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

Raphael J. L. Townshend, Rishi Bedi, Patricia Suriana, Ron O. Dror

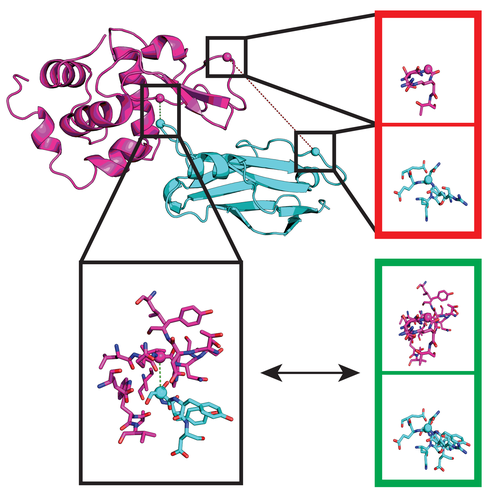
Department of Computer Science, Stanford University

Stanford Computer Science

Abstract

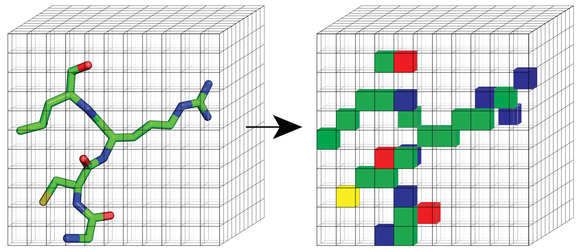
Despite an explosion in the number of experimentally determined, atomically detailed structures of biomolecules, many critical tasks in structural biology remain data-limited. Whether performance in such tasks can be improved by using large repositories of tangentially-related structural data remains an open question. To address this question, we focused on a central problem in biology: predicting how proteins interact with one another—that is, which surfaces of one protein bind to which surfaces of another protein. We built a training dataset that contains significant biases but is two orders of magnitude larger than those used previously. We found that these biases significantly degrade the performance of existing methods on gold-standard data. Hypothesizing that assumptions baked into the hand-crafted features on which these methods depend were the source of the problem, we developed the first end-to-end learning model for protein interface prediction, the Siamese Atomic Surfacelet Network (SASNet). Using only spatial coordinates and identities of atoms, SASNet outperforms state-of-the-art methods trained on gold-standard structural data, even when trained on only 3% of our new dataset.

Task: Protein Interface Prediction

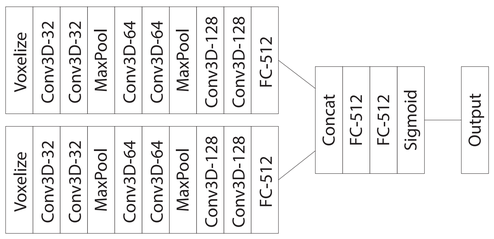


• Given a pair of amino acids from two different proteins, predict if they will interact (green pair) or not (red pair) when the two proteins bind

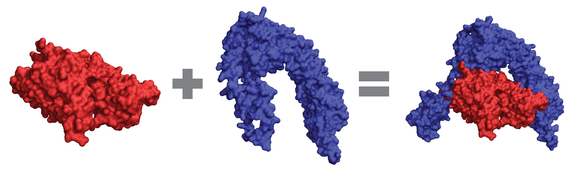
SASNet: Learn Directly on Atoms



• Store as 3D grid, encoding element type (C, N, O, S)

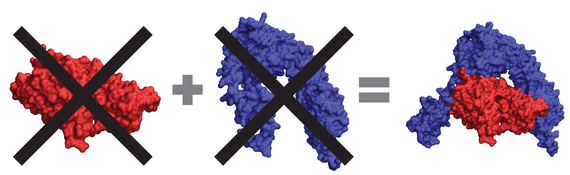


Small Amount of Gold-Standard Data



• Binding often causes large changes in shape
• Existing methods train on sets of ~200 proteins
• Know proteins' individual and bound shapes

DIPS: A New, Larger Training Dataset



• ~50k protein pairs where only the shape of proteins bound together is known—significant bias
• Overemphasis on shape complementarity
• 20x increase in overall dataset size

State-of-the-Art Interface Prediction

Method	CAUROC
NGF [Duvenaud *et al.* 2015]	0.843 (0.851 +/- 0.010)
DTNN [Schütt *et al.* 2017]	0.861 (0.861 +/- 0.004)
Node+Edge Average [Fout *et al.* 2017]	0.844 (0.850 +/- 0.004)
Order Dependent [Fout *et al.* 2017]	0.857 (0.864 +/- 0.006)
Node Average [Fout *et al.* 2017]	0.876 (0.877 +/- 0.005)
BIPSPI [Sanchez-Garcia *et al.* 2018]	0.878 (0.878 +/- 0.003)
SASNet	**0.892 (0.885 +/- 0.009)**

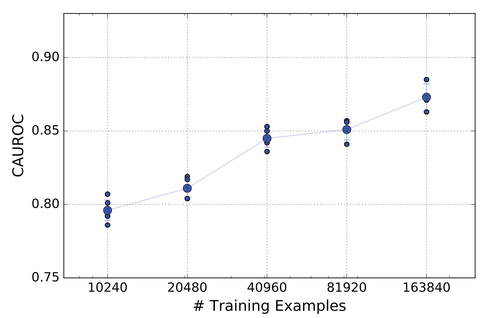
• No related structure information provided (e.g., sequence conservation, templates)
• SASNet currently trained on only 3% of DIPS

Existing Methods not Robust to DIPS

Method	DB5 Trained	DIPS Trained
Node Average [Fout *et al.* 2017]	0.876 (0.877 +/- 0.005)	0.712 (0.714 +/- 0.022)
BIPSPI [Sanchez-Garcia *et al.* 2018]	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)

• Previous state-of-the-art algorithms fall in performance when retrained on DIPS
• SASNet is competitive when trained on DB5

Accuracy Scales with Dataset Size



• Further improvements in accuracy possible through further use of DIPS dataset and mining of more related data