TrustAffinity: accurate, reliable and scalable out-of-distribution protein-ligand binding affinity prediction using trustworthy deep learning

Amitesh Badkul[1], Li Xie[1], Shuo Zhang[1,3], Lei Xie[1, 2, 3]

[1]Hunter College, City University of New York [2]The Graduate Center, City University of New York
[3]Weill Cornell Medicine, Cornell University

Motivation

Despite, the integration of AI for accelerating drug discovery, the existing algorithms fail:

- To **generalize** for understudied proteins or compounds from unlabeled protein families or chemical scaffolds.
- To **quantify uncertainty** associated with their predictions.
- To achieve **scalability** to billions of compounds.

Overcoming these challenges is key to successfully integrating machine learning with drug discovery with higher efficacy and safety.

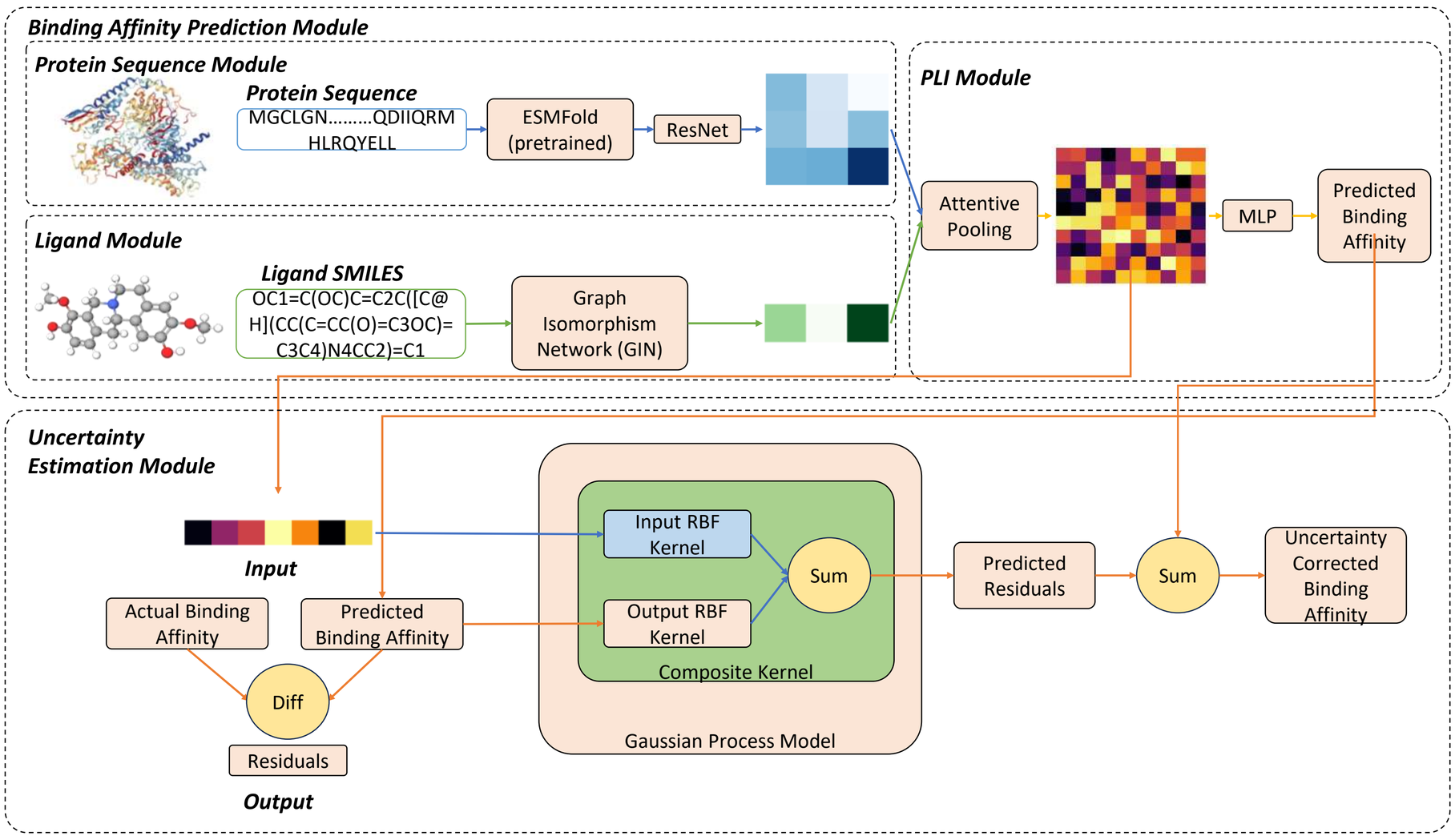
Contributions

We address these limitations by introducing **TrustAffinity**:

- A **sequence based** deep learning framework with efficient uncertainty quantification based on residue-estimation.
- Rigorously validated TrustAffinity across various **OOD scenarios**, surpassing **SOTA methods**.
- We demonstrate the practicality of our framework through **Opioid Use Disorder lead discovery**.

These are vital for the successful incorporation integrating machine learning with drug discovery.

Method

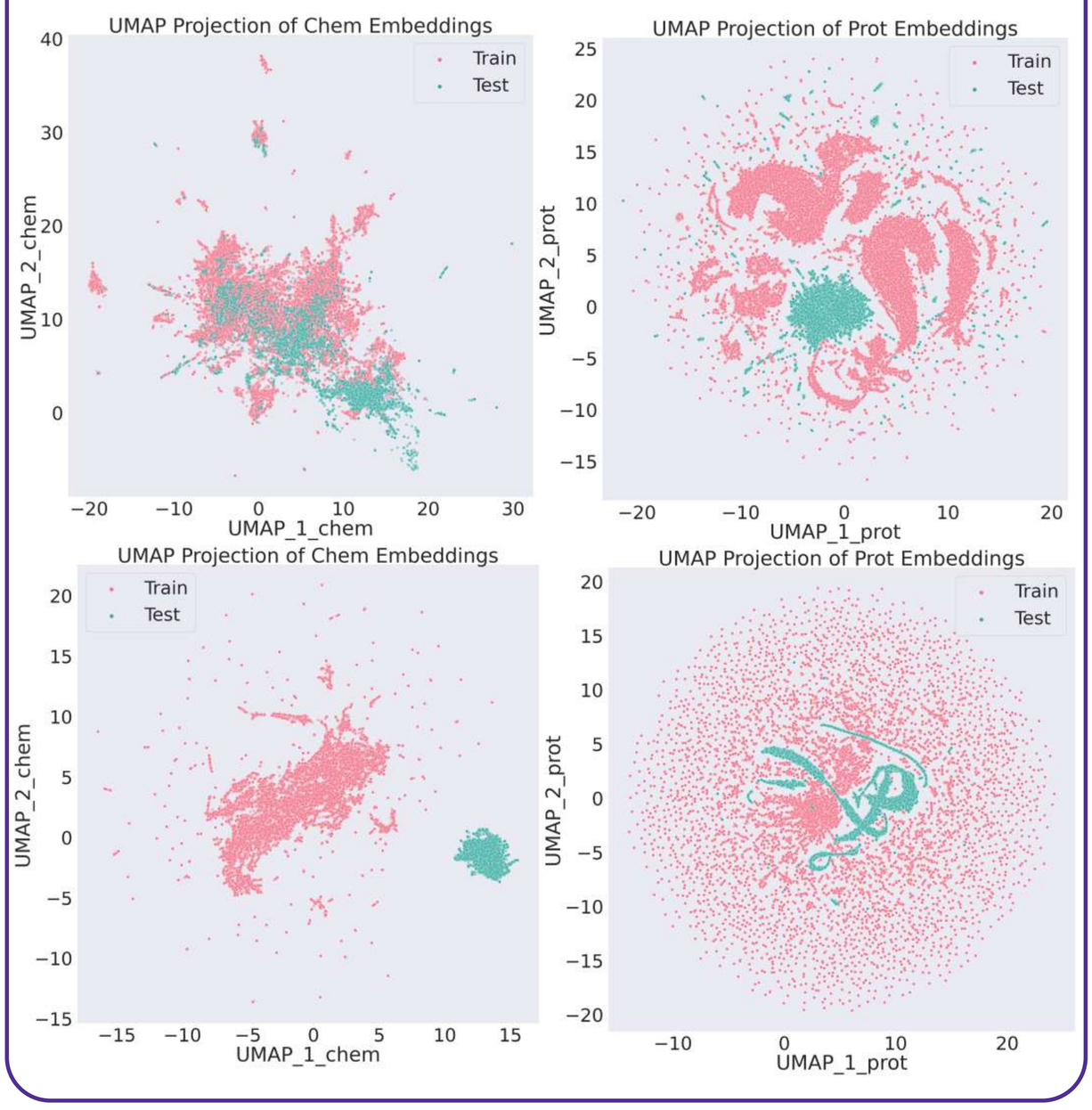


TrustAffinity is validated through:

- Pfam (protein family) split
- Scaffold (chemical scaffold) split
- Random split
- Opioid Use Disorder (lead discovery)

Read more here!

Results



Opioid Use Disorder (OUD) Results

Method	RMSE	MAE	r	ρ
AutoDock Vina	1.179	1.031	0.308	0.334
BACPI	2.523	2.181	0.103	0.122
TrustAffinity (Y_{true})	**0.384**	**0.312**	**0.856**	**0.820**
TrustAffinity (Y_{pred})	0.846	0.65	0.612	0.667

Conclusion

TrustAffinity successfully:

- Displays ability to differentiate in Pfam and Scaffold **OOD scenarios**.
- Achieves **high confidence** performance when compared to **machine learning** and **docking based** methods.
- Exhibits **highest performance** for OUD lead discovery case.
- Runs at **three orders** of magnitude faster than traditional protein-ligand docking.