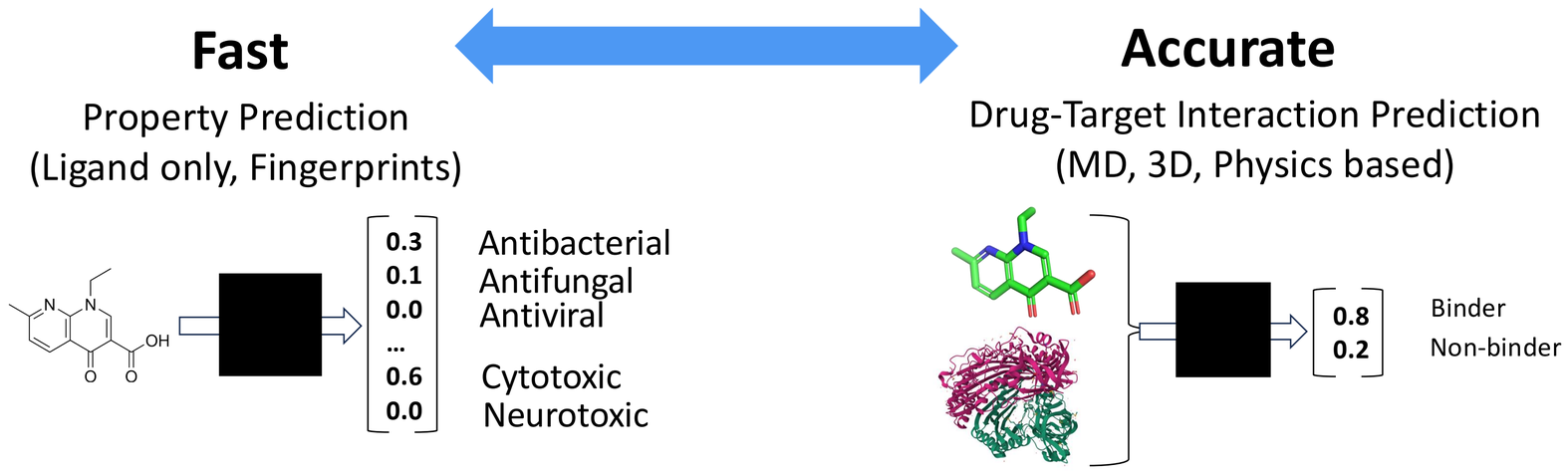
SPRINT: Ultra-Fast Virtual Screening

Andrew T. McNutt[1] Abhinav K. Adduri[2] Caleb N. Ellington[2,3] Monica T. Dayao[2] Eric P. Xing[2,3,4] Hosein Mohimani[2] David R. Koes[1]

1 University of Pittsburgh 2 Carnegie Mellon University 3 GenBio AI 4 MBZUAI

Virtual Screening

Experimental screening of small molecules is essential for drug discovery and development, but *in vitro* screening is a difficult and time-consuming process. Virtual screening of these drugs against protein targets can inform experiments by predicting drug-target interactions. However, the size of molecular libraries (~10^9) are now surpassing the capabilities of high accuracy structure-based methods.

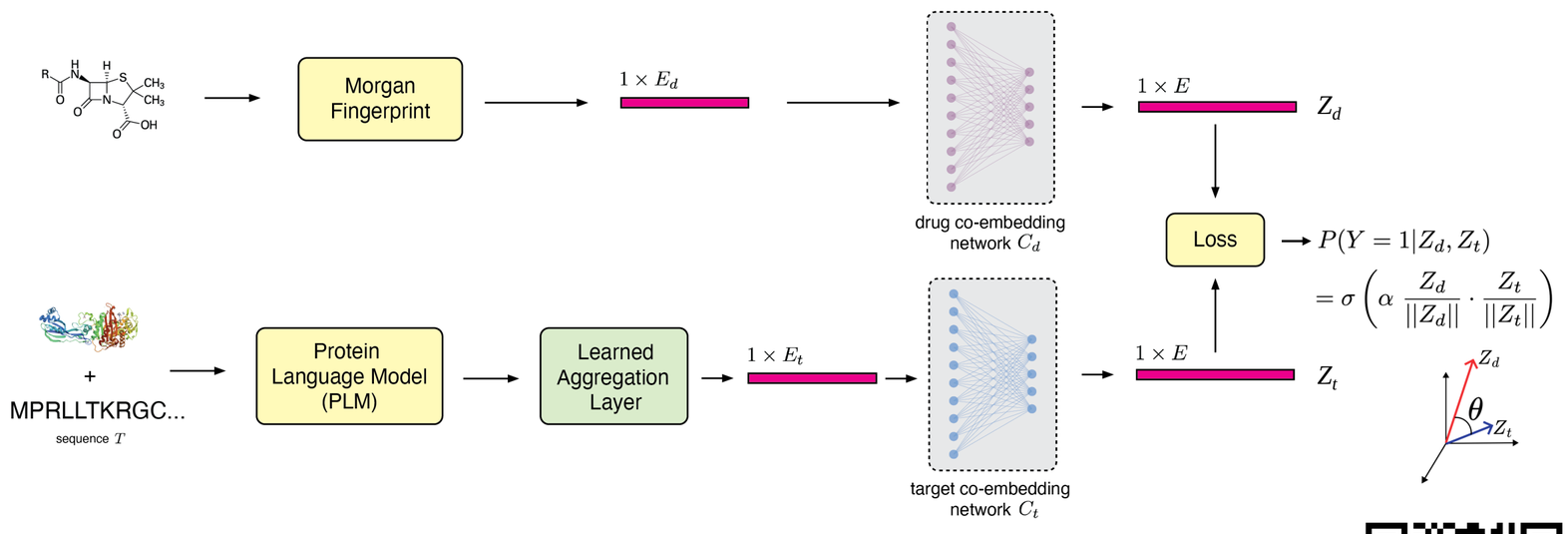


SPRINT: Ultra-Fast Virtual Screening Mines Massive Databases for New Drug Mechanisms

We seek to identify novel antimicrobial mechanisms by **screening entire chemical databases (~10^9 entries) against all known bacterial, fungal, and human proteomes (~10^6 proteins).**

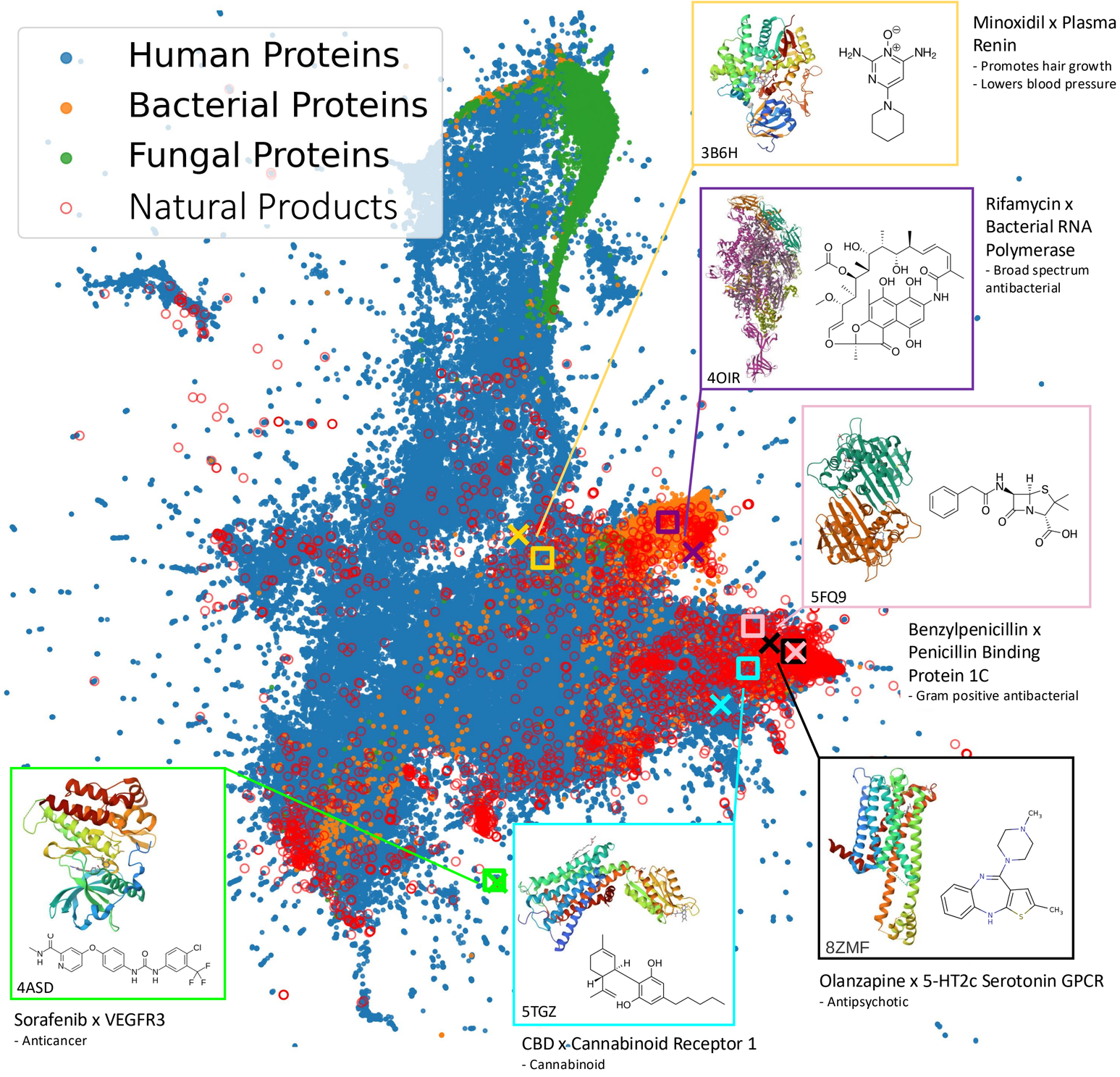
To do this, we develop Structure-aware PRotein ligand INTeraction (SPRINT) by

1. Improving metric-based representation learning methods [1] to achieve a new SOTA for DTI prediction.
2. Integrating structure-aware PLMs to use 3D structural information with efficient 1D vector computation.
3. Applying efficient vector retrieval methods from NLP to predict binding partners.



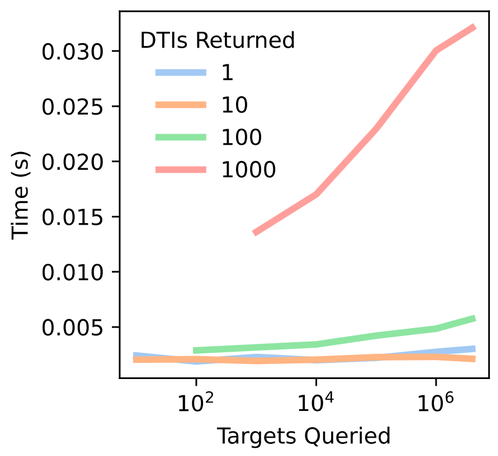
Try it with Colab Screen! *bit.ly/colab-screen*

arXiv



How Fast is SPRINT?

Ultra. Co-embedding enables SPRINT to use lightning-fast vector databases for prediction. Querying a ligand for the top 10 binders in UniProt (10^7 proteins) takes **<0.01s**. Querying a protein for the top 100 ligands in ChEMBL (10^6 molecules) also takes **<0.01s.** Screening the whole human proteome against the ENAMINE Real Database (6.7B drugs) for the 100 most likely binders per protein takes **16 minutes**.



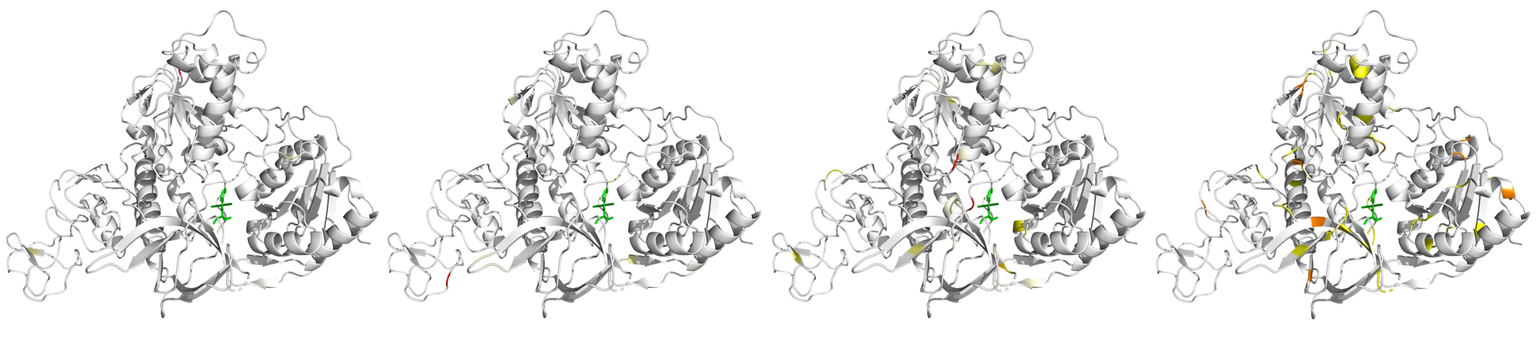
How Accurate is SPRINT?

We benchmark on LitPCBA, a challenging virtual screening dataset. Surprisingly, SPRINT outperforms computationally intensive structure-based methods. Without a structure-aware PLM, SPRINT-ProtBert, performance decreases showing the utility of structure for virtual screening.

	AUROC (%)	BEDROC (%)	EF		
			0.5%	1%	5%
Surflex [21]	51.47	-	-	2.50	-
Glide-SP [22]	53.15	4.00	3.17	3.41	2.01
Planet [23]	57.31	-	4.64	3.87	2.43
GNINA [24]	60.93	5.40	-	4.63	-
DeepDTA [25]	56.27	2.53	-	1.47	-
BigBind [26]	60.80	-	-	3.82	-
DrugCLIP [9]	57.17	6.23	8.56	5.51	2.27
SPRINT-Average (15.7M)	67.49	7.80	7.23	6.26	3.71
SPRINT-ProtBert (13.4M)	**73.4**	11.9	11.68	10.19	5.27
SPRINT-sm (16M)	**73.4**	**12.3**	**15.90**	**10.78**	**5.29**

Learned aggregation mapped to the structure

We can visualize the attention of the learned aggregation layer on the structure of the protein to understand the residues which were most relevant to the decision. Here we visualize the target structure of CACHE challenge 2 (PDB ID: 5RLZ).



What else does SPRINT do?

Many molecular "properties" (e.g. bioactivity, toxicity) result from interactions within a biological system. Co-embedding localizes drug-target interactions and improves property prediction in terms of F1 score.

Task	Morgan Fingerprint	Co-embedding
Antibacterial	0.564 ± 0.031	**0.571** ± 0.028
Antifungal	0.365 ± 0.094	**0.366** ± 0.081
Antiviral	0.266 ± 0.159	**0.293** ± 0.105
Toxicity†	0.611 ± 0.068	**0.622** ± 0.073