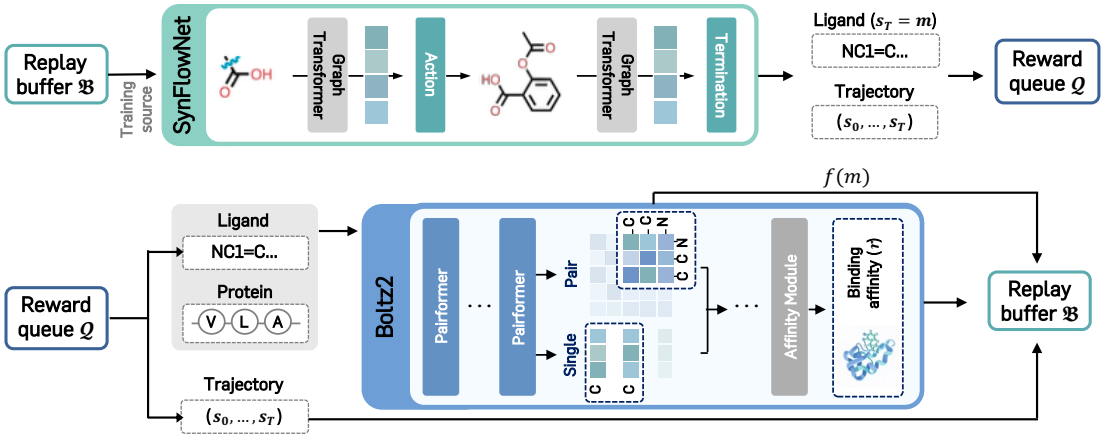


Figure 7. **Online SynFlowNet-Boltz ligand discovery pipeline.**

C. Experimental Setup of Structure-guided Ligand Discovery

We use SynFlowNet-Boltz ligand discovery pipeline introduced in prior work (Passaro et al., 2025; Cretu et al., 2025). The task is formulated as an online reinforcement learning problem, where a molecular policy is iteratively updated based on binding affinity rewards computed by Boltz2. In this experiment, we use the official codebase of SynFlowNet-Boltz.

Policy network. At each generation step, the policy P_F sequentially selects reaction templates and reactants to construct a molecule, yielding a trajectory $\tau = (s_0, \ldots, s_T)$ of molecular construction steps. The policy is parameterized by a four-layer graph transformer (Yun et al., 2019), and we do not modify the original implementation (Passaro et al., 2025; Cretu et al., 2025). The graph transformer outputs 128-dimensional atom-wise single representations, which are pooled and passed through an MLP to produce a distribution over valid actions. The action space is the same as that of SynFlowNet (Cretu et al., 2025), including reactant and reaction template selections. The backward policy in SynFlowNet is also implemented as a four-layer graph transformer and trained using a maximum likelihood–based approach, following prior implementations (Cretu et al., 2025; Jang et al., 2024a).

Reward computation. For each generated terminal molecule $s_T = m$, Boltz2 computes a reward r as follows:

$$r = \max\Big(\frac{-\text{affinity} + 2}{4}, 0\Big) \cdot \text{likelihood}$$

where affinity is the Boltz2-predicted binding affinity score for molecule m, and likelihood denotes the Boltz2-predicted likelihood of binding to the corresponding binding site of the target protein.

Here, Boltz2 uses single and pair representations, along with 3D structure prediction, for protein–ligand binding affinity predictions. Thus, we can naturally obtain single representations of ligand molecules within the SynFlowNet–Boltz pipeline.

Representation alignment. In addition to scalar rewards, we incorporate representation alignment to provide auxiliary supervision for policy training. Specifically, for each terminal-state molecule, we extract single representations from Boltz2 that are computed during the binding affinity prediction process. We align these representations with the hidden single representations from the second layer of the policy network by maximizing cosine similarity. Here, we basically consider representation alignment to terminal-state molecules $s_T = m$:

$$\mathcal{L}_{\text{align.}}(m, f(m)) = -\lambda \cdot \cos\big(h_\theta(m), f(m)\big),$$

where $h_\theta(m)$ denotes the hidden representation of the graph transformer, and $f(m)$ denotes the frozen Boltz2 representation. Here, representation alignment is applied at the second graph transformer layer of the policy model, aligning its representations with those obtained from the Boltz2 Pairformer trunk. Since the graph transformer in the policy network outputs only single atom-wise representations, alignment is enforced only for single representations, and Boltz2 pair representations are omitted. In this setting, Boltz2 representations span both ligand atoms and protein residues. We therefore slice the representations to retain only the ligand atom representations $f(m)$.

For distillation, we also introduce a lightweight distillation network parameterized as a two-layer MLP with a hidden dimension of 512. We set the representation alignment coefficient to $\lambda = 10$.

Algorithm 1 Learning SynFlowNet with Representation Alignment

1: Initialize replay buffer $\mathcal{B}$, trajectory queue $\mathcal{Q}$, and SynFlowNet policy P_θ
2: **repeat**
3: Sample a batch of molecular generative trajectories $\{\tau^{(k)} = (s_0^{(k)}, \ldots, s_T^{(k)} = m^{(k)})\}_{k=1}^K$ from the policy P_θ
4: Update $\mathcal{Q} \leftarrow \mathcal{Q} \cup \{\tau^{(k)}\}_{k=1}^K$
5: Sample a batch of trajectories with rewards and molecular representations $\{\tau^{(k)}, r^{(k)}, f(m^{(k)})\}_{k=1}^K$ from $\mathcal{B}$
6: Update P_F by minimizing $\mathcal{L}_{\text{SynFlowNet}}(\tau, r) + \mathcal{L}_{\text{align.}}(m, f(m))$ using $\{\tau^{(k)}, r^{(k)}, f(m^{(k)})\}_{k=1}^K$
7: **until** converged

Algorithm 2 Asynchronous Boltz2 Worker

1: **repeat**
2: Sample the latest trajectories $\{\tau^{(k)}\}_{k=1}^K$ from the queue $\mathcal{Q}$
3: Compute rewards and molecular representations $\{(r^{(k)}, f(m^{(k)}))\}_{k=1}^K$ using Boltz2
4: Update $\mathcal{B} \leftarrow \mathcal{B} \cup \{\tau^{(k)}, r^{(k)}, f(m^{(k)})\}_{k=1}^K$
5: **until** converged

We also study intermediate-state distillation as an ablation in Appendix D.1. This is defined as follows:

$$\mathcal{L}_{\text{align.}}^{\text{inter.}}(\tau = (s_0, \ldots, s_T), f(m)) = -\lambda \cdot \sum_{t=0}^{T} \cos\big(h_\theta(s_t), f(m)_{\mathcal{I}(s_t)}\big).$$

Here, $\mathcal{I}(s_t)$ denotes the index set of atoms in the terminal molecule s_T that correspond to the molecular substructure present at the intermediate state s_t, and $f(m)_{\mathcal{I}(s_t)}$ denotes the representations of this corresponding substructure. In this ablation objective, an intermediate molecule s_t may lead to multiple terminal molecules through different trajectory actions, resulting in multiple possible alignment targets for $h_\theta(s_t)$. To address this, we restrict alignment to representations of terminal molecules in the top 10% by reward.

Optimization and training. The policy is trained using SynFlowNet objective $\mathcal{L}_{\text{SynFlowNet}}$, augmented with the representation alignment loss $\mathcal{L}_{\text{align.}}$. As illustrated in Figure 7, Algorithm 1, and Algorithm 2, policy optimization and reward computation of SynFlowNet-Boltz are decoupled and executed asynchronously (Passaro et al., 2025). The policy process continuously samples trajectories, pushes them to the reward queue $\mathcal{Q}$, and trains on trajectories, rewards, and representations sampled from the replay buffer $\mathcal{B}$. The Boltz2 worker process computes binding affinity rewards and representations for queued trajectories and pushes them to the replay buffer $\mathcal{B}$. We use four NVIDIA B200 GPUs for policy training and run 16 parallel Boltz2 worker processes for asynchronous reward computation.

Note that we initialize the replay buffer $\mathcal{B}$ with $5{,}000$ trajectories sampled from multiple warm-up policies. All other implementation details, including replay buffer size, reward normalization, and exploration strategy, follow prior work without modification (Passaro et al., 2025).

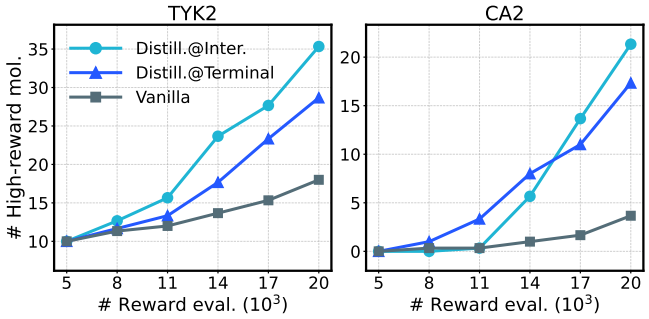


Figure 8. **Representation alignment on intermediate molecules.**

D. Additional Results

D.1. Intermediate State Distillation

We further conduct an ablation study that extends representation alignment-based distillation from generated molecules to intermediate molecules produced by the policy before generation, for example, the intermediate molecular representations shown in Figure 7. To be specific, we consider alignment loss $\mathcal{L}_{\text{align.}}^{\text{inter.}}$ described in Appendix C. We report the results in Figure 8. Here, one can observe that distillation to intermediate molecules yields consistent improvements, indicating that the benefits of representation alignment are robust to algorithmic design choice.

D.2. Full Results of Representation Ensembling

We report the full results in Table 8. Here, one can observe that Boltz$^{\text{MolE}}$ outperforms Boltz$^{\text{Mini.}}$ on some benchmarks, although MiniMol shows strong standalone performance compared to MolE.

Table 8. **Additional results on ADMET property prediction benchmark.**

Dataset	Boltz2	Boltz2$^{\text{Mini.}}$	Boltz2$^{\text{MolE}}$
Absorption			
Caco2 ↓	0.30	0.30	0.30
HIA ↑	0.99	0.99	0.99
Pgp ↑	0.93	0.93	0.93
Bioavailability ↑	0.75	0.77	0.72
Lipophilicity ↓	0.45	0.41	0.43
Solubility ↓	0.66	0.64	0.65
Distribution			
BBB ↑	0.93	0.94	0.92
PPBR ↓	7.65	7.59	7.34
VDss ↑	0.74	0.75	0.75
Metabolism			
CYP2C9 V. ↑	0.82	0.86	0.88
CYP2D6 V. ↑	0.69	0.72	0.77
CYP3A4 V. ↑	0.85	0.88	0.89
CYP2C9 S. ↑	0.36	0.36	0.37
CYP2D6 S. ↑	0.52	0.51	0.55
CYP3A4 S. ↑	0.62	0.60	0.64
Excretion			
Half Life ↑	0.62	0.65	0.66
Clearance H. ↑	0.62	0.60	0.57
Clearance M. ↑	0.61	0.65	0.65
Toxicity			
LD50 ↓	0.40	0.40	0.40
hERG ↑	0.86	0.86	0.83
AMES ↑	0.91	0.91	0.92
DILI ↑	0.89	0.87	0.87

Reproducibility Statement

Code and datasets. We provide the code and datasets for ADMET property prediction at https://anonymous.4open.science/r/ADMET-E838/README.MD.

Boltz2. We modify Boltz2 to operate in a ligand-only setting. The modification is as follows:

- Boltz2 yields errors when the input protein sequence is empty. We therefore use a single `X` token as the protein sequence input and modify the Pairformer trunk to ignore it by slicing the corresponding indices.
- Boltz2 uses 3D conformations computed with the Universal Force Field (UFF) implemented in RDKit. For molecules whose 3D conformations cannot be initialized with UFF, we initialize conformations by sampling from $[-1, 1]$.

In the code, we include the modified Boltz2 implementation used in our work.