Enhancing Bioactivity Prediction via Spatial Emptiness Representation of Protein-ligand Complex and Union of Multiple Pockets

Zhiyuan Zhou, Yueming Yin, Yiming Yang, Yuguang Mu, Hoi Yeung Li, Adams Wai-Kin Kong

Nanyang Technological University, Singapore

Research Background

- **Bioactivity** is the measurable effect a compound exerts on a biological system

- Accurate prediction of bioactivity can help identify the most promising compounds

- The task is challenging, and there are multiple labels

Conventional methods: rely on the intrinsic properties of small molecules (Ligand only). E.g., QSAR models

Issue:
A compound cannot function biologically without its target and environment.

Recent methods: using protein information and constructing interaction graphs

Issues:
1. Each pocket–ligand pair is treated as an **independent data**
2. Only concentrate on **local** binding region

Ligand only ➡ **Protein-ligand interaction based**

Motivation

- Bioactivity arises from specific biological interactions, and the **geometric information** of binding pocket is crucial

- The principle of **competitive inhibition**: drug molecules bind to the active site of a protein and block access for endogenous substrates

- Key geometric info: **Vacant space**. How much empty space remains, and to what extent does the ligand occupy the pocket?

- Vacant space small → endogenous substrate blocked → inhibition successful
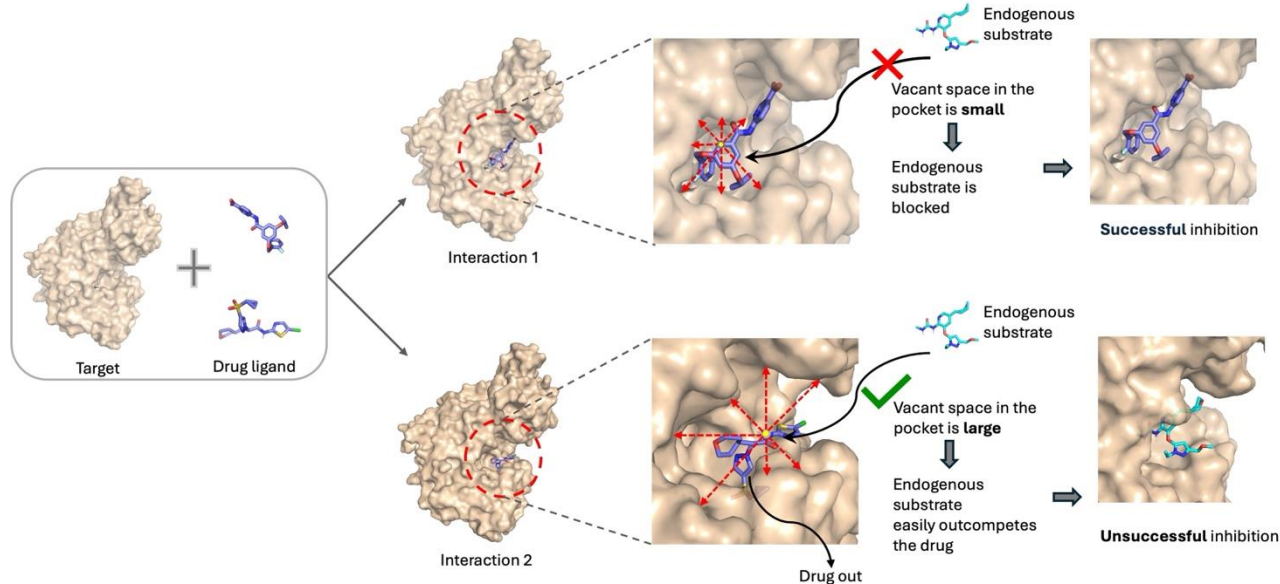
Main Contributions

LigoSpace:

- **Geo**metric **R**epresentation of spatial **E**mptiness in protein-ligand **C**omplex (GeoREC)

    - Describe the surroundings of each node in the protein-ligand graph representation

    - Quantify atomic-level empty space within the complex

- Union-Pocket

    - The entire pocket space can be accessible

    - Keep the protein information consistent, encourage the model to focus on the ligand

- Incorporate a **pairwise loss**

    - Traditional loss (MSE) is symmetric and treats each data point independently

    - Preserve the relative order among samples

Method

- ➤ GeoREC (**Geo**metric **R**epresentation of spatial **E**mptiness in protein-ligand **C**omplex)
  - ➤ Characterizes spatial configuration inside binding pockets by dividing surrounding space into cones, identifying nearest atoms in each region, and connecting them as geometric edges
- ➤ Global interaction model with Union-Pocket
  - ➤ Addresses fragmented local pockets from previous methods by defining **Union-Pocket** as the union of all atoms from candidate pockets for a given protein
- ➤ Pairwise loss
  - ➤ Predict the difference between any pair of data
  - ➤ Preserves relative order information, improving suitability for drug discovery
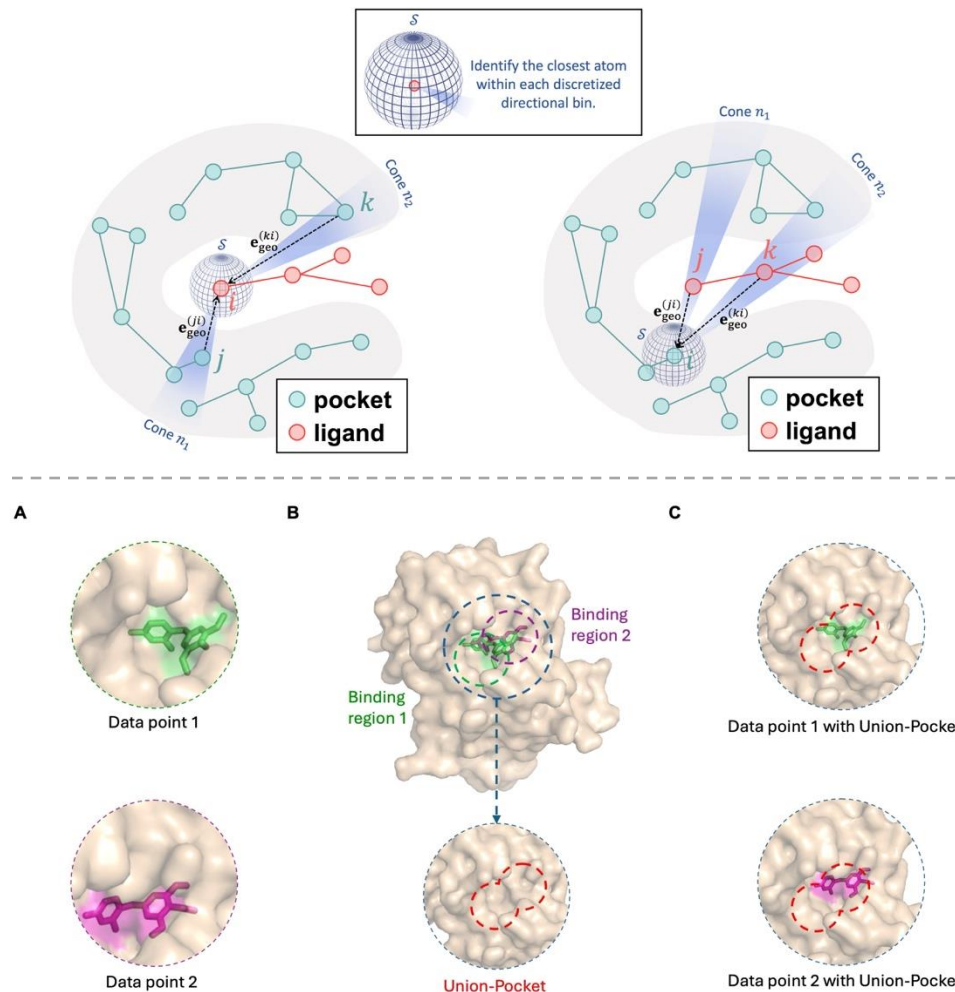
Result

➤ Tested against 6 baselines (DTIGN, GIGN, GAT, EHIGN, SIGN, MBP) on DTIGN & SIU datasets

➤ The labels in the dataset include Kd, Ki, IC50 and EC50

➤ LigoSpace consistently enhances the baseline across most datasets and metrics

➤ Table below shows the results on DTIGN dataset with 8 protein targets

Model	Metric	Dataset								Avg	Avg Imp%
		I1	I2	I3	I4	I5	E1	E2	E3		
DTIGN baseline	RMSE ↓	1.1977	0.7952	1.1443	0.9396	0.9518	0.9086	0.7765	1.0869	0.9751	-
	r ↑	0.3547	0.7128	0.7839	0.6177	0.6227	0.3363	0.6946	0.4825	0.5757	-
	τ ↑	0.2445	0.4922	0.5991	0.4574	0.4691	0.2139	0.4917	0.3786	0.4183	-
DTIGN+LigoSpace	RMSE ↓	**1.0364**	**0.6507**	**1.0245**	**0.8124**	**0.8966**	**0.8645**	**0.7262**	**1.0150**	**0.8783**	9.93%
	r ↑	**0.5895**	**0.7888**	**0.8251**	**0.7334**	**0.6986**	**0.4969**	**0.7275**	**0.5924**	**0.6815**	18.39%
	τ ↑	**0.4239**	**0.5454**	**0.6360**	**0.5205**	**0.5083**	**0.3705**	**0.5243**	**0.4563**	**0.4982**	19.09%
GIGN baseline	RMSE ↓	1.2365	0.7487	1.0677	0.8560	0.8700	0.9563	0.7545	0.9704	0.9325	-
	r ↑	0.4357	0.7234	0.8080	0.6888	0.7302	0.4435	0.7246	0.6285	0.6478	-
	τ ↑	0.3216	0.5083	0.6310	0.4920	0.5439	0.3443	0.5043	0.4664	0.4765	-
GIGN+LigoSpace	RMSE ↓	**0.9515**	**0.6880**	**0.9403**	**0.8111**	**0.8302**	**0.8733**	0.8737	**0.9161**	**0.8605**	7.72%
	r ↑	**0.6861**	**0.7735**	**0.8539**	**0.7450**	**0.7513**	**0.4912**	0.6637	**0.6860**	**0.7063**	9.03%
	τ ↑	**0.4899**	**0.5378**	**0.6673**	**0.5439**	**0.5591**	**0.3786**	**0.5511**	**0.5228**	**0.5313**	11.51%
GAT baseline	RMSE ↓	1.1801	0.9270	1.3433	1.0200	1.0547	0.8962	0.8579	1.1716	1.0564	-
	r ↑	0.4147	0.4771	0.6663	0.5091	0.5035	0.3771	0.5836	0.3468	0.4848	-
	τ ↑	0.2267	0.3115	0.4618	0.3685	0.3428	0.2451	0.3917	0.2938	0.3302	-
GAT+LigoSpace	RMSE ↓	**1.1767**	**0.8906**	**1.3352**	**1.0037**	**1.0342**	0.9252	0.8587	**1.1476**	**1.0465**	0.93%
	r ↑	**0.4511**	**0.5437**	**0.6813**	**0.5570**	**0.5302**	0.3604	**0.5838**	**0.4060**	**0.5142**	6.07%
	τ ↑	**0.2922**	**0.3593**	**0.4823**	**0.4075**	**0.3690**	**0.2497**	0.3819	**0.3330**	**0.3594**	8.82%

Ablation study

➢ **Setup:** each ablation removes exactly one module from the LigoSpace model to isolate its effect.

➢ **GeoREC (G):** removing G causes the largest performance drop — GeoREC is the single most critical component.

➢ **Union-Pocket (U):** removing U yields consistent, moderate declines and shows strong synergy with GeoREC.

➢ **Hybrid loss (L):** removing L mainly harms ranking/order consistency and slightly reduces overall accuracy.

➢ **Overall:** the full configuration (G + U + L) performs best — components are complementary and work best together.

Table 3: Ablation study on DTIGN baseline (each variant excludes one component)

Dataset	DTIGN-GUL			DTIGN-GU(L)			DTIGN-(G)UL			DTIGN-G(U)L		
	RMSE↓	R↑	τ↑	RMSE↓	R↑	τ↑	RMSE↓	R↑	τ↑	RMSE↓	R↑	τ↑
I1	**1.0364**	0.5895	**0.4239**	1.1163	0.5259	0.3494	1.1341	0.4721	0.3249	1.0739	**0.5941**	0.4128
I2	**0.6507**	**0.7888**	**0.5454**	0.7292	0.7284	0.5081	0.7724	0.7077	0.4665	0.7395	0.7331	0.5265
I3	**1.0245**	**0.8251**	**0.6360**	1.0557	0.8153	0.6175	1.1170	0.8196	0.6348	1.2061	0.7486	0.5446
I4	**0.8124**	**0.7334**	**0.5205**	0.8550	0.7072	0.5023	0.9149	0.6666	0.4725	0.8364	0.7149	0.5067
I5	0.8966	0.6986	0.5083	**0.8625**	**0.7127**	**0.5225**	0.9591	0.6343	0.4561	0.9017	0.6928	0.4994
E1	0.8645	**0.4969**	**0.3705**	**0.8586**	0.4513	0.3498	0.9513	0.3064	0.1727	0.9243	0.3176	0.1983
E2	**0.7262**	0.7275	0.5243	0.7600	0.7083	0.5066	0.8263	0.6384	0.4300	0.7323	**0.7314**	**0.5332**
E3	1.0150	0.5924	**0.4563**	**1.0119**	**0.5955**	0.4524	1.0679	0.5198	0.3917	1.0438	0.5656	0.4332
Avg	**0.8783**	**0.6815**	**0.4982**	0.9062	0.6556	0.4761	0.9679	0.5956	0.4187	0.9323	0.6373	0.4568

Thanks for your attention!