End-to-End Learning on 3D Protein Structure for Interface Prediction

Raphael J. L. Townshend
Stanford University
raphael@cs.stanford.edu

Rishi Bedi
Stanford University
rbedi@cs.stanford.edu

Patricia A. Suriana
Stanford University
psuriana@stanford.edu

Ron O. Dror
Stanford University
rondror@cs.stanford.edu

Presented by: Rahmatullah Roche

COMPUTER SCIENCE
VIRGINIA TECH

The Problem

- Protein function largely depends on their binding to one another
- This work addresses paired protein interface prediction
- If a non-H atom is within 6 Angstrom of any of atoms of its pairing protein, it's in interface

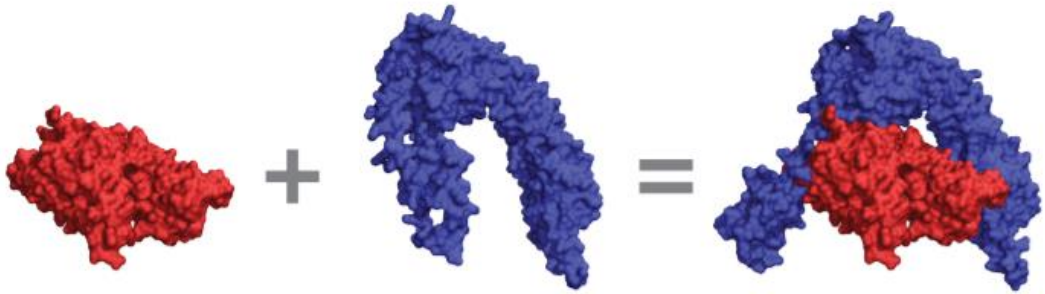


Figure 1: Protein Binding. The BNI1 protein (blue) opens up to bind to actin (red). While our method is trained only using structures of complexes such as the one at right, without any information on how the individual proteins deformed upon binding, we test on pairs of unbound structures such as those at left with minimal loss in performance.

The Approach

- SASNet: an end-to-end learning method applied to interface prediction
- Instead of hand-crafted features, directly uses atomic positions and ids
- Voxelizes the local atomic environments, or "surfacelets“
- Applies a siamese-like 3D CNN

Related Works

- Graph-based approaches
  - deriving properties of small molecules
  - graph policy networks to generate new molecules
- Symmetry functions for protein-ligand binding affinity prediction
- 3D convolutional networks for protein-ligand binding affinity prediction
- Graph CNN and Xgboost for interface prediction
- Single interface prediction/binding site prediction

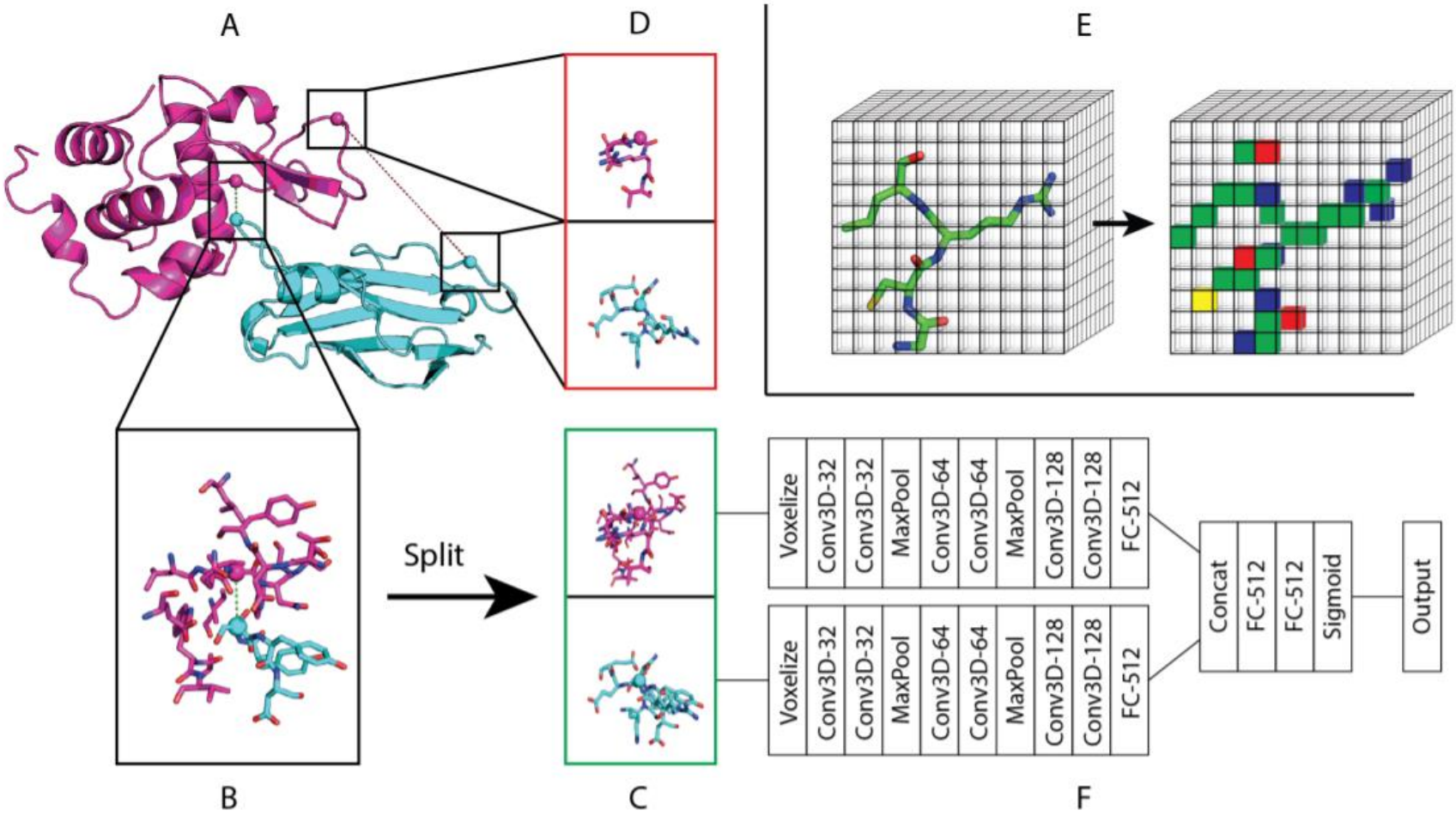
Dataset

- Docking Benchmark 5 (DB5)
  - DB5-train: training/validation set of 175 complexes
  - DB5-test: 55 complexes (the complexes added in the update from DB4 to DB5)
- Database of Interacting Protein Structures (DIPS)
  - ≥ 500 Å2 buried surface area, better than 3.5 Å resolution
  - <30% sequence identity

Dataset	# Binary Complexes	# Amino Acid Interactions
DB5	230	21,091
DIPS	42,826	5,767,093

Method

Rotation invariance: for training, randomly rotated, for testing, rotated 20 times and their results averaged



Performance Comparison

Method	CAUROC
NGF [4]	0.843 (0.851 +/- 0.010)
DTNN [35]	0.861 (0.861 +/- 0.004)
Node+Edge Average [23]	0.844 (0.850 +/- 0.004)
Order Dependent [23]	0.857 (0.864 +/- 0.006)
Node Average [23]	0.876 (0.877 +/- 0.005)
BIPSPI [24]	0.878 (0.878 +/- 0.003)
SASNet	**0.892 (0.885 +/- 0.009)**

Table 2: DB5-test CAUROC performance. For each method we report the CAUROC of the best replicate (as selected by DIPS validation loss for SASNet, and DB5-train loss for others) as well as mean and standard deviation of CAUROC across training seeds (see section 5.1). We note that while competing methods have used all available training data, due to computational limitations our SASNet model is trained on less than 3% of our dataset, suggesting an opportunity for further performance improvements.

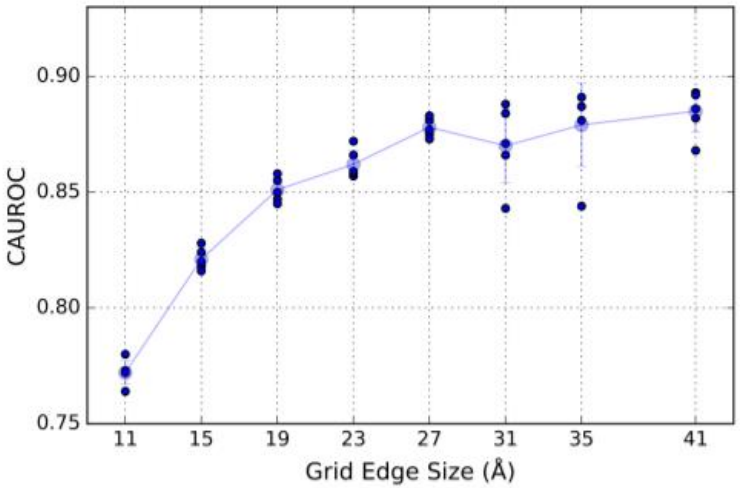
Performance Comparison Cont.

Dose the performance improvement is for DIPS training?

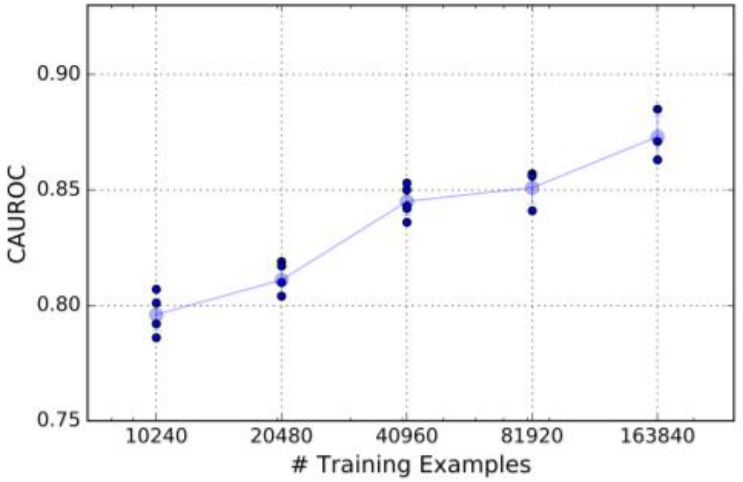
Method	DB5 Trained	DIPS Trained
Node Average [23]	0.876 (0.877 +/- 0.005)	0.712 (0.714 +/- 0.022)
BIPSPI [24]	0.878 (0.878 +/- 0.003)	0.836 (0.836 +/- 0.001)
SASNet	0.876 (0.864 +/- 0.037)	0.892 (0.885 +/- 0.009)

Table 3: DB5-test CAUROC for leading methods trained on DB5-train and DIPS. Competing methods with hand-engineered features experience a large drop in performance when trained on DIPS, despite its greater size. This indicates the assumptions embedded in their high-level features are not suited to the DIPS dataset. SASNet, on the other hand, increases in performance when trained on DIPS.

Ablation Study



(A) Grid size tests, dataset size fixed to 81920.



(B) Dataset size tests, grid size fixed to 23 Å.

Figure 3: SASNet benefits from large input sizes (A), and has potential for being further scaled (B). We plot the DB5-test CAUROC mean and standard deviation over five different training seeds.

Conclusion

- A method SASNet and a dataset DIPS
- My thoughts
  - Paper is well written, easy to read
  - Computation cost for random rotation in train and test?
  - Overlap of DIPS train and DB 5-test?