A Systematic Evaluation of Co-folding Model Representations for Small-Molecule Learning

Hyosoon Jang **Hyunjin Seo** **Honghui Kim** **Seonghyun Park** **Taewon Kim**
Yunhui Jang **Sungsoo Ahn**
KAIST
{hyosoon.jang,sungsoo.ahn}@kaist.ac.kr

Abstract

Small-molecule foundation models are typically pretrained on standalone molecular data, unlike vision and language models that often benefit from cross-modal or relational supervision. Protein-ligand co-folding provides a molecular analogue of such supervision by exposing models to atom-level ligand-protein interactions, raising the question of whether co-folding models can yield strong small-molecule representations. We study this question using Boltz2, a modern co-folding model, by transferring its atom-level ligand representations to standalone small-molecule tasks. Through systematic probing and distillation, we show that Boltz2 representations match or outperform existing models on the ADMET benchmark, accelerate molecular generative modeling, and improve sample efficiency in structure-guided ligand optimization. We further find that Boltz2 representations are complementary to those learned from conventional standalone molecular supervision, including 3D conformers, bioassay labels, and quantum-chemical properties. Finally, we extend representation alignment to reinforcement learning, showing that dense representation-level supervision can complement scalar rewards in molecular discovery. These results identify protein-ligand co-folding as a promising pretraining paradigm for small-molecule representation learning and position Boltz2 as a strong, off-the-shelf molecular foundation model.[1]

1 Introduction

Large-scale representation learning has shown strong transferability across a wide range of downstream tasks, particularly in language [1, 2] and vision domains [3, 4]. Foundation models with large-scale pretraining have become essential for achieving state-of-the-art performance in downstream applications, serving as powerful feature extractors [5] for predictive tasks. Beyond predictive tasks, researchers have leveraged their representations to enhance generative modeling, including distillation into denoising models [6] and use as latent embeddings for latent diffusion models [7].

In the small-molecule domain, this paradigm has motivated foundation models that learn transferable atom-level representations for drug discovery tasks such as ADMET prediction [8]. Most existing models learn from molecules considered in isolation: each training example is a small molecule associated with molecule-level supervision such as bioassay labels and quantum-chemical properties [9, 10, 11, 12, 13, 14]. Although these models differ in architecture and objective, they share a common design choice: the small molecule itself is the primary object of representation learning, without explicit modeling of an interacting molecular partner.

This contrasts with vision and language, where strong representations often arise from supervision beyond standalone data, such as cross-modal or relational signals [5, 15, 16]. We argue that the

[1] Code: https://github.com/hsjang0/boltz-as-FM.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

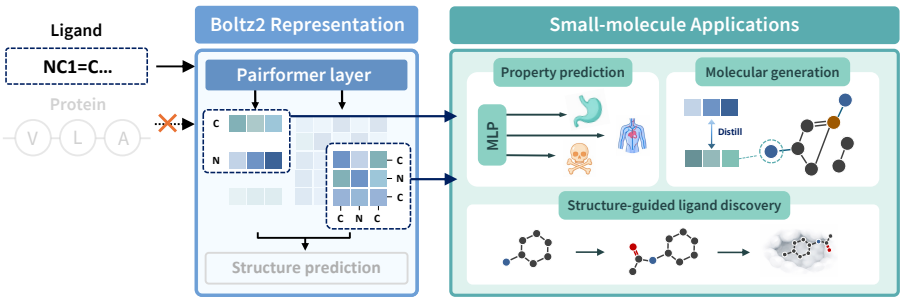


Figure 1: **Boltz2 as an atom-level molecular foundation model.** We repurpose Boltz2, trained for protein–ligand co-folding, as a small-molecule representation model by leveraging its ligand representations and evaluating them on downstream small-molecule tasks.

molecular domain has an analogous but underexplored source of supervision: protein-ligand co-folding structures [17], in which a small molecule is observed within an interacting protein context.

We hypothesize that protein-ligand structures provide richer supervision than standalone molecular data by exposing atom-level interactions in bound conformations, including hydrogen bonding and electrostatic interactions. Timely, modern protein models, such as AlphaFold3 [18] and its open-weight counterpart Boltz [19, 20], have transitioned from residue-level to atom-level representation modeling for protein-ligand structure prediction. Although designed for structure prediction, we can transfer their atomistic representations to small-molecule tasks. We therefore ask:

Can protein-ligand co-folding models provide strong atom-level representations for standalone small-molecule tasks?

Although repurposing protein–ligand co-folding models for small-molecule representation learning is conceptually straightforward, it has not been systematically studied. This gap is important because the broader utility of co-folding models remains debated: recent studies suggest that models such as AlphaFold3 may rely heavily on MSA-derived evolutionary patterns rather than a genuine atom-level understanding of molecular interactions [21, 22]. Establishing this baseline therefore allows us to test whether the interaction-aware supervision learned by modern co-folding models transfers beyond their original structure-prediction objective.

Contribution. In this work, we investigate the potential of a co-folding model, Boltz2, as a source of representations for small-molecule tasks, bridging cutting-edge protein-centric models and atom-level representation learning for small molecules. As illustrated in Figure 1, we extract atom-level representations from Boltz2 for a given small molecule and evaluate their expressiveness across a broad range of small-molecule downstream tasks. Notably, our work is the first to show that informative molecular representations can improve molecular generation and optimization.

Through systematic experiments, we identify four key findings that position modern co-folding models, in particular Boltz2, as a strong baseline for small-molecule foundation models.

- **Predictive transfer (Section 3)**: Boltz2 representations match or outperform specialized small-molecule foundation models on the extensive TDC ADMET benchmark.
- **Generative transfer (Section 4):** Boltz2 representations improve molecular generative models with representation alignment-based distillation [6].
- **Optimization transfer (Section 5):** Boltz2 representations enhance sample efficiency for designing ligands to maximize Boltz2's binding affinity score using a new distillation method.
- **Representation analysis (Section 6):** Boltz2 yields a representation space that complements existing molecular models. We further analyze the effects of protein context and layer depth.

Finally, we provide actionable insights for the molecular machine learning community. First, we extend representation alignment to online reinforcement learning for molecular discovery, showing how dense signals from the reward model can be used to improve sample efficiency beyond scalar rewards alone. Second, we show that combining weakly aligned representations, e.g., Boltz2 with MolE [9], outperforms more strongly aligned combinations, e.g., Boltz2 with MiniMol [10], suggesting a practical strategy for representation fusion.

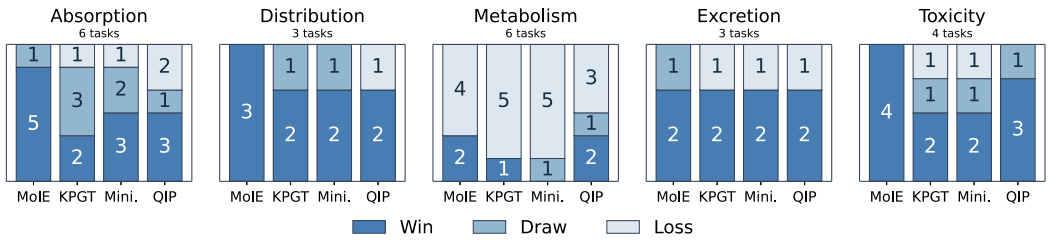


Figure 2: **Boltz2 vs. ADMET specialist foundation models.** Bars show raw win/draw/loss counts. As illustrated, Boltz2 shows competitive performance compared with existing foundation models specialized for ADMET property prediction in four of the five domains.

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 using a large collection of small molecules. MolE [9] uses masked atom prediction, MiniMol [10] leverages large-scale bioassay supervision, QIP [11] incorporates quantum-chemical property regression to capture electronic structure information relevant to molecular properties, and UniMol [12, 23] learns representations through 3D molecular structure prediction. Other approaches incorporate chemical knowledge or scalability: KPGT [13] injects knowledge-guided objectives based on molecular structure and functional groups, UniQSAR [24] advances ADMET prediction using UniMol, and MolGPS [14] studies scalable pretraining across datasets and model sizes.

2.2 Atom-level Representations in Modern 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 [19, 20] and AlphaFold3 [18] are trained to predict 3D structures of protein-ligand complexes at atomic resolution for both proteins and ligands. 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.

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 primarily evaluate Boltz2 [20] in this paper as an atom-level foundation model for small molecules. The model has approximately 200M parameters and a 64-layer Pairformer trunk that produces pair and single representations over protein residues and ligand atoms. 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. For standalone small-molecule tasks (Sections 3 and 4), we omit protein inputs and consider ligands only. For structure-guided ligand discovery (Section 5), we retain protein inputs and use Boltz2 in its native protein-ligand setting.

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 [9, 10, 11, 13, 14]. This benchmark covers absorption, distribution, metabolism, excretion, and toxicity prediction tasks, which depend on both global molecular structure and local atomic interactions.

Table 1: **Results on ADMET property prediction benchmark.** †Other denotes the highest score reported on the TDC ADMET leaderboards as of Jan 2026, excluding the baselines in the table. We evaluate the 1B-scale UniMol2 model. Mini. denotes MiniMol. 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, while underlined numbers indicate the best performance without ensemble. Overall, Boltz2 shows competitive performance compared to baselines. *This performance can be improved by *incorporating relevant protein* as input to Boltz2 (Table 3). See Table 6 in Appendix A for the standard deviations.

		w/ Pretraining						w/ Representation Ensembling		
Dataset	**†Other**	**MolE**	**KPGT**	**Mini.**	**QIP**	**UniMol2**	**Boltz2**	**MolGPS**	**UniQSAR**	**Boltz2$^{Mini.}$**
Absorption										
Caco2 ↓	**0.26**	0.31	0.28	0.35	0.27	0.41	0.30	0.29	0.27	0.30
HIA ↑	**0.99**	0.96	0.98	**0.99**	**0.99**	0.87	**0.99**	0.98	**0.99**	**0.99**
Pgp ↑	0.94	0.92	0.94	0.94	0.93	0.90	0.93	**0.95**	0.93	0.93
Bioavail. ↑	0.75	0.65	0.75	0.69	0.73	0.64	0.75	0.70	0.73	**0.77**
Lipophilicity ↓	0.47	0.47	0.45	0.46	0.44	0.61	0.45	**0.39**	0.42	0.41
Solubility ↓	0.76	0.79	0.71	0.74	0.70	0.81	0.66	0.68	0.68	**0.64**
Distribution										
BBB ↑	0.92	0.90	0.91	0.92	0.90	0.86	0.93	**0.94**	0.93	**0.94**
PPBR ↓	7.44	8.07	7.68	7.70	7.36	8.95	7.65	**6.46**	7.53	7.59
VDss ↑	0.71	0.65	0.63	0.54	0.61	0.70	0.74	0.65	0.73	**0.75**
Metabolism										
CYP2C9 V. ↑	**0.86**	0.80	0.80	0.82	0.78	0.77	0.82*	0.84	0.80	**0.86**
CYP2D6 V. ↑	**0.79**	0.68	0.72	0.72	0.66	0.62	0.69*	0.75	0.74	0.72
CYP3A4 V. ↑	**0.92**	0.87	0.89	0.88	0.87	0.81	0.85*	0.90	0.89	0.88
CYP2C9 S. ↑	0.44	0.45	0.45	0.48	**0.52**	0.33	0.36	0.46	0.45	0.36
CYP2D6 S. ↑	**0.74**	0.70	**0.74**	0.73	0.67	0.44	0.52	0.71	0.72	0.51
CYP3A4 S. ↑	0.67	0.67	**0.73**	0.64	0.62	0.58	0.62	0.68	0.65	0.60
Excretion										
Half Life ↑	0.58	0.55	0.53	0.50	0.53	0.48	0.62	0.63	0.61	**0.65**
Clearance H. ↑	0.54	0.38	0.42	0.45	0.50	0.45	**0.62**	0.57	0.49	0.60
Clearance M. ↑	0.63	0.61	0.64	0.63	**0.66**	0.62	0.61	0.63	0.65	0.65
Toxicity										
LD50 ↓	0.55	0.82	0.55	0.59	0.56	0.50	**0.40**	0.56	0.55	**0.40**
hERG ↑	**0.88**	0.81	0.85	0.85	0.82	0.75	0.86	0.86	0.86	0.86
AMES ↑	0.87	0.88	0.87	0.85	0.86	0.88	**0.91**	0.86	0.88	**0.91**
DILI ↑	0.93	0.58	0.93	**0.96**	0.89	0.83	0.89	0.94	0.94	0.87
# 1st (# 1st w/o Ensembling)		0 (0)	2 (4)	2 (3)	3 (5)	0 (0)	4 (9)	4	1	9
# 2nd (# 2nd w/o Ensembling)		1 (2)	4 (7)	3 (5)	2 (4)	1 (1)	5 (4)	8	8	4

3.1 Experimental Setup

Datasets. We conduct experiments on the Therapeutics Data Commons (TDC) ADMET benchmark datasets [8], which consist of 22 ADMET property prediction tasks for small molecules represented as SMILES [25]. We follow the standard benchmark protocol, including data splits, evaluation metrics, and random seeds, where dataset statistics and benchmark settings are provided in Appendix A.

Implementation details. We apply probing to Boltz2 molecular representations for property prediction. We provide SMILES strings of molecules to the ligand modality of Boltz2, and protein sequence inputs are omitted. Then, we concatenate atom-wise pair representations from the $\{16, 32, 48, 64\}$-th layers of the 64-layer Pairformer trunk. The pooled representation is fed into a probing network for target label prediction. We provide full details in Appendix A.[2]

Baselines. We consider seven foundation models specialized for ADMET property prediction (See comparison with three more foundation models in Table 7 of Appendix B.1):

- **MolE** [9]: pretrained via masked language modeling on atom tokens with auxiliary losses to predict molecular properties and fingerprints.
- **KPGT** [13]: pretrained with knowledge-guided objectives that enforce consistency with chemical structures, functional groups, and expert-defined rules.

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

- **MiniMol** [10]: pretrained to predict various labels spanning quantum chemistry, biological assays, and transcriptomic responses.
- **QIP** [11]: pretrained to approximate electronic structure properties such as orbital interactions.
- **UniMol2** [12, 23]: pretrained with a 1B Pairformer to predict standalone 3D molecular structures. Since no official results are available, we evaluate it under the same settings as ours.
- **MolGPS** [14]: ensembles three 1B-scale pretrained models, i.e., 3B scale, trained to predict bioassay labels or quantum properties.
- **UniQSAR** [24]: advances UniMol [12], which is pretrained to predict 3D structures of standalone molecules, by ensembling with various pretrained models to solve ADMET.

Among these, MolGPS and UniQSAR are ensemble-based methods over multiple pretrained models. For a direct comparison, we combine Boltz2 representations with those from MiniMol (Boltz2$^{\text{Mini.}}$).

3.2 Results

As reported in Table 1, Boltz2 representations exhibit competitive performance. Among non-ensemble methods, Boltz2 achieves the best average performance on 9 of the 22 tasks. When evaluated under ensemble settings, Boltz2$^{\text{Mini.}}$ likewise attains the best results on 9 tasks.Figure 2 summarizes task-level wins, draws, and losses using statistical Welch's t-tests over five seeds. Here, Boltz2 outperforms existing models on the majority of tasks. These results indicate that Boltz2 representations transfer effectively to standalone small-molecule property prediction, despite having no standalone small-molecule pretraining.

In Appendix B, we further isoloate the contribution of co-folding supervision. We train ligand-only and co-folding denoising models with matched architecture, ligand data, initialization, and objective. Co-folding improves all five tasks in this controlled toy experiment (Table 8 in Appendix B). AlphaFold3, ESMFold2, and Protenix-Mini also outperform UniMol2 on the same task subset under our downstream protocol (Table 9 in Appendix B), supporting transfer beyond Boltz2.

In the analysis section (Section 6), we also analyze how Boltz2 provides representations that are complementary to those of existing molecular foundation models beyond the observed improvement from ensembling (Boltz2$^{\text{Mini.}}$). We also find that the Boltz2 performance can be further improved by incorporating relevant protein context, as shown in Table 3 of Section 6.

Intriguingly, both Boltz2 and UniMol2 both underperform on metabolism substrate prediction, suggesting a limitation of current structure-based representations in capturing the local atomistic mechanisms required by these tasks [26]. Bioassay-label leakage may also affect pretrained baselines [14, 27]. We therefore assess molecular and CYP-bound structural overlap separately in Appendix C and discuss local mechanistic understanding as a limitation in Section 7.

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 [6]. Following this approach, we distill Boltz2 representations into state-of-the-art molecular generative models and assess their representation quality via acceleration of generative training.

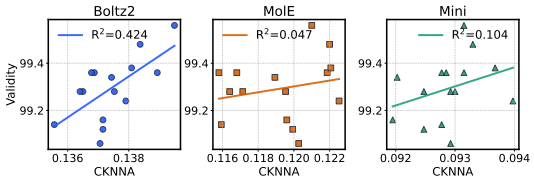


Figure 3: **Representation alignment vs. generation quality.** Each point corresponds to a GruM* run. Stronger alignment with Boltz2 correlates with higher generation quality.

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 [6].

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. Representation alignment with Boltz2 accelerates the training of molecular generative models (Figure 4, 2× faster training) to produce higher-quality molecules compared to the baselines. We report the standard deviation in Table 12 of Appendix D.

Method	Valid ↑	FCD ↓	NSPDK ↓	Novel ↑	Unique ↑	Scaffold ↑	Fragment ↑	SNN ↑
State-of-the-art diffusion-based graph generative models								
GruM	98.65	2.26	0.0015	99.98	99.97	0.5299	–	–
GBD	97.87	2.25	0.0018	–	–	0.5042	–	–
DeFoG	99.22	1.42	0.0008	–	99.99	**0.5903**	–	–
TopBF	99.37	1.39	0.0008	–	–	0.5372	–	–
Marg. SID	99.50	2.01	0.0021	–	–	–	–	–
GruM*	99.36	1.49	0.0007	99.99	**100.00**	0.4923	0.9852	0.3697
*Applying representation alignment-based distillation to GruM**								
MolE	99.51	1.41	0.0006	**100.00**	**100.00**	0.4932	0.9864	0.3737
KPGT	99.37	1.46	0.0006	99.99	**100.00**	0.4800	0.9860	0.3702
Mini.	99.48	1.44	0.0007	99.99	**100.00**	0.4812	0.9856	0.3710
QIP	99.37	1.43	0.0007	**100.00**	**100.00**	0.5242	0.9864	0.3718
UniMol2	99.60	1.40	0.0006	**100.00**	**100.00**	0.4864	0.9867	0.3721
Boltz2	**99.65**	**1.31**	**0.0005**	**100.00**	**100.00**	0.5064	**0.9881**	**0.3766**

To be specific, following prior work [6], we measure the representation alignment between Boltz2 and a state-of-the-art molecular generation model, GruM [28]. Specifically, we adopt CKNNA [29], which quantifies the alignment between representations from molecular foundation models and generative models. Details of CKNNA are provided in Appendix D.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 [30, 31, 32, 33, 28, 34]. 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 [6] to existing diffusion-based molecular generative models, specifically GruM [28]. 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 Boltz2. Implementation details are provided in Appendix D.2.

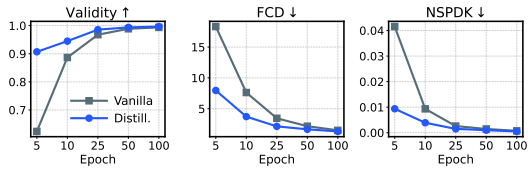


Figure 4: **2× faster training using Boltz2.** Representation alignment with Boltz2 makes generative model training 2× faster.

Baselines. We consider five state-of-the-art diffusion-based graph generative models: original GruM [28], GBD [35], DeFoG [36], TopBF [37], and marginal SID [38]. We also apply representation alignment to GruM* using molecular representations from MolE, KPGT, MiniMol, QIP, and UniMol2 to compare Boltz2-based alignment with alternative representation sources.

Metrics. In this experiment, we evaluate 10,000 generated molecules using eight metrics: chemical validity, Fréchet ChemNet Distance (FCD) [39], the neighborhood subgraph pairwise distance kernel (NSPDK), novelty with respect to the training molecules, uniqueness (Unique), and structural similarity metrics, including scaffold similarity (Scaffold), fragment similarity (Fragment), and similarity to the nearest neighbor (SNN).

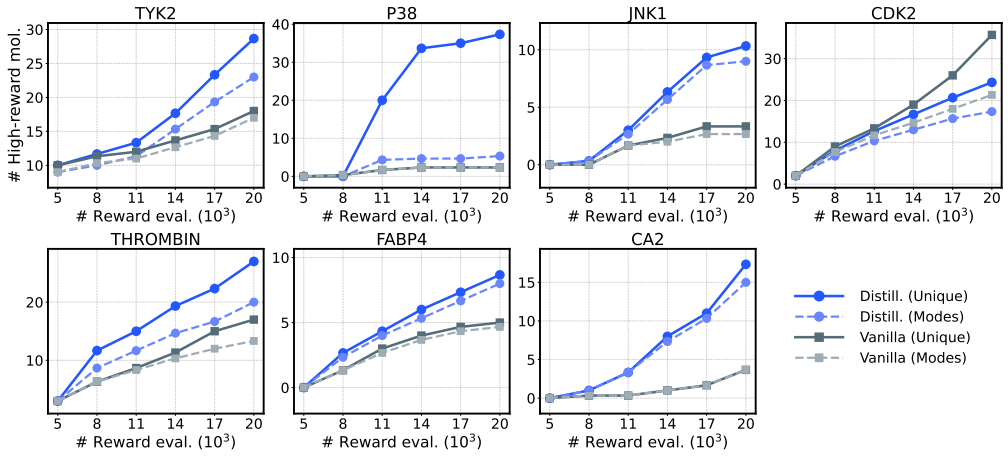


Figure 5: **Results on structure-guided ligand discovery.** The results are averaged over three random seeds. Representation alignment with Boltz2 improves the sample efficiency for discovering high-score molecules that bind to target structures.

4.3 Results

In Figure 4 and Table 2, we report the performance of generative models trained via Boltz2-based distillation. 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, and Boltz2, trained for protein-ligand co-folding, provides stronger structural supervision for molecular generation. Notably, none of the recent state-of-the-art graph diffusion models achieves dominant performance across validity, FCD, and NSPDK simultaneously, with each method excelling on only a subset of these metrics. Boltz2-based distillation, in contrast, improves all three, indicating that representation alignment yields coherent gains rather than trade-offs across metrics.

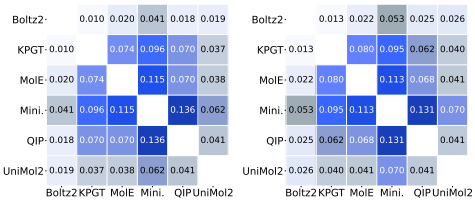
We further evaluate representation alignment on a larger generative model with an eight-layer Pairformer architecture. As demonstrated in Table 13 in Appendix D, Boltz2-based distillation consistently improves generation quality, indicating that the benefit of Boltz2-based distillation is not limited to low-capacity generative models.

5 White-box Structure-guided Ligand Optimization

We further evaluate Boltz2 representations in an online structure-guided ligand discovery setting [20, 40]. The task is formulated as online reinforcement learning, where a molecular generative policy is iteratively updated using predicted binding affinities of generated molecules for a target protein. The objective is to optimize generated molecules with respect to Boltz2 affinity as the reward signal. Here, we treat Boltz2 as a white-box reward teacher rather than only a black-box scalar oracle to better maximize the reward. For each generated molecule, Boltz2 produces both the scalar affinity reward and intermediate ligand representations. We use these representations as dense supervision for the policy, improving credit assignment during policy optimization. To our knowledge, this is the first use of representation alignment [6] to improve online reinforcement learning.

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 [20]: 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.



(a) CKNNA ($K = 5$) (b) CKNNA ($K = 50$)

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$_{(+.03)}$

Figure 6: **Representation complementarity of Boltz2 with existing molecular foundation models. Left:** Boltz2 exhibits relatively weak CKNNA with existing molecular foundation models. **Right:** Combining Boltz2 with a less aligned representation (MolE) yields larger gains on some benchmarks than combining with a more aligned representation (MiniMol).

Implementation details. We use the SynFlowNet-Boltz pipeline introduced in the Boltz2 paper [20]. In this pipeline, the SynFlowNet policy, parameterized by a four-layer graph transformer [41, 40], 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.

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 [20], with details provided in Appendix E.1.

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 [20] 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.

These experiments evaluate how effectively dense representation supervision improves optimization of Boltz2's own affinity score within SynFlowNet's reaction-based search space. To assess transfer to an independent scorer, we additionally re-score the generated TYK2 molecules with GNINA. Distillation increases the mean number of high-scoring unique molecules and modes under both Vina and CNNaffinity (Appendix E.2). Overall, these results show that exposing Boltz2's intermediate ligand representations provides a stronger training signal than using its binding affinity rewards alone in online structure-guided discovery. In Figure 8 of Appendix E.3, we additionally conduct ablation studies on variants of representation alignment.

6 Analyses on Boltz2 Representations

Boltz2 yields a distinct representation space that complements foundation models trained on standalone molecules. We analyze representation alignment between Boltz2 and existing molecular foundation models using the CKNNA described in Section 4. According to the results in the left figure of Figure 6, Boltz2 shows relatively low average alignment with existing molecular foundation models trained on standalone molecular data. This suggests that Boltz2's co-folding learns a representation

space that is distinct from those of existing foundation models, while consistently achieving strong performance across 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 the right table of Figure 6, the Boltz2+MolE ensemble yields larger gains on several benchmarks than Boltz2+MiniMol, despite MiniMol exhibiting stronger performance than MolE. We provide full results in Table 6 of Appendix A.

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

Table 3: **Boltz2 performance with protein-context.** Including protein context improves performance on enzyme inhibition tasks.

Method	Without protein	With protein
Molecular (global) structure inhibits protein		
CYP2C9 V. ↑	0.82	0.83(+.01)
CYP2D6 V. ↑	0.69	0.72(+.03)
CYP3A4 V. ↑	0.85	0.87(+.02)

We report the results in Table 3. 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.Decoy-protein controls in Table 15 show that ligand-only performance is comparable to that with irrelevant protein context, while the relevant CYP context gives the clearest gains on the three inhibition 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 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.

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	32nd	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

7 Conclusion

In this work, we identify protein-ligand co-folding structures as a new source of supervision for atom-level representation learning. We show that Boltz2 representations transfer to small-molecule tasks and achieve strong performance on molecular property prediction, molecular generation, and structure-guided ligand discovery, while complementing representations from existing standalone molecular foundation models. Our results highlight co-folding-based pretraining as an effective strategy for small-molecule representation learning and position Boltz2 as a strong off-the-shelf baseline for small-molecule foundation models. 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. 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.

Limitations. First, while we position Boltz2 as an atom-level foundation model, strong downstream performance does not imply complete atom-level understanding [22]. The shared weakness of Boltz2 and UniMol2 on metabolism substrate prediction points to the need for representations that better capture local atomistic mechanisms [26]. Second, we use the standard frozen-backbone setting through probing and distillation, and leave backbone fine-tuning for future work.

Acknowledgments

This research was supported by the Artificial Intelligence Specialized Foundation Model Project of the National IT Industry Promotion Agency (NIPA), funded by the Ministry of Science and ICT (MSIT) of the Republic of Korea (No. PJT-25-100009). This research was also supported by Institute for Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (RS-2019-II190075, Artificial Intelligence Graduate School Program (KAIST), and RS-2025-02304967, AI Star Fellowship (KAIST)). Additional support was provided by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT (MSIT) (No. RS-2022-NR072184), and a grant of the Korea Machine Learning Ledger Orchestration for Drug Discovery Project (K-MELLODDY), funded by the Ministry of Health & Welfare (MOHW) and Ministry of Science and ICT (MSIT), Republic of Korea (No. RS-2025-21063030, Developing a Highly Multimodal (Drug, Protein, Gene, Cell Imaging, Literature) Foundation Model for ADMET Property Prediction). This research was also supported by the Advanced GPU Utilization Support Program, funded by the Government of the Republic of Korea (Ministry of Science and ICT).

References

[1] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[4] 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.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[6] 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 *International Conference on Learning Representations*, 2025.

[7] Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025.

[8] Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf H Roohani, Jure Leskovec, Connor W. Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. 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.

[9] Oscar Méndez-Lucio, Christos A Nicolaou, and Berton Earnshaw. Mole: a foundation model for molecular graphs using disentangled attention. *Nature Communications*, 15(1):9431, 2024.

[10] Kerstin Klaser, Blazej Banaszewski, Samuel Maddrell-Mander, Callum McLean, Luis Müller, Ali Parviz, Shenyang Huang, and Andrew W Fitzgibbon. Minimol: A parameter-efficient foundation model for molecular learning. In *ICML 2024 Workshop on Efficient and Accessible Foundation Models for Biological Discovery*, 2024.

[11] Jungwoo Kim, Woojae Chang, Hyunjun Ji, and InSuk Joung. Quantum-informed molecular representation learning enhancing admet property prediction. *Journal of Chemical Information and Modeling*, 64(13):5028–5040, 2024.

[12] 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 *International Conference on Learning Representations*, 2023.

[13] Han Li, Ruotian Zhang, Yaosen Min, Dacheng Ma, Dan Zhao, and Jianyang Zeng. A knowledge-guided pre-training framework for improving molecular representation learning. *Nature Communications*, 14(1):7568, 2023.

[14] Maciej Sypetkowski, Frederik Wenkel, Farimah Poursafaei, Nia Dickson, Karush Suri, Philip Fradkin, and Dominique Beaini. On the scalability of gnns for molecular graphs. In *Advances in Neural Information Processing Systems*, 2024.

[15] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[16] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18123–18133, 2022.

[17] Stephen K Burley, Charmi Bhikadiya, Chunxiao Bi, Sebastian Bittrich, Li Chen, Gregg V Crichlow, Cole H Christie, Kenneth Dalenberg, Luigi Di Costanzo, Jose M Duarte, et al. Rcsb protein data bank: powerful new tools for exploring 3d structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic acids research*, 49(D1):D437–D451, 2021.

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

[19] Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Noah Getz, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Liam Atkinson, Tally Portnoi, Itamar Chinn, Jacob Silterra, Tommi Jaakkola, and Regina Barzilay. Boltz-1: Democratizing biomolecular interaction modeling. *bioRxiv*, 2024.

[20] Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, David Kwabi-Addo, Dominique Beaini, Tommi Jaakkola, and Regina Barzilay. Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv*, 2025.

[21] Kolja Stahl, Andrea Graziadei, Therese Dau, Oliver Brock, and Juri Rappsilber. Protein structure prediction with in-cell photo-crosslinking mass spectrometry and deep learning. *Nature Biotechnology*, 41(12):1810–1819, 2023.

[22] Z Faidon Brotzakis, Shengyu Zhang, Mhd Hussein Murtada, and Michele Vendruscolo. Alphafold prediction of structural ensembles of disordered proteins. *Nature Communications*, 16(1):1632, 2025.

[23] Xiaohong Ji, Zhen Wang, Zhifeng Gao, Hang Zheng, Linfeng Zhang, Guolin Ke, and Weinan E. Exploring molecular pretraining model at scale. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[24] Zhifeng Gao, Xiaohong Ji, Guojiang Zhao, Hongshuai Wang, Hang Zheng, Guolin Ke, and Linfeng Zhang. Uni-qsar: an auto-ml tool for molecular property prediction. *arXiv preprint arXiv:2304.12239*, 2023.

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

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

[27] Ihor Koleiev, Roman Stratiichuk, Nazar Shevchuk, Mykola Melnychenko, Oleksiy Nyporko, Daniil Todoryshyn, Vladyslav Husak, Sergii Starosyla, Semen Yesylevskyy, and Alan Nafiiev. Critical assessment of ml models for admet prediction in tdc leaderboards. *bioRxiv*, pages 2026–02, 2026.

[28] Jaehyeong Jo, Dongki Kim, and Sung Ju Hwang. Graph generation with diffusion mixture. In *International Conference on Machine Learning*, 2024.

[29] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

[30] Jaehyeong Jo, Seul Lee, and Sung Ju Hwang. Score-based generative modeling of graphs via the system of stochastic differential equations. In *International Conference on Machine Learning*, 2022.

[31] Clement Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. Digress: Discrete denoising diffusion for graph generation. In *International Conference on Learning Representations*, 2023.

[32] Yunhui Jang, Seul Lee, and Sungsoo Ahn. A simple and scalable representation for graph generation. In *International Conference on Learning Representations*, 2024.

[33] Lingkai Kong, Jiaming Cui, Haotian Sun, Yuchen Zhuang, B Aditya Prakash, and Chao Zhang. Autoregressive diffusion model for graph generation. In *International Conference on Machine Learning*, 2023.

[34] Hyunjin Seo, Taewon Kim, Sihyun Yu, and SungSoo Ahn. Learning flexible forward trajectories for masked molecular diffusion. *arXiv preprint arXiv:2505.16790*, 2025.

[35] Xinyang Liu, Yilin He, Bo Chen, and Mingyuan Zhou. Advancing graph generation through beta diffusion. In *The Thirteenth International Conference on Learning Representations*.

[36] Yiming Qin, Manuel Madeira, Dorina Thanou, and Pascal Frossard. Defog: Discrete flow matching for graph generation. In *International Conference on Machine Learning*, pages 50269–50326. PMLR, 2025.

[37] Yida Xiong, Jiameng Chen, Kun Li, Hongzhi Zhang, Xiantao Cai, Jia Wu, and Wenbin Hu. Transport-coupled bayesian flows for molecular graph generation, 2026.

[38] Yoann Boget. Simple and critical iterative denoising: A recasting of discrete diffusion in graph generation. In *International Conference on Machine Learning*, pages 4713–4736. PMLR, 2025.

[39] Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Gunter Klambauer. Fréchet chemnet distance: a metric for generative models for molecules in drug discovery. *Journal of chemical information and modeling*, 58(9):1736–1741, 2018.

[40] Miruna Cretu, Charles Harris, Ilia Igashov, Arne Schneuing, Marwin Segler, Bruno Correia, Julien Roy, Emmanuel Bengio, and Pietro Lio. Synflownet: Design of diverse and novel molecules with synthesis constraints. In *International Conference on Learning Representations*, 2025.

[41] Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J Kim. Graph transformer networks. In *Advances in Neural Information Processing Systems*, 2019.

[42] Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda. Attentive statistics pooling for deep speaker embedding. *arXiv preprint arXiv:1803.10963*, 2018.

[43] Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow network based generative models for non-iterative diverse candidate generation. *Advances in neural information processing systems*, 34:27381–27394, 2021.

[44] Nikolay Malkin, Moksh Jain, Emmanuel Bengio, Chen Sun, and Yoshua Bengio. Trajectory balance: Improved credit assignment in gflownets. *Advances in Neural Information Processing Systems*, 35:5955–5967, 2022.

[45] Yoshua Bengio, Salem Lahlou, Tristan Deleu, Edward J Hu, Mo Tiwari, and Emmanuel Bengio. Gflownet foundations. *Journal of Machine Learning Research*, 24(210):1–55, 2023.

[46] Minsu Kim, Taeyoung Yun, Emmanuel Bengio, Dinghuai Zhang, Yoshua Bengio, Sungsoo Ahn, and Jinkyoo Park. Local search gflownets. In *International Conference on Learning Representations*, 2024.

[47] Hyosoon Jang, Minsu Kim, and Sungsoo Ahn. Learning energy decompositions for partial inference in gflownets. In *International Conference on Learning Representations*, 2024.

[48] Seonghwan Seo, Minsu Kim, Tony Shen, Martin Ester, Jinkyoo Park, Sungsoo Ahn, and Woo Youn Kim. Generative flows on synthetic pathway for drug design. In *The Thirteenth International Conference on Learning Representations*.

[49] Hyosoon Jang, Yunhui Jang, Minsu Kim, Jinkyoo Park, and Sungsoo Ahn. Pessimistic backward policy for gflownets. In *Advances in Neural Information Processing Systems*, 2024.

A ADMET Benchmark Details and Full Results

A.1 TDC ADMET Benchmark

Table 5: **Data statistics and evaluation metrics of the TDC ADMET benchmark.** TDC ADMET benchmark is distributed through TDC, whose codebase is MIT-licensed.

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)	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 Life	Regression	Spearman	1,039
Clearance H. (Hepatocyte)	Regression	Spearman	1,102
Clearance M. (Microsome)	Regression	Spearman	1,102
LD50 (Acute Toxicity)	Regression	MAE	7,385
hERG (hERG Blockade)	Classification	AUROC	648
AMES (Mutagenicity)	Classification	AUROC	6,512
DILI (Drug-Induced Liver Injury)	Classification	AUROC	475

A.2 Implementation Details

Representation extraction. In this section, we describe the pooling strategy for ADMET property prediction. Note that Boltz does not include a pooling layer and is not trained on pooled representations to extract informative fixed-length representations for downstream tasks, unlike existing small-molecule foundation models. A naive choice would be simple global mean pooling over all atom-pair entries, but this risks ignoring local features that are critical for tasks such as substrate prediction. We therefore apply a hybrid pooling strategy to capture diverse structural statistics of pair representations. Note that we apply the identical pooling strategy to UniMol2 for a fair comparison.

For Boltz2 and UniMol2, we concatenate pair representations from the $\{16, 32, 48, 64\}$-th layers, corresponding to the $1/4$, $2/4$, $3/4$, and final layers of the 64-layer Pairformer. We then concatenate pooled representations over diagonal entries, bonded atom-pair entries, and all entries in the pair representation matrix. Specifically, we apply statistic pooling [42].

Importantly, even with reduced pooling, Boltz representations outperform existing molecular representation models (Appendix G). However, our hybrid pooling does not recover performance on substrate prediction tasks, suggesting that the substrate gap is unlikely to be a pooling artifact and instead reflects a more fundamental limitation of structure-based representations on local atomistic mechanisms (see limitations in Section 7).

Probing network. The probing network is implemented as a multi-layer perceptron (MLP), following prior works [10, 14]. The hidden dimension and the number of layers of the MLP are selected from $\{512, 1024, 2048\}$ and $\{2, 3, 4\}$, respectively. The hidden representation at each layer is concatenated with the input, following the model design in prior work [10]. 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 batch size 32. Early stopping is applied with a patience of 25 epochs, and a cosine learning-rate scheduler is used. Hyperparameters are selected based on validation performance.

A.3 Full Results on The TDC ADMET Benchmark

Table 6: **Full results of representation ensembling on the ADMET benchmark.**

Dataset	Boltz2	Boltz2$^{\text{Mini.}}$	Boltz2$^{\text{MolE}}$
Absorption			
Caco2 ↓	0.30 ± .004	0.30 ± .009	0.30 ± .005
HIA ↑	0.99 ± .001	0.99 ± .001	0.99 ± .001
Pgp ↑	0.93 ± .002	0.93 ± .002	0.93 ± .003
Bioavailability ↑	0.75 ± .009	0.77 ± .006	0.72 ± .009
Lipophilicity ↓	0.45 ± .005	0.41 ± .003	0.43 ± .007
Solubility ↓	0.66 ± .013	0.64 ± .017	0.65 ± .003
Distribution			
BBB ↑	0.93 ± .003	0.94 ± .001	0.92 ± .005
PPBR ↓	7.65 ± .063	7.59 ± .090	7.34 ± .074
VDss ↑	0.74 ± .010	0.75 ± .007	0.75 ± .008
Metabolism			
CYP2C9 V. ↑	0.82 ± .003	0.86 ± .003	0.88 ± .002
CYP2D6 V. ↑	0.69 ± .002	0.72 ± .004	0.77 ± .001
CYP3A4 V. ↑	0.85 ± .001	0.88 ± .002	0.89 ± .002
CYP2C9 S. ↑	0.36 ± .015	0.36 ± .029	0.37 ± .021
CYP2D6 S. ↑	0.52 ± .012	0.51 ± .021	0.55 ± .024
CYP3A4 S. ↑	0.62 ± .011	0.60 ± .012	0.64 ± .013
Excretion			
Half Life ↑	0.62 ± .010	0.65 ± .015	0.66 ± .030
Clearance H. ↑	0.62 ± .009	0.60 ± .018	0.57 ± .013
Clearance M. ↑	0.61 ± .008	0.65 ± .011	0.65 ± .008
Toxicity			
LD50 ↓	0.40 ± .004	0.40 ± .006	0.40 ± .003
hERG ↑	0.86 ± .009	0.86 ± .008	0.83 ± .008
AMES ↑	0.91 ± .002	0.91 ± .003	0.92 ± .001
DILI ↑	0.89 ± .008	0.87 ± .009	0.87 ± .009

We report the full results of representation ensembling in Supplementary Table 6. Boltz$^{\text{MolE}}$ outperforms Boltz$^{\text{Mini.}}$ on some benchmarks, although MiniMol shows stronger standalone ADMET prediction performance than MolE.

B Comparisons of Pretraining and Model Scale

B.1 Comparison with MolGPS Backbones

Table 7: **Boltz2 and individual MolGPS backbones (different parameter scales).** Boltz2 has approximately 500M parameters. The MolGPS backbones (MPNN++, Transformer, GPS++) are each approximately 1B parameters. We use numbers from [14]. **Bold** indicates the best performance.

Dataset	MPNN++	Transformer	GPS++	Boltz2
Absorption				
Caco2 ↓	0.58	0.36	0.40	**0.30**
HIA ↑	0.96	0.93	0.96	**0.99**
Pgp ↑	**0.94**	0.92	**0.94**	0.93
Bioavailability ↑	0.70	0.70	0.69	**0.75**
Lipophilicity ↓	0.55	0.52	0.49	**0.45**
Solubility ↓	0.86	0.75	0.78	**0.66**
Distribution				
BBB ↑	0.90	0.91	0.90	**0.93**
PPBR ↓	9.09	13.36	10.91	**7.65**
VDss ↑	0.60	0.65	0.65	**0.74**
Metabolism				
CYP2C9 V. ↑	**0.85**	0.82	0.82	0.82
CYP2D6 V. ↑	0.70	**0.71**	0.69	0.69
CYP3A4 V. ↑	**0.88**	**0.88**	**0.88**	0.85
CYP2C9 S. ↑	0.37	**0.41**	0.38	0.36
CYP2D6 S. ↑	**0.68**	0.66	0.66	0.52
CYP3A4 S. ↑	0.65	**0.72**	0.69	0.62
Excretion				
Half Life ↑	0.44	0.56	0.53	**0.62**
Clearance H. ↑	0.39	0.31	0.36	**0.62**
Clearance M. ↑	0.61	0.50	**0.62**	0.61
Toxicity				
LD50 ↓	0.72	0.67	0.69	**0.40**
hERG ↑	0.83	0.81	0.79	**0.86**
AMES ↑	0.81	0.83	0.83	**0.91**
DILI ↑	0.85	**0.90**	**0.90**	0.89

B.2 Controlled Toy Experiment on Co-Folding Supervision

Table 8: **Controlled toy-scale comparison of denoising objectives.** Results are mean ± standard deviation. Arrows indicate the favorable direction.

Task	Metric	Ligand-only	Co-folding
Solubility	MAE ↓	0.821 ± 0.006	$\mathbf{0.799 \pm 0.003}$
BBB	AUROC ↑	0.850 ± 0.005	$\mathbf{0.892 \pm 0.009}$
CYP2C9 Veith	AUPRC ↑	0.747 ± 0.002	$\mathbf{0.760 \pm 0.002}$
Half Life	Spearman ↑	0.563 ± 0.011	$\mathbf{0.580 \pm 0.019}$
LD50	MAE ↓	0.622 ± 0.004	$\mathbf{0.589 \pm 0.005}$

To isolate co-folding supervision more closely, we train two models on approximately 100,000 protein-ligand entries from RCSB. Both use the UniMol2 codebase, the same ligand examples, feature initialization, architecture, and denoising objective. The ligand-only model sees only the ligand, while the co-folding model sees the full complex. The Pairformer has 12 layers, 128-dimensional single and 64-dimensional pair representations. Both models are trained for 10 epochs with batch size 64 and learning rate 3×10^{-4}. The co-folding variant uses spatial cropping to at most 128 tokens and Cα coordinates. Attention biases encode molecular modality and relative residue-index differences. As shown in Table 8, the co-folding model improves all five tasks in the subset used for the ablation. This controlled comparison supports the benefit of co-folding supervision at toy scale, complementing the comparisons between pretrained models.

B.3 Comparison with Other Co-Folding Models

Table 9: **Transfer from different co-folding models.** All models are evaluated on the five-task subset with the same downstream protocol. Arrows indicate the favorable direction.

Task	Metric	UniMol2	Protenix-Mini	ESMFold2	AF3	Boltz2
Solubility	MAE ↓	0.81	0.70	0.68	0.67	**0.66**
BBB	AUROC ↑	0.86	0.90	0.92	**0.93**	**0.93**
CYP2C9 Veith	AUPRC ↑	0.77	0.80	**0.83**	0.82	0.82
Half Life	Spearman ↑	0.48	0.56	0.60	**0.63**	0.62
LD50	MAE ↓	0.50	0.43	0.43	0.42	**0.40**

We also apply the same downstream evaluation protocol to three further co-folding models. Table 9 shows that no one model dominates all five tasks. Protenix-Mini is weaker than the larger co-folding models on this subset but exceeds UniMol2 on each reported task. Together with the controlled toy experiment, these results support the usefulness of co-folding representations beyond Boltz2's particular training recipe.

C Pretraining Overlap and Data Leakage

Table 10: **Pretraining-corpus overlap with TDC test molecules.** Exact / connectivity-block matches (%), averaged by category. Task counts are in parentheses.

Category	Boltz-CCD	PCBA (MiniMol, MolGPS)	ChEMBL (KPGT)
Absorption (6)	19.5 / 29.5	38.6 / 50.0	71.4 / 82.9
Distribution (3)	10.0 / 26.3	27.1 / 43.3	71.1 / 95.4
Metabolism (6)	23.1 / 28.8	77.2 / 89.0	87.4 / 97.4
Excretion (3)	24.3 / 30.7	35.8 / 41.9	99.9 / 99.9
Toxicity (4)	14.0 / 28.4	34.0 / 50.6	51.7 / 70.1

We compare pretraining corpora with TDC test molecules using exact InChIKey matches and connectivity-block matches (Table 10). Boltz-CCD has the lowest overlap in every category under both criteria. Boltz2 is pretrained on protein-ligand structures without the downstream TDC assay labels. Molecular overlap therefore has a different implication from exposure to the corresponding assay labels during bioassay pretraining.

Table 11: **Overlap with CYP-bound ligands.** Test molecules found in RCSB structures of the corresponding enzyme.

Task	Test molecules	Exact	Connectivity block
CYP2C9 Veith	2418	0	0
CYP2D6 Veith	2626	0	0
CYP3A4 Veith	2466	0	0
CYP2C9 Substrate	134	0	0
CYP2D6 Substrate	133	0	1
CYP3A4 Substrate	134	0	0

We also scan RCSB structures of CYP2C9, CYP2D6, and CYP3A4 for test molecules bound to the corresponding enzyme. Across the six metabolism tasks, we find no exact matches and only one connectivity-block match (Table 11). Almost none of these test molecules were therefore observed bound to their relevant CYP.

The overlap is not limited to metabolism. In particular, ChEMBL overlaps with almost all test molecules in excretion, where Boltz2 remains competitive despite its smaller overlap. Thus, molecular overlap alone does not explain the relative performance across ADMET categories. It also does not establish that a model has seen the downstream assay labels. Assessing label leakage requires examining the assays and labels used in pretraining, beyond the molecular identities measured here.

For metabolism substrate prediction, possible leakage in bioassay-pretrained baselines and limitations of structure-based representations remain separate considerations. The weakness shared by Boltz2 and UniMol2 suggests that learning 3D structures may not fully capture the local mechanisms required for substrate prediction [26]. Our overlap analysis provides context for this comparison, but does not determine how much of the performance gap is attributable to either factor.

D Molecular Generation Details and Full Results

D.1 CKNNA Metric

We measure alignment using Centered Kernel Nearest-Neighbor Alignment metric (CKNNA) [29], 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:

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

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. CKNNA is then computed as

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

D.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 two-layer MLPs. Single representations are initialized from atom features and the time conditioning signal.

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, scaling, and noise scheduling, follow the default GruM configuration [28].

Representation alignment. We adopt a representation alignment-based distillation objective to train molecular generative models, specifically GruM, with auxiliary supervision from Boltz2 representations. Given a noisy molecule m_t and its corresponding clean molecule m, we define the training objective 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 denoising loss of GruM, $s_\theta(m_t)$ is the denoising prediction, $h_\theta(m_t)$ denotes the output of the distillation network given the hidden representation of the generative model, and $f(m)$ denotes the frozen Boltz2 representation of the clean molecule. For distillation, we introduce a lightweight distillation network, 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.

To be more specific, 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.

D.3 Full Results on Unconditional Graph Generation

Table 12: **Full results on unconditional molecular generation.** GruM* denotes GruM parameterized with a Pairformer. Mean ± standard deviation over three random seeds is reported. Entries with ±x.xx denote that the standard deviation is not reported in the original paper. **Bold** numbers indicate the best performance. Representation alignment with Boltz2 accelerates the training of molecular generative models to produce higher-quality molecules compared to the baselines.

Method	Valid ↑	FCD ↓	NSPDK ↓	Novel ↑	Unique ↑	Scaffold ↑	Fragment ↑	SNN ↑
State-of-the-art diffusion-based graph generative models								
GruM	$98.65_{\pm 0.25}$	$2.26_{\pm 0.08}$	$0.0015_{\pm 0.0003}$	$99.98_{\pm 0.02}$	$99.97_{\pm 0.03}$	$0.5299_{\pm 0.0441}$	–	–
GBD	$97.87_{\pm \text{x.xx}}$	$2.25_{\pm \text{x.xx}}$	$0.0018_{\pm \text{x.xxxx}}$	–	–	$0.5042_{\pm \text{x.xxxx}}$	–	–
DeFoG	$99.22_{\pm 0.08}$	$1.42_{\pm 0.02}$	$0.0008_{\pm 0.0001}$	–	$99.99_{\pm 0.01}$	$\mathbf{0.5903}_{\pm 0.0099}$	–	–
TopBF	$99.37_{\pm \text{x.xx}}$	$1.39_{\pm \text{x.xx}}$	$0.0008_{\pm \text{x.xxxx}}$	–	–	$0.5372_{\pm \text{x.xxxx}}$	–	–
Marg. SID	$99.50_{\pm \text{x.xx}}$	$2.01_{\pm \text{x.xx}}$	$0.0021_{\pm \text{x.xxxx}}$	–	–	–	–	–
GruM*	$99.36_{\pm 0.09}$	$1.50_{\pm 0.07}$	$0.0007_{\pm 0.0001}$	$99.99_{\pm 0.01}$	$\mathbf{100.00}_{\pm 0.00}$	$0.4923_{\pm 0.0074}$	$0.9852_{\pm 0.0017}$	$0.3697_{\pm 0.0028}$
*Applying representation alignment-based distillation to GruM**								
+UniMol2	$99.60_{\pm 0.07}$	$1.42_{\pm 0.03}$	$0.0006_{\pm 0.0000}$	$\mathbf{100.00}_{\pm 0.00}$	$\mathbf{100.00}_{\pm 0.00}$	$0.4864_{\pm 0.0133}$	$0.9867_{\pm 0.0007}$	$0.3721_{\pm 0.0002}$
+MolE	$99.55_{\pm 0.02}$	$1.40_{\pm 0.01}$	$0.0006_{\pm 0.0000}$	$\mathbf{100.00}_{\pm 0.00}$	$\mathbf{100.00}_{\pm 0.00}$	$0.4932_{\pm 0.0041}$	$0.9864_{\pm 0.0004}$	$0.3737_{\pm 0.0007}$
+KPGT	$99.37_{\pm 0.05}$	$1.46_{\pm 0.05}$	$0.0006_{\pm 0.0000}$	$99.99_{\pm 0.01}$	$99.99_{\pm 0.01}$	$0.4800_{\pm 0.0299}$	$0.9860_{\pm 0.0009}$	$0.3702_{\pm 0.0020}$
+Mini.	$99.48_{\pm 0.09}$	$1.45_{\pm 0.04}$	$0.0007_{\pm 0.0001}$	$99.99_{\pm 0.01}$	$\mathbf{100.00}_{\pm 0.00}$	$0.4812_{\pm 0.0234}$	$0.9856_{\pm 0.0009}$	$0.3710_{\pm 0.0023}$
+QIP	$99.37_{\pm 0.09}$	$1.43_{\pm 0.05}$	$0.0007_{\pm 0.0001}$	$\mathbf{100.00}_{\pm 0.00}$	$\mathbf{100.00}_{\pm 0.00}$	$0.5242_{\pm 0.0071}$	$0.9864_{\pm 0.0003}$	$0.3718_{\pm 0.0015}$
+Boltz2	$\mathbf{99.65}_{\pm 0.11}$	$\mathbf{1.31}_{\pm 0.01}$	$\mathbf{0.0005}_{\pm 0.0000}$	$\mathbf{100.00}_{\pm 0.00}$	$\mathbf{100.00}_{\pm 0.00}$	$0.5064_{\pm 0.0187}$	$\mathbf{0.9881}_{\pm 0.0003}$	$\mathbf{0.3766}_{\pm 0.0014}$

Table 13: **Distillation into a large-scale model.** Boltz2 representation improves molecular generation performance on a large-scale backbone.

Method	Validity ↑	FCD ↓	NSPDK ↓
PairFormer	99.41	1.49	0.0007
Align. w/ Boltz2	99.65	1.31	0.0005
Pairformer$^{\text{large}}$	99.56	1.35	0.0006
w/ Boltz2	**99.76**	**1.28**	**0.0005**

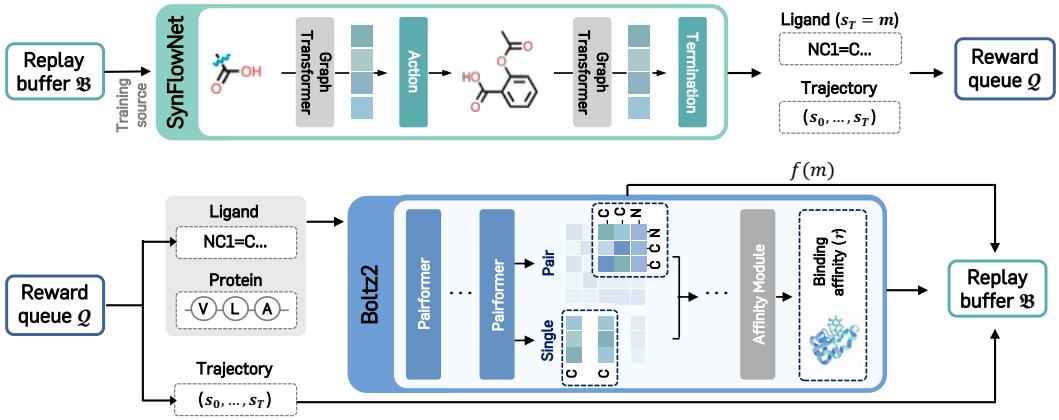


Figure 7: **Online SynFlowNet-Boltz ligand discovery pipeline.** The top and bottom panels illustrate the processes of molecular generation with SynFlowNet and evaluation with Boltz2, respectively. Both processes are asynchronous.

E Details in SynFlowNet-Boltz

E.1 Implementation Details

We use SynFlowNet-Boltz ligand discovery pipeline introduced in prior work [20, 40]. 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.

SynFlowNet. We use SynFlowNet [40], a widely studied GFlowNet framework [43, 44, 45, 46, 47, 48], for synthesizable molecular generation, which defines a reaction-based action space.

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 [41], and we do not modify the original implementation [20, 40]. 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 [40], 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 [40, 49].

Reward computation. For each generated terminal molecule $s_T = m$, Boltz2 computes a reward r (a screening score) 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 $s_T = m$:

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

Algorithm 1 Learning SynFlowNet with Representation Alignment

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

Algorithm 2 Asynchronous Boltz2 Worker

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

where $h_\theta(m)$ denotes the output of the distillation network given the hidden representation of the policy network, and $f(m)$ denotes the frozen Boltz2 representation. 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 $\lambda = 10$.

To be specific, 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 $f(m)$.

As an ablation study, we also study intermediate-state distillation as an ablation in Appendix E.3. 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 [20]. 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 [20].

Table 14: **Independent GNINA re-scoring for TYK2.** Vina cutoffs are in kcal/mol and CNNaffinity cutoffs are in pK_d. Results are three-seed mean $\pm$ standard deviation. "Modes" counts distinct strong-scoring modes. "Unique" counts distinct strong-scoring molecules.

		Modes		**Unique molecules**	
Scorer	**Cutoff**	**Baseline**	**Distilled**	**Baseline**	**Distilled**
Vina	< -9.5	27.0 ± 1.7	$\mathbf{37.3 \pm 5.1}$	35.7 ± 7.0	$\mathbf{43.0 \pm 9.5}$
Vina	< -10.0	8.7 ± 1.2	$\mathbf{10.7 \pm 4.7}$	9.0 ± 1.0	$\mathbf{10.7 \pm 4.7}$
CNNaffinity	> 8.0	9.0 ± 2.6	$\mathbf{17.0 \pm 5.0}$	10.7 ± 4.0	$\mathbf{22.7 \pm 11.6}$
CNNaffinity	> 8.5	0.7 ± 1.2	$\mathbf{2.7 \pm 2.9}$	0.7 ± 1.2	$\mathbf{3.0 \pm 3.5}$

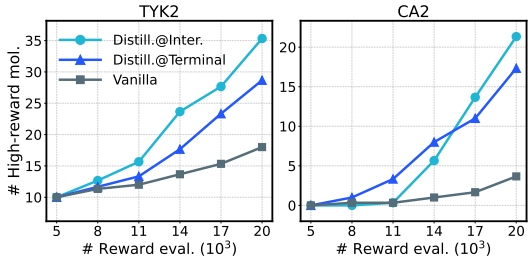


Figure 8: Representation alignment on intermediate molecules.

E.2 Independent Docking Evaluation on TYK2

The optimization experiments in Section 5 use Boltz2's affinity score as the reward and its intermediate representations as auxiliary supervision. To test whether the generated molecules also score well under a different model, we re-evaluate the TYK2 candidates with GNINA using two scoring functions, Vina and CNNaffinity. Table 14 reports the numbers of unique molecules and modes above two strong-binding thresholds for each scorer, averaged over three seeds. Distillation improves the mean counts at all four thresholds, although the strictest thresholds yield few candidates and substantial variability. This evaluation provides independent computational support.

E.3 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. Specifically, we consider the following alignment loss:

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

where $\mathcal{I}(s_t)$ 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. An intermediate molecule s_t may lead to multiple terminal molecules through different trajectory actions, resulting in multiple possible alignment targets. To address this, we restrict alignment to representations of terminal molecules in the top 10% by reward. We report the results of this Figure 8

F Effect of Protein Context

Table 15: **Effect of decoy protein context.** α-syn. denotes α-synuclein. Relevant denotes the corresponding CYP enzyme and is not applicable to solubility.

Task	Metric	Ligand only	+poly-X	+Lysozyme	+α-syn.	+Relevant
CYP2C9 Veith	AUPRC ↑	0.821	0.820	0.825	0.818	**0.834**
CYP2D6 Veith	AUPRC ↑	0.693	0.690	0.708	0.709	**0.722**
CYP3A4 Veith	AUPRC ↑	0.854	0.850	0.862	0.855	**0.873**
Solubility	MAE ↓	0.661	0.667	0.650	**0.640**	n/a

We extract the same ligand representations with an artificial poly-X chain, folded but irrelevant lysozyme, or intrinsically disordered α-synuclein as context. Table 15 shows generally similar results to ligand-only extraction. The corresponding CYP enzyme gives the highest AUPRC on all three inhibition tasks. An unrelated context can also help: α-synuclein improves solubility MAE from 0.661 to 0.640. These results support ligand-only extraction as a useful default and motivate the exploration of protein context as a task-specific prompt (Section 7).

G Ablations on Pooling

Table 16: **Ablations on pooling strategy. Bold** indicates the best performance. Boltz2 consistently shows strong performance compared to the baselines without advanced pooling strategy.

	Baselines				(Ours) Pooling w/o			
Tasks	**MolE**	**KGPT**	**Mini.**	**QIP**	**Bond**	**Diag**	**Std**	∅
Solubility ↓	0.79	0.71	0.74	0.70	0.66	0.66	**0.64**	0.66
BBB ↑	0.90	0.91	0.92	0.90	0.92	0.92	0.92	**0.93**
CYP2C9 V. ↑	0.80	0.80	**0.82**	0.78	**0.82**	**0.82**	**0.82**	**0.82**
Half Life ↑	0.55	0.53	0.50	0.53	0.61	0.63	**0.64**	0.62
LD50 ↓	0.82	0.55	0.59	0.56	**0.40**	**0.40**	0.41	**0.40**

We conduct an ablation study on pooling strategy in Supplementary Table 16. Our hybrid pooling captures diverse structural statistics of pair representations, it yields only marginal improvements over simpler pooling schemes. Importantly, even with reduced pooling, Boltz representations generally outperform existing molecular representation models.

H Boltz2 Modification

We modify the official Boltz2 code [20], which is released under the MIT license, 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.

I Broader Impact

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.