What Ails Generative Structure-based Drug Design?

Rafał Karczewski [1] Samuel Kaski [1, 2] Markus Heinonen [1] Vikas Garg, [1, 3]

[1]Aalto University [2]University of Manchester [3]YaiYai Ltd.

Overview

- SBDD models are growing in size, yet performance remains suboptimal;
- Overemphasis on expressivity via large GNNs might be misguided—these models face inherent limits (shown empirically and theoretically);
- We advocate focusing on **generalization instead of expressivity**;
- **Our model is 100x smaller, 100–1000x faster**, and matches or improves on SOTA;
- **Binding is not the only goal**: synthesizability, drug-likeness, and toxicity matter too;
- Our approach frees resources for more precise steps, like molecular simulations.

What is structure-based drug design?

- SBBD - finding molecules likely to attach to a given protein;
- Solved by generative models learning from available protein-molecule pairs;
- Chemical software evaluates the quality of generated molecules, and strength of binding.

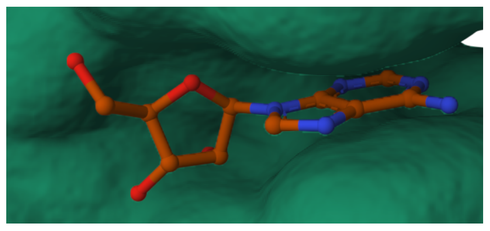


Figure 1. A molecule (ligand) in a protein pocket.

What affects binding?

For G^M molecule, and G^P protein, binding is estimated by:

$$\beta(G^M, G^P) = \min_T f(T(G^M), G^P),$$

where T are 3D modifications of G^M and f a physics based scoring function. To study factors impacting binding predictions, we decompose the molecule into:

$$\underbrace{G^M}_{\text{molecule}} = (\underbrace{\mathcal{U}^M}_{\text{topology}}, \underbrace{\mathbf{a}^M}_{\text{atom types}}, \underbrace{\mathbf{s}^M}_{\text{3D coords}}),$$

where topology is the 2D unlabeled molecular graph. We measure how each component impacts binding predictions. $\mathbf{s}^M$ has a negligible effect (redocking). Surprisingly, we find:

$$\mathrm{Corr}(\beta(G^M, G^P), \beta(\mathbf{\Pi}(G^M), G^P)) = 0.85,$$

where $\mathbf{\Pi}$ randomly changes atom types. **Binding can be predicted from topology alone!**

SimpleSBDD: topology first, atoms later

Idea: surrogate model predicting binding solely from topology

$$g_\theta(\mathcal{U}^M, G^P) \approx \beta(G^M, G^P),$$

which allows generating molecules in a 2-step procedure:

- First generate topology $\mathcal{U}^M$ **optimized for binding** using g_θ;
- Then atom types $\mathbf{a}^M$ **optimized for other properties** (pretrained model).

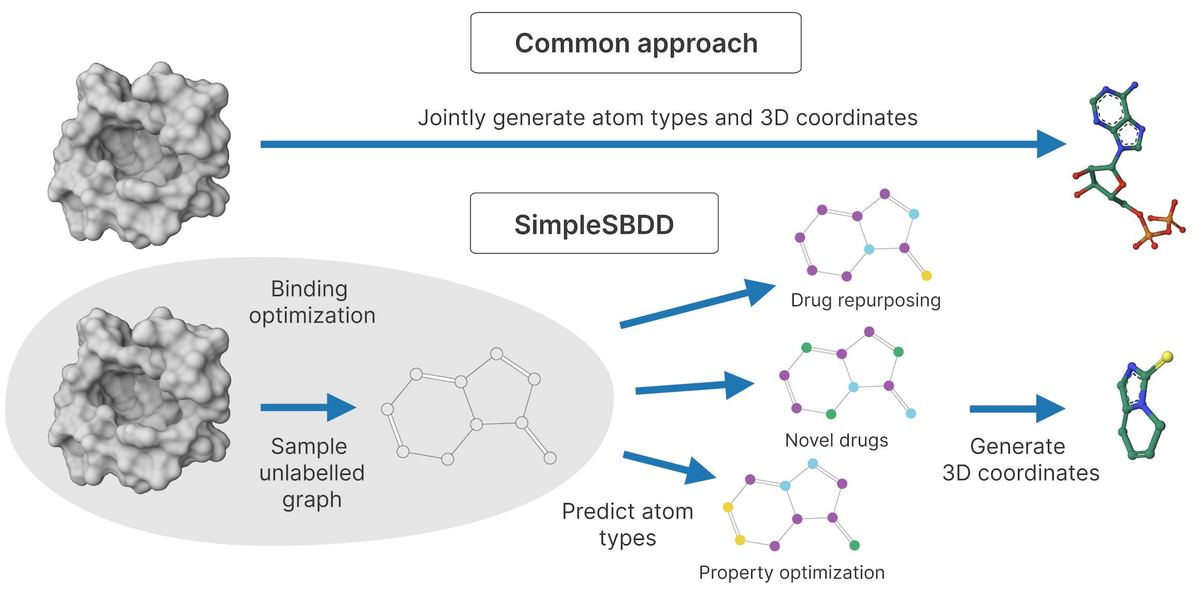
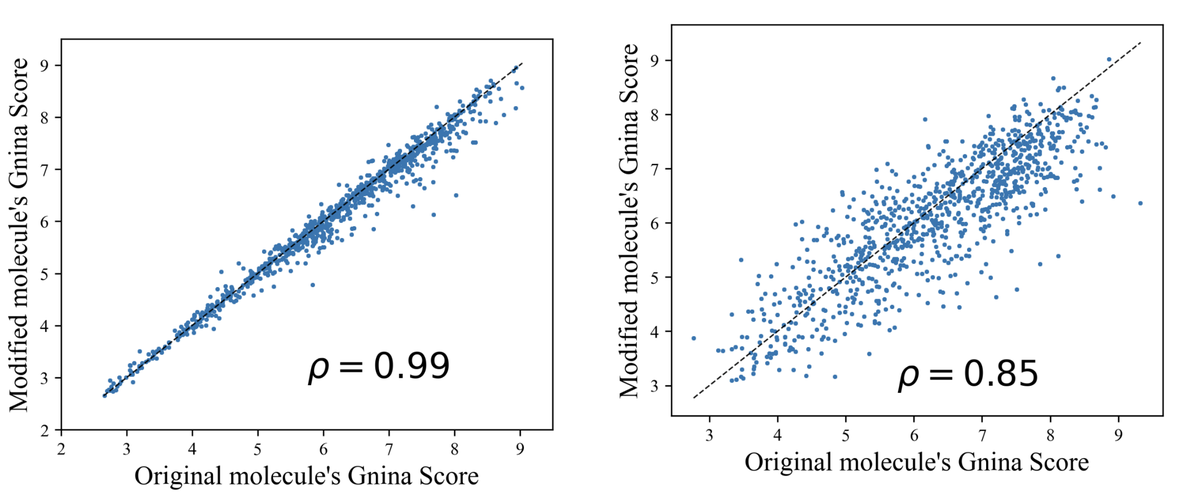


Figure 2. SimpleSBDD predicits topology optimized for binding, and then predicts atom types.



(a) Changing initial 3D configuration

(b) Changing atom types and initial 3D configuration

Figure 3. Substantial binding signal remains after changing 3D configuration and atom assignments.

Wait, can you represent topology with GNNs?

We want to predict binding from topology. However, **features that characterize the topology** (intuitively its shape and flexibility), such as:

- diameter, number of rings;
- ring sizes, number of rotatable bonds

are unlearnable by GNNs!

Protein context or persistent homology (PH) do not help

We show that, even with the additional protein context or PH features, GNNs cannot distinguish graphs which differ in topological features.

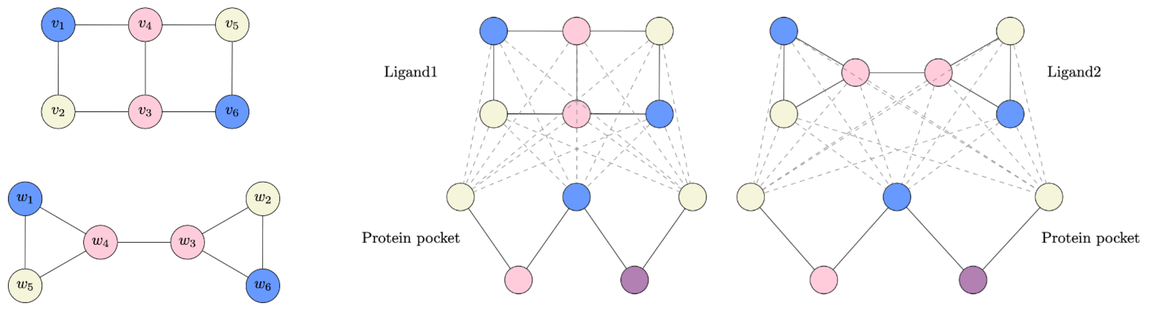


Figure 4. GNNs cannot distinguish graphs even when provided with additional context

Solution: represent topology with features unlearnable by GNNs!

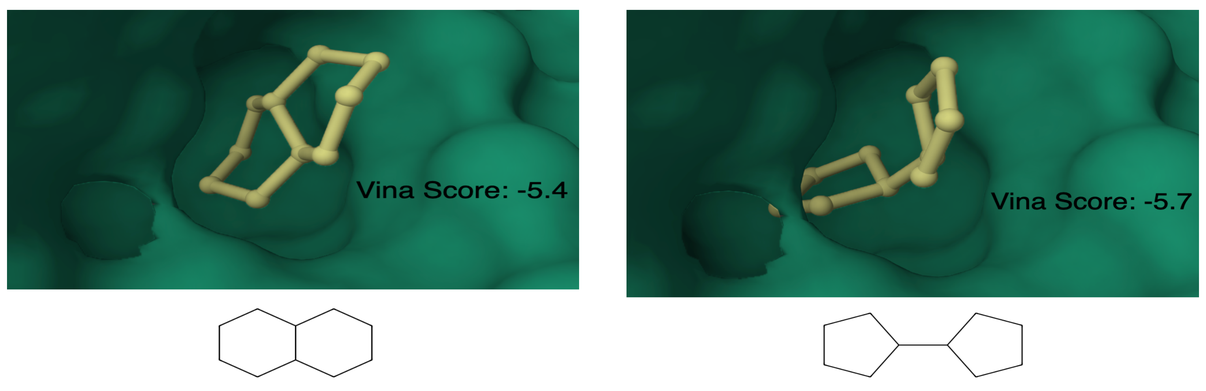


Figure 5. Two different molecules with different binding affinities, identical to GNNs.

Results

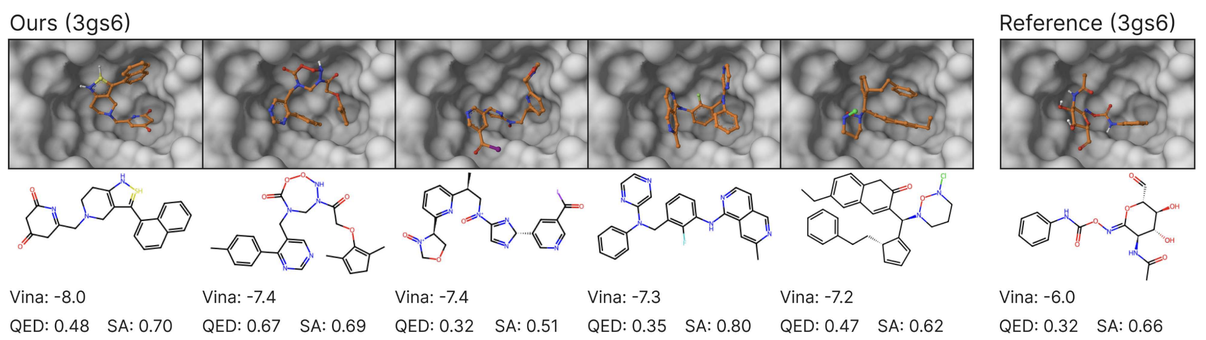


Figure 6. SimpleSBDD generates diverse high quality molecules with strong predicted binding.

Generative methods

Table 1. SimpleSBDD finds high-quality drug candidates up to 1000x faster than competing methods.

	Vina Score (kcal/mol, ↓)	High Affinity (↑)	QED (↑)	SA (↑)	Diversity (↑)	Novelty (↑)	#Params (↓)	Time (s, ↓)
Test set	-6.99 ± 2.16	-	0.48 ± 0.21	0.73 ± 0.14	-	-	-	-
DiffSBDD	-6.29 ± 1.93	0.37 ± 0.31	0.49 ± 0.19	0.63 ± 0.14	**0.79** ± 0.07	**0.54** ± 0.14	3.5M	135 ± 52
Pocket2Mol	-7.10 ± 2.56	0.55 ± 0.31	0.57 ± 0.16	0.74 ± 0.13	0.72 ± 0.16	0.45 ± 0.16	3.7M	2504 ± 220
FLAG	-7.25 ± 2.25	0.58 ± 0.24	0.50 ± 0.17	0.75 ± 0.16	0.70 ± 0.15	0.44 ± 0.17	11M	1048 ± 682
DrugGPS	-7.28 ± 2.14	0.57 ± 0.23	**0.61** ± 0.22	0.74 ± 0.18	0.68 ± 0.15	0.47 ± 0.15	14.7M	1008 ± 554
TargetDiff	-6.91 ± 2.25	0.52 ± 0.32	0.48 ± 0.20	0.58 ± 0.13	0.72 ± 0.09	0.47 ± 0.14	2.5M	3428 ± NA
DecompDiff	-6.76 ± 1.64	0.46 ± 0.36	0.45 ± 0.21	0.61 ± 0.14	0.68 ± 0.10	0.52 ± 0.13	5.0M	6189 ± NA
D3FG	-6.96 ± NA	0.46 ± NA	0.50 ± NA	**0.84** ± NA	-	-	-	-
EQGAT-diff	-7.42 ± 2.33	-	0.52 ± 0.18	0.70 ± 0.20	0.74 ± 0.07	-	12.3M	-
SimpleSBDD (Ours)	**-7.78** ± 1.47	**0.71** ± 0.34	**0.61** ± 0.18	0.69 ± 0.09	0.68 ± 0.06	0.51 ± 0.10	**23K**	**3.9**† ± 0.9

Optimization-based methods

Table 2. SimpleSBDD is significantly faster than optimization-based methods.

	Vina Score (kcal/mol, ↓)	High Affinity (↑)	QED (↑)	SA (↑)	Diversity (↑)	#Params (↓)	Time (s, ↓)
Test set	-6.99 ± 2.16	-	0.48 ± 0.21	0.73 ± 0.14	-	-	-
RGA	-6.93 ± 1.17	0.53 ± 0.41	0.46 ± 0.15	**0.80** ± 0.07	**0.76** ± 0.01	341K	11576 ± 3717
3D-MCTS	-7.55 ± 1.32	0.66 ± 0.38	0.65 ± 0.14	0.78 ± 0.07	0.62 ± 0.07	**0**	4150 ± 313
AutoGrow4	**-8.33** ± 1.55	**0.81** ± 0.28	0.36 ± 0.17	0.67 ± 0.10	0.65 ± 0.06	**0**	10800† ± 0
SimpleSBDD–$\mathcal{PO}$	-7.98 ± 1.46	0.75 ± 0.35	**0.80** ± 0.10	0.73 ± 0.08	0.67 ± 0.06	23K	**115**† ± 11

Limitations

- Docking software is only a proxy - molecular dynamics simulations are more accurate, but too expensive;
- Theory restricted to 2D graphs. What are representational limits for 3D?