Fusion of Structure Based Deep Learning to Accelerate Molecular Docking Predictions

Derek Jones

LLNL-PRES-764149

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Lawrence Livermore National Laboratory

Artificial intelligence is infiltrating into our daily lives

- Image classification
- Image segmentation
- Speech recognition
- Speech synthesis
- Search
- Ad recommendation
- Games
- Image enhancement
- Synthetic image generation
- Identification of planets
- Predicting elections
- Self driving car
- Path planning
- Visual recognition

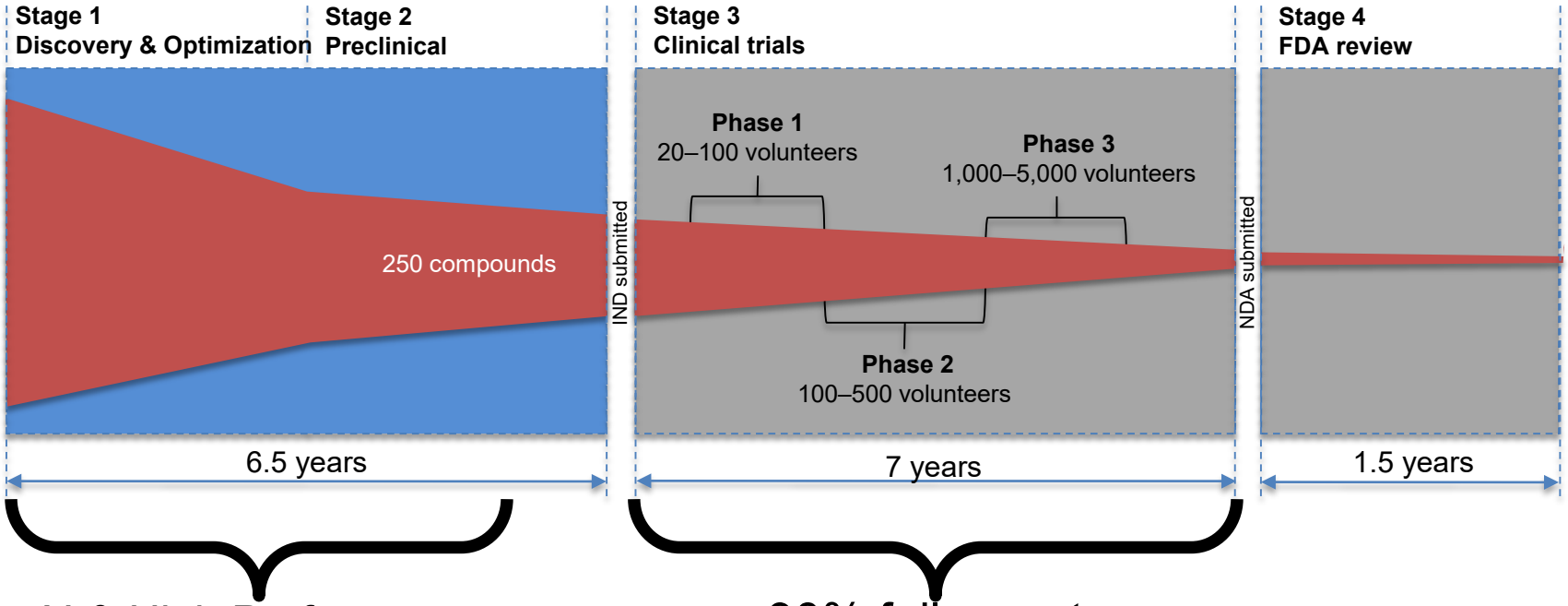
1990s

Sample of cats & dogs images from Kaggle Dataset

2016

2020??

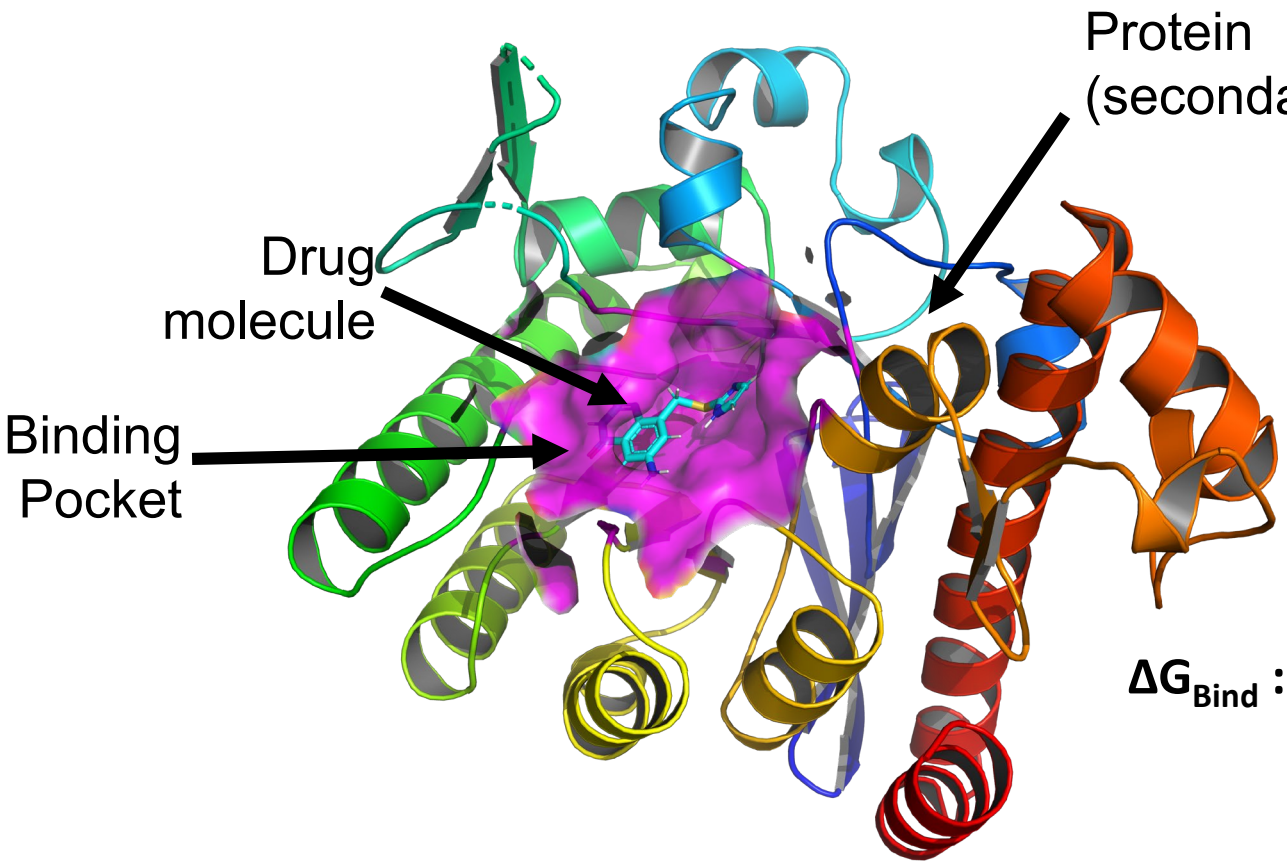
We need to predict human outcome to drug candidates prior to clinical trials



Challenges and Opportunities Report, FDA, 2004

Need to prioritize new therapeutic candidates and mitigate the risk of failure in clinical trials

Can AI design our next drugs?



ΔG_{Bind} : -Docking
-MM/GBSA
-Free energy (FEP)

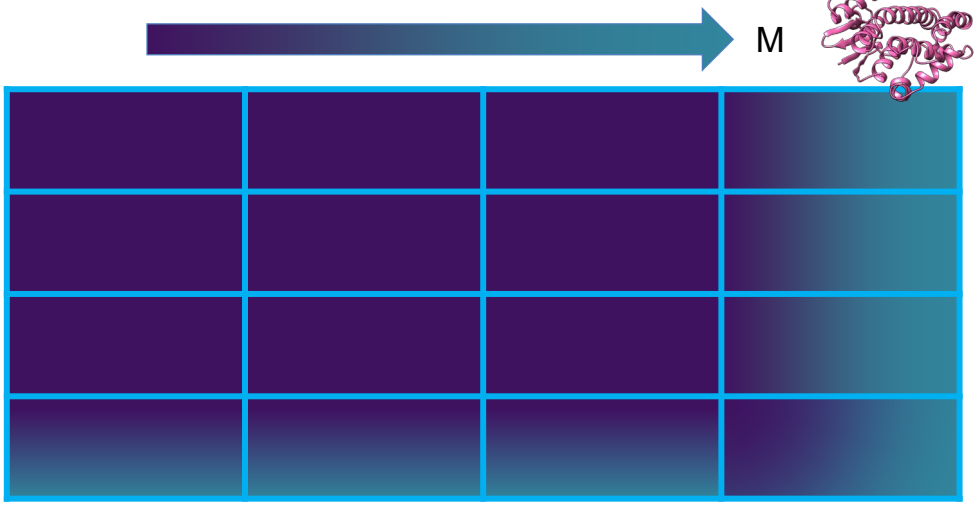
Binding pocket of complex 1q63 from PDBBIND 2007

We are creating a queryable protein-small molecule atlas

American Heart Association.

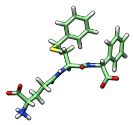
- Matrix is **sparse!**
- Simulations can be used to circumvent costly experimental steps in the drug discovery process
- **Approach:** Create a large database of computed protein ligand interactions

Human Proteins: 28,658 models for 14,813 proteins



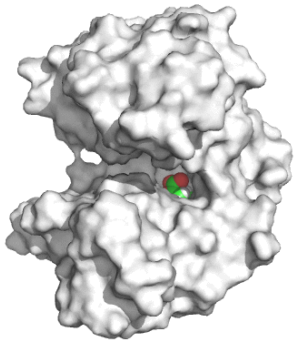
Small Molecule Compounds: ~700M

Enamine	~679M
eMolecules	~18M
ChEMBL24:	1.8M
Approved drugs:	5,557
Foodome:	24,114
NCI60:	244,800
PDBBind:	14,744



Physics based protein-ligand binding affinity does not scale to modeling billions of interactions

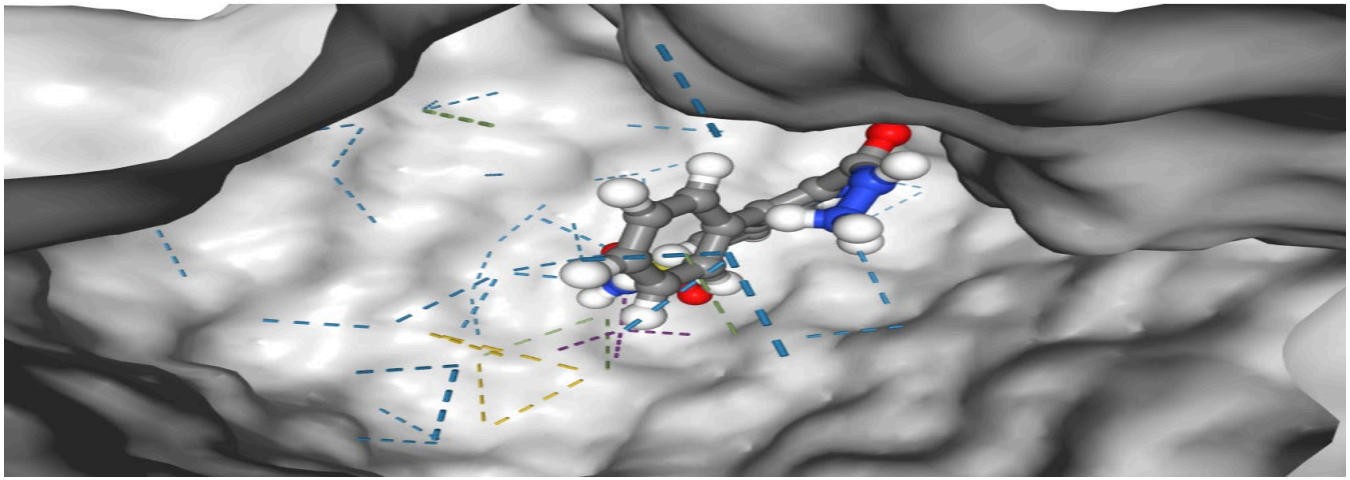
- Vina – speed=fast (14 minutes)
- GBSA – speed=moderate (62 minutes)



- GBSA does not scale well for large virtual screens (e.g. millions of molecules against the human proteome).
- Docking offers a clear speed advantage at the cost of accuracy (best used to "enrich" a large set of molecules)

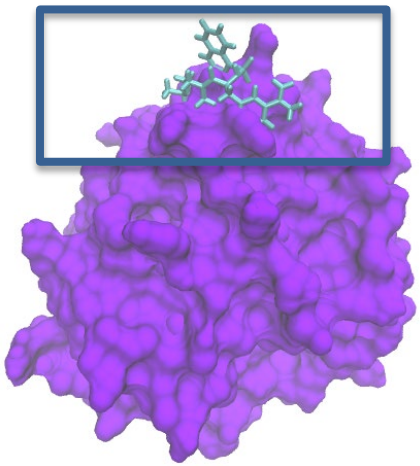
Machine learning models can scale to screen large compound sets but…struggle to maintain accuracy

- Key hypotheses we are testing
  1. Modeling spatial features from both the ligand and target enable
     - Training on a corpus of crystal structures, which generalize to new molecule / protein pocket combinations
  2. Deep learning models meet or exceed docking based scoring models in some cases.
     - Key future challenge is to measure confidence in ML predictions versus physics based calculations



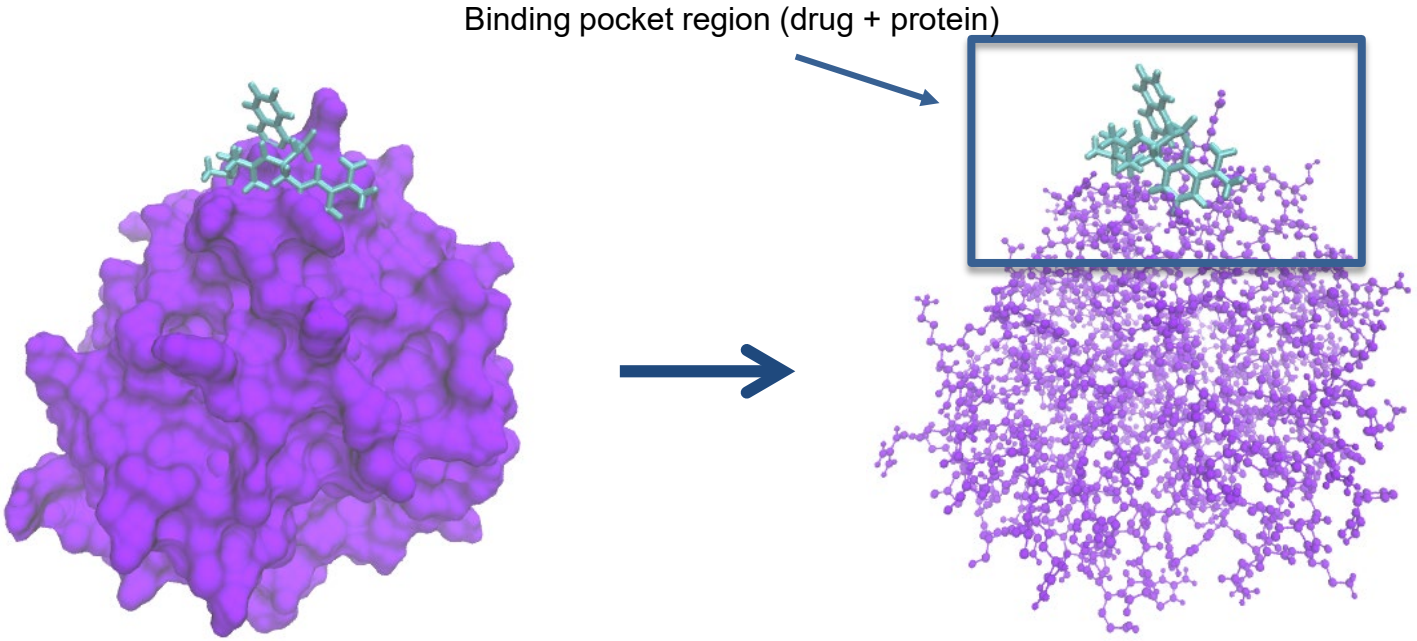
3D CNN and SGCNN share common base representations of the binding complex

- Atom type: one-hot encoding of B, C, N, O, P, S, Se, halogen or metal (9 bits)
- Atom Hybridization (1, 2, or 3) (1 integer)
- Number of heavy atom bonds (i.e. heavy valence) (1 integer)
- Number of bonds with other heteroatoms (i.e. hetero valence) (1 integer)



- Structural properties: bit vector (1 where present) encoding of hydrophobic, aromatic, acceptor, donor, ring (5 bits)
- Partial Charge (1 float)
- Molecule type to indicate protein atom versus ligand atom (-1 for protein, 1 for ligand) (1 integer)
- Van der Waals radius (1 float)

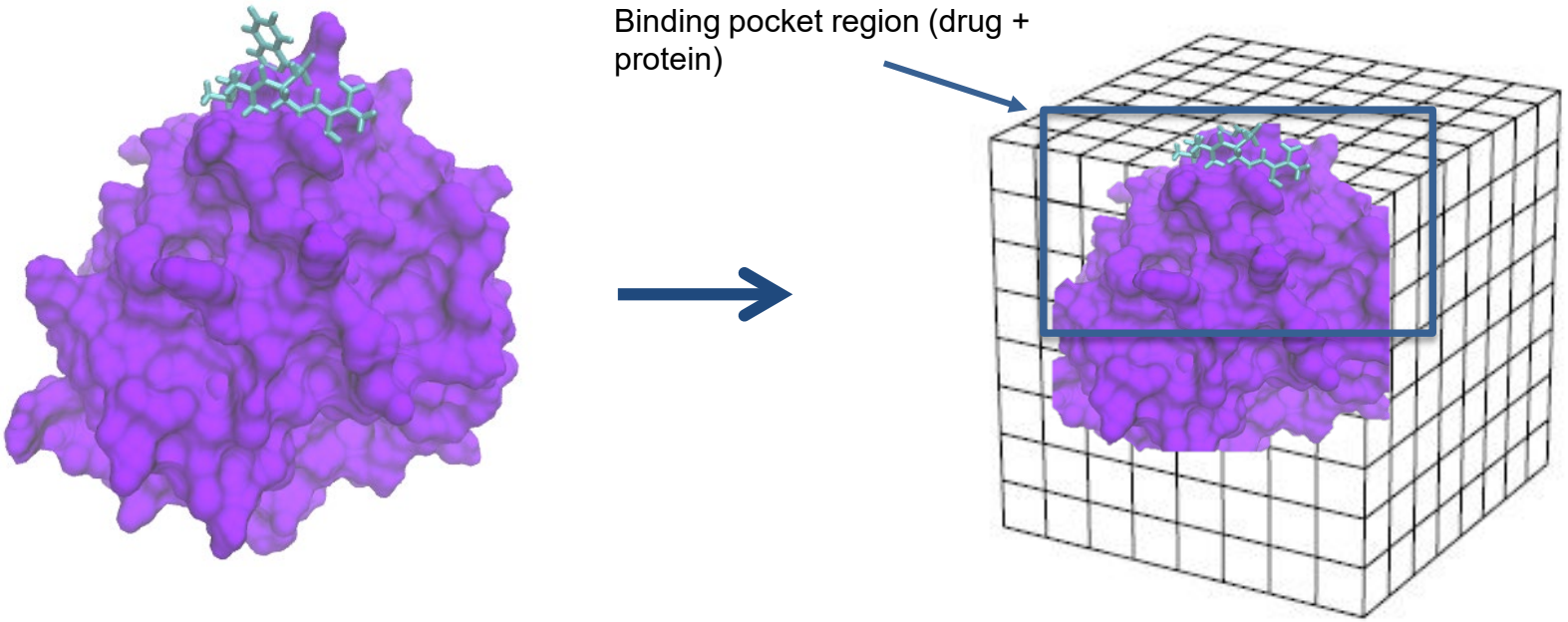
Spatial GCNN Feature Extraction



The result of the feature extraction which uses the atom and bond information of the binding complex to construct a graph representation

*based on “PotentialNet for Molecular Property Prediction” by Feiberg et. al.

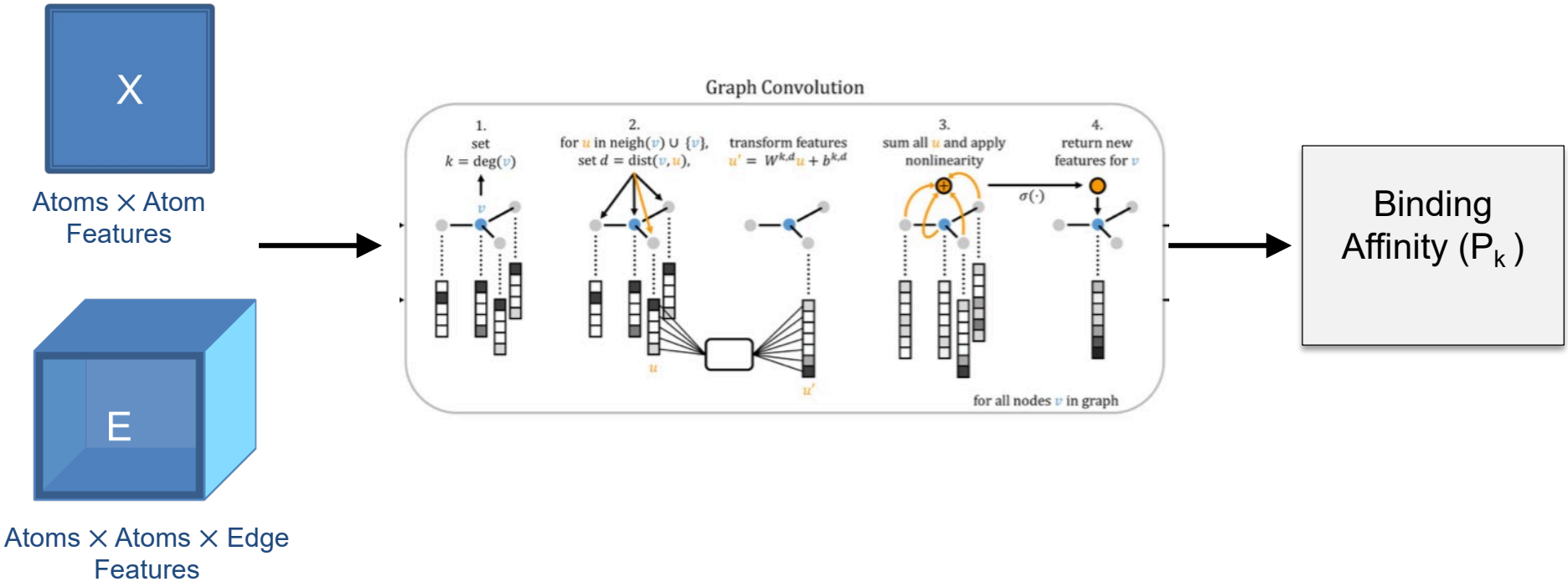
3D-CNN Feature Extraction



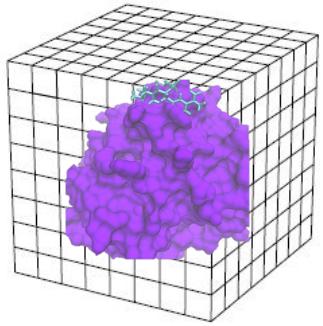
Binding complex atoms are placed into 3D grid based on coordinates, with a feature channel for each atom feature type

Based on KDEEP: Protein–Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks
By jiminez et. al.

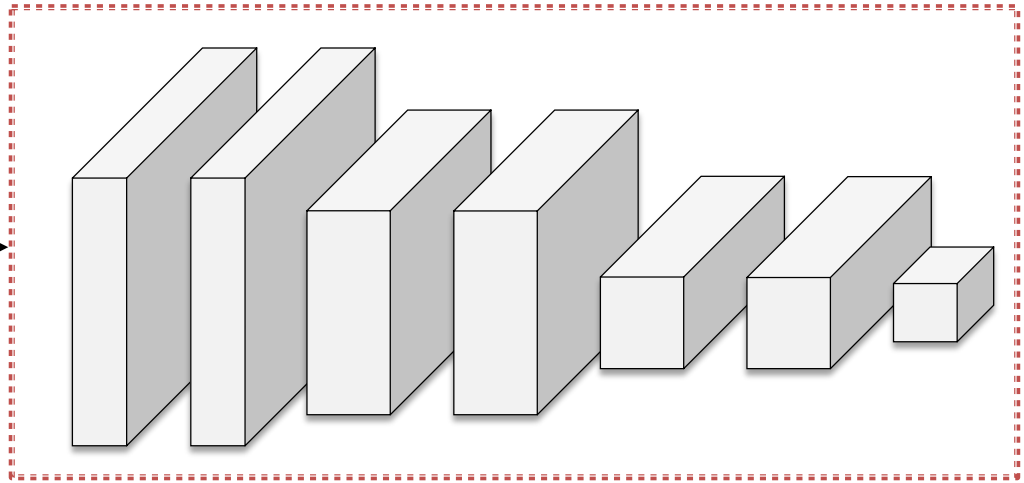
SGCN utilizes graph representation to infer spatial connectivity and interactions between atoms



3D CNN incorporates spatial information of protein-ligand atoms to capture spatially-coherent features



3D voxel representation of protein-ligand complex

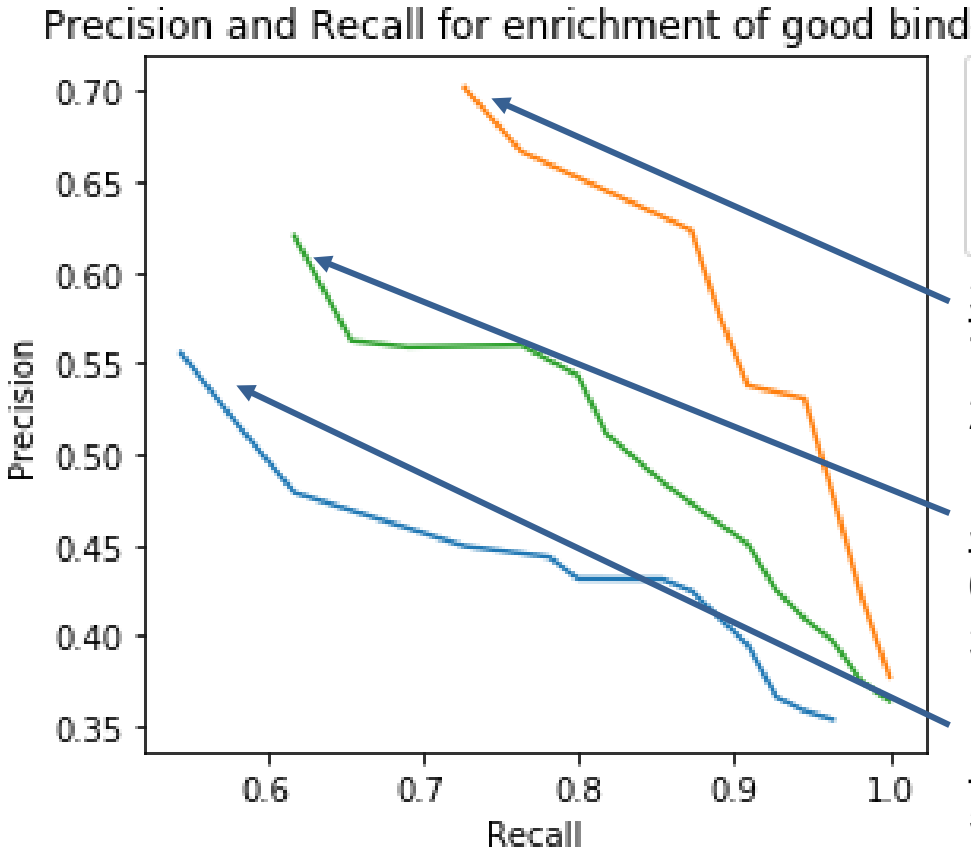


3D Convolutional Layers

Fully Connected Layers

Binding Affinity (P_k)

Deep learning models improve scoring accuracy over traditional docking scores



3D CNN
70% correctly labeled
28% of true binders missed

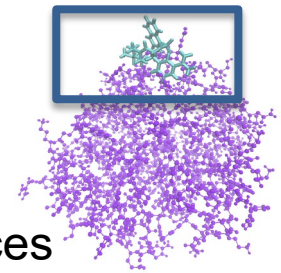
Spatial Graph
63% correctly labeled
38% of true binders missed

Docking
55% correctly labeled
48% of true binders missed

Curve shows performance as the scoring threshold is adjusted

A novel deep learning fusion model integrates multiple views of the protein-ligand interaction

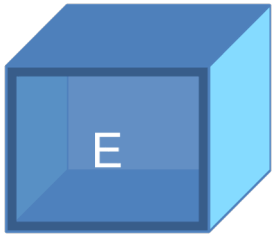
Spatial Graph: record explicit atom – atom interactions as ***graph***



Vertices

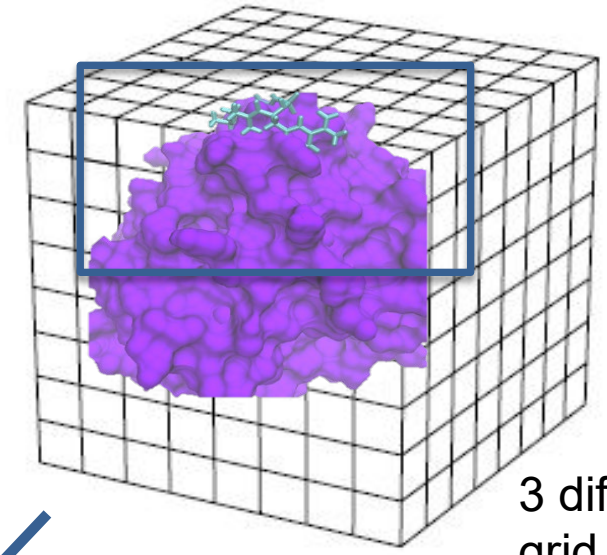
Atoms × Atom Features

Edges



Atoms × Atoms × Edge Features

3D-CNN: record implicit atom-atom interactions as ***3D image***



3 different grid sizes are used

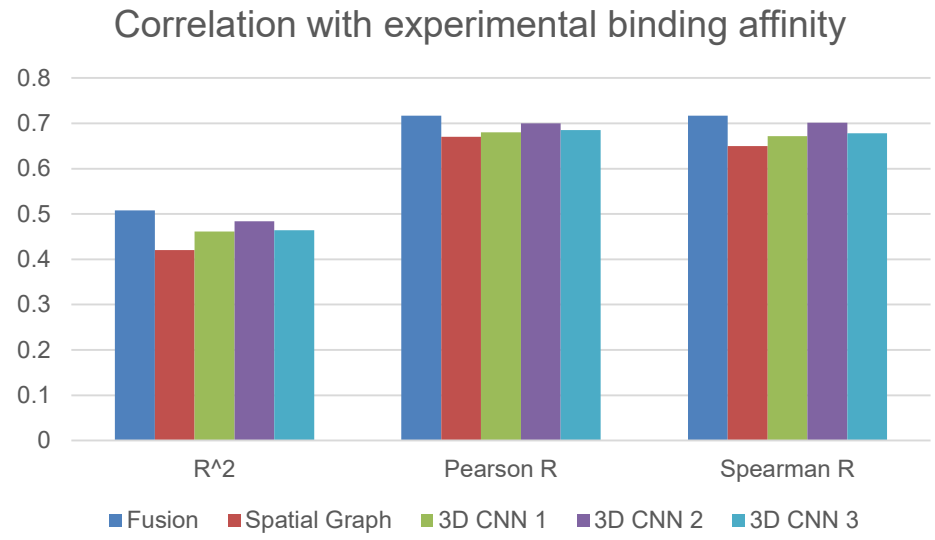
Deep Learning Fusion Model

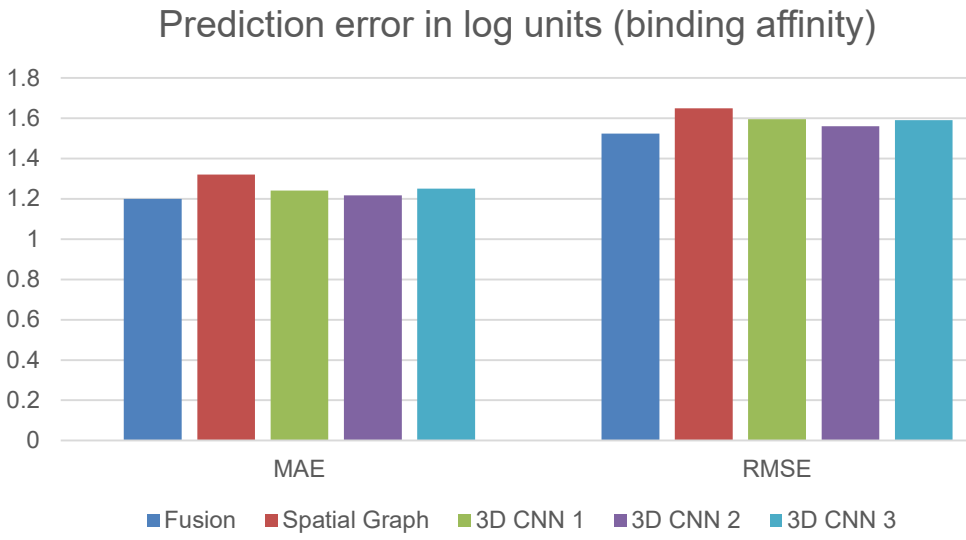
Binding Affinity (P_k)

pocket

Previous efforts have used one method or the other but not both together

Fusion model shows modest but consistent improvement over all accuracy metrics

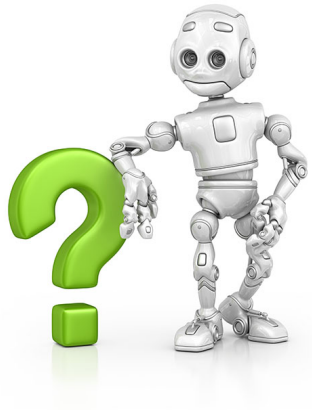




Fusion model combines Spatial Graph and 3D CNN representations to make a single bind affinity prediction on newly docked complexes

Next steps

- A more accurate scoring function that is cheaper to evaluate – greater throughput and performance in simulation steps, enabling greater scaling efficiency in HPC environments
- Integration of deep learning based scoring function into HPC docking pipeline (ConveyorLC)* to enhance mechanistic modeling capabilities
- Quantifying model uncertainty to help inform docking pipelines and users of the ML predictions in broader applications



LLNL's sierra cluster

* https://github.com/XiaohuaZhangLLNL/conveyorlc

Team

Drew Bennett

Brian Bennion

Adam Zemla

Xiaohua Zhang

Dan Kirshner

Sergio Wong

Fangqiang Zhu

Jonathan Allen

Hyojin Kim

Amanda Minnich

Derek Jones

Marisa Torres

Aseeva Masha

Ed Lau

American Heart Association®

Questions?

Lawrence Livermore
National Laboratory

Deep Learning Libraries FYI

https://pytorch.org/

https://github.com/rusty1s/pytorch_geometric

https://www.tensorflow.org/