DeepAffinity: Interpretable Deep Learning of Compound-Protein Affinity through Unified Recurrent and Convolutional Neural Networks

MOSTAFA KARIMI[1,3] , DI WU[1], ZHANGYANG WANG[2], YANG SHEN[1,3]

[1]Department of Electrical and Computer Engineering, [2]Department of Computer Science and Engineering, [3]TEES-AgriLife Center for Bioinformatics and Genomic Systems Engineering, Texas A&M University, USA.

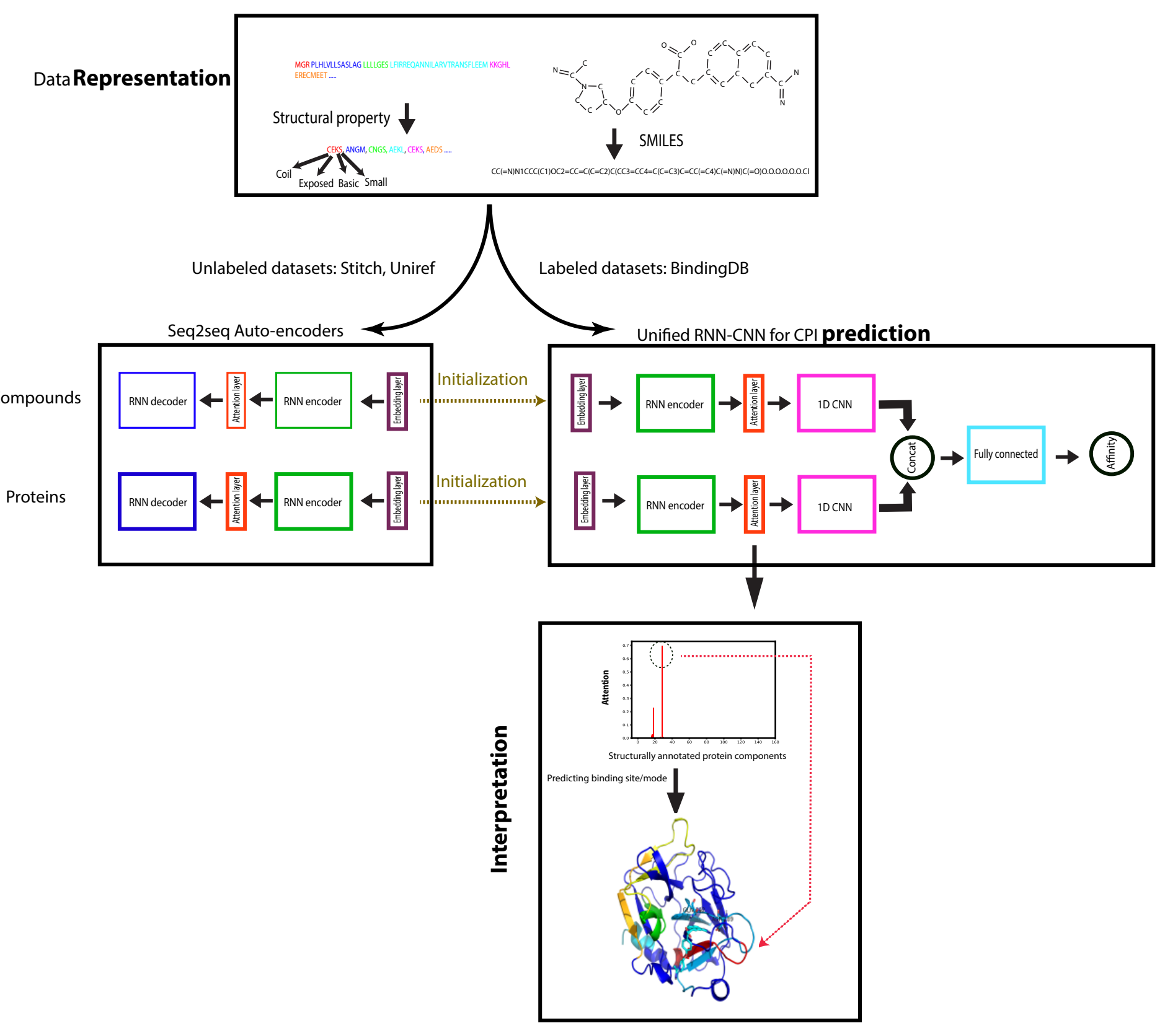
ABSTRACT

- **High-throughput drug discovery** through deep learning.
- **Novel representation** of structurally-annotated protein sequences.
- We present a **semi-supervised deep learning** model that unifies recurrent and convolutional neural networks to exploit both unlabeled and labeled data.
- **Transfer learning** for new protein classes with few labeled data.
- Embedded **attention mechanism** to gain **interpretability**.
- Our models **outperform** conventional options in achieving relative error in IC$_{50}$ within 5 to 10-fold.

Availability: https://github.com/Shen-Lab/DeepAffinity

METHODS

Overall scheme


Data Representation
MGR PLHLVLLSASLAG LLLLGES LFIRREQANNILARVTRANSFLEEM KKGHL ERECMEET ....
Structural property
CEKS, ANGM, CNGS, AEKL, CEKS, AEDS ....
Coil
Exposed
Basic
Small
SMILES
CC(=N)N1CCC(C1)OC2=CC=C(C=C2)C(CC3=CC4=C(C=C3)C=CC(=C4)C(=N)N)C(=O)O.O.O.O.O.O.Cl
Unlabeled datasets: Stitch, Uniref
Labeled datasets: BindingDB
Seq2seq Auto-encoders
Unified RNN-CNN for CPI prediction
Compounds
Proteins
RNN decoder
Attention layer
RNN encoder
Embedding layer
Initialization
1D CNN
Concat
Fully connected
Affinity
Interpretation
Attention
Structurally annotated protein components
Predicting binding site/mode

Data representation

- Compound: SMILES strings
- Protein: We developed Structural property sequence (SPS) based on predicted secondary structure elements (SSEs), solvent accessibility, physicochemical characteristics and lengths of SSEs.

Semi-supervised deep learning model

- Unsupervised learning: Seq2seq autoencoder models with attention mechanism to exploit abundant unlabeled data.
- Supervised learning: Unified recurrent and convolutional neural networks with attention mechanism are jointly trained starting with pre-trained encoder part of seq2seq
- Interpretability through the embedded attention mechanism
- Deep transfer learning

RESULT

Novel Representations v.s. Baseline Pfam/Fingerprints

	Baseline representations			Novel representations		
	Ridge	Lasso	RF	Ridge	Lasso	RF
Training	1.16 (0.60)	1.16 (0.60)	0.76 (0.86)	1.23 (0.54)	1.22 (0.55)	**0.63** (0.91)
Testing	1.16 (0.60)	1.16 (0.60)	**0.91** (0.78)	1.23 (0.54)	1.22 (0.55)	**0.91** (0.78)
ER	1.43 (0.30)	1.43 (0.30)	1.44 (0.37)	1.46 (0.18)	1.48 (0.18)	**1.41** (0.26)
Ion Channel	1.32 (0.22)	1.34 (0.20)	1.30 (0.22)	1.26 (0.23)	1.32 (0.17)	**1.24** (0.30)
GPCR	**1.28** (0.22)	1.30 (0.22)	1.32 (0.28)	1.34 (0.20)	1.37 (0.17)	1.40 (0.25)
Tyrosine Kinase	**1.16** (0.38)	1.16 (0.38)	1.18 (0.42)	1.50 (0.11)	1.51 (0.10)	1.58 (0.11)
Time (core hours)	3.5	7.4	1239.8	0.47	2.78	668.7
Memory (GB)	7.6	7.6	8.3	7.3	7.3	6.3

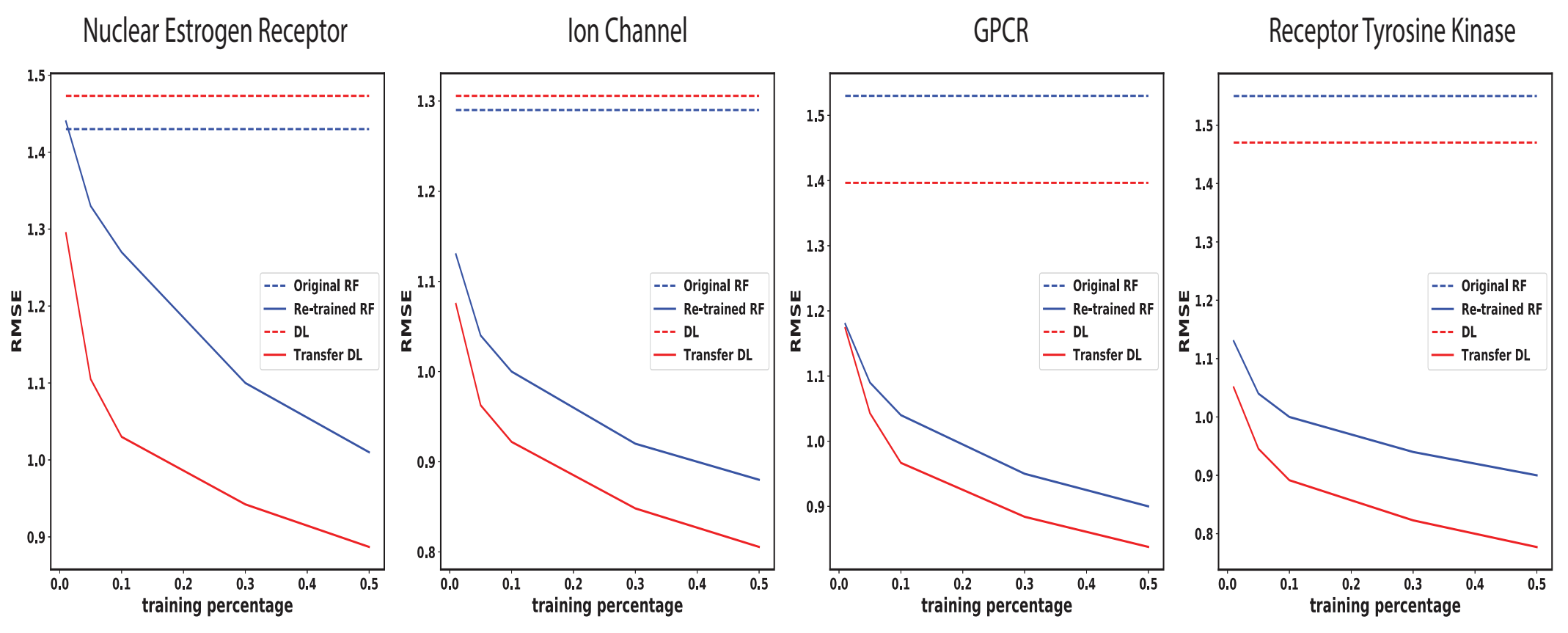
SPS representation saves 40% training time and 20% memory while achieving the similar or better performances over test set and lowered RMSE for generalization sets

Shallow Models v.s. Deep Models

	RF	Separate RNN-CNN Models			Unified RNN-CNN Models		
		single	parameter ensemble	parameter+NN ensemble	single	parameter ensemble	parameter+NN ensemble
Training	0.63 (0.91)	0.68 (0.88)	0.67 (0.90)	0.68 (0.89)	0.47 (0.94)	0.45 (0.95)	**0.44** (0.95)
Testing	0.91 (0.78)	0.94 (0.76)	0.92 (0.77)	0.90 (0.79)	0.78 (0.84)	0.77 (0.84)	**0.73** (0.86)
Generalization – ER	**1.41** (0.26)	1.45 (0.24)	1.44 (0.26)	1.43 (0.28)	1.53 (0.16)	1.52 (0.19)	1.46 (0.30)
Generalization – Ion Channel	**1.24** (0.30)	1.36 (0.18)	1.33 (0.18)	1.29 (0.25)	1.34 (0.17)	1.33 (0.18)	1.30 (0.18)
Generalization – GPCR	1.40 (0.25)	1.44 (0.19)	1.41 (0.20)	1.37 (0.23)	1.40 (0.24)	1.40 (0.24)	**1.36** (0.30)
Generalization – Tyrosine Kinase	1.58 (0.11)	1.66 (0.09)	1.62 (0.10)	1.54 (0.12)	1.24 (0.39)	1.25 (0.38)	**1.23** (0.42)

Unified RNN-CNN models outperform random forest and separate RNN-CNN models. Averaging ensembles of models lower RMSE by reducing the variance of model.

Deep transfer learning for new classes of protein targets



Deep transfer learning models increasingly improved the predictive performance compared to the original deep learning models, given increasing amount of labeled data. Even few labeled data is enough for significant improvement.

Predicting target selectivity of drugs

Protein-tyrosine phosphatase (PTP) family:

	Baseline rep. + RF			Novel rep. + RF			Novel rep. + DL (sep. attn.)			Novel rep. + DL (joint attn.)		
Protein	Comp1	Comp2	Comp3	Comp1	Comp2	Comp3	Comp1	Comp2	Comp3	Comp1	Comp2	Comp3
PTP1B	4.15	3.87	5.17	6.70	6.55	6.71	3.76	**3.84**	**3.92**	2.84	**4.10**	**4.04**
PTPRA	4.15	3.87	5.17	6.29	6.59	6.27	2.73	2.90	3.44	2.39	2.62	2.12
PTPRC	4.15	3.87	5.17	**6.86**	6.73	**6.87**	3.37	3.25	3.19	3.36	3.49	2.97
PTPRE	4.15	3.87	5.17	6.79	6.68	6.81	**3.83**	3.75	3.85	2.75	2.93	2.61
SHP1	4.15	3.87	5.17	6.71	**6.74**	6.73	3.37	3.38	3.89	**3.42**	3.52	3.22

- Random forest using baseline representations cannot tell target specificity within the PTP family as the proteins' Pfam descriptions are almost indistinguishable.
- Using novel representations, random forest correctly predicted PTP1B selectivity for compounds 1 and 3 but not compound 2, whereas unified RNN-CNN models correctly did so for all three compounds.

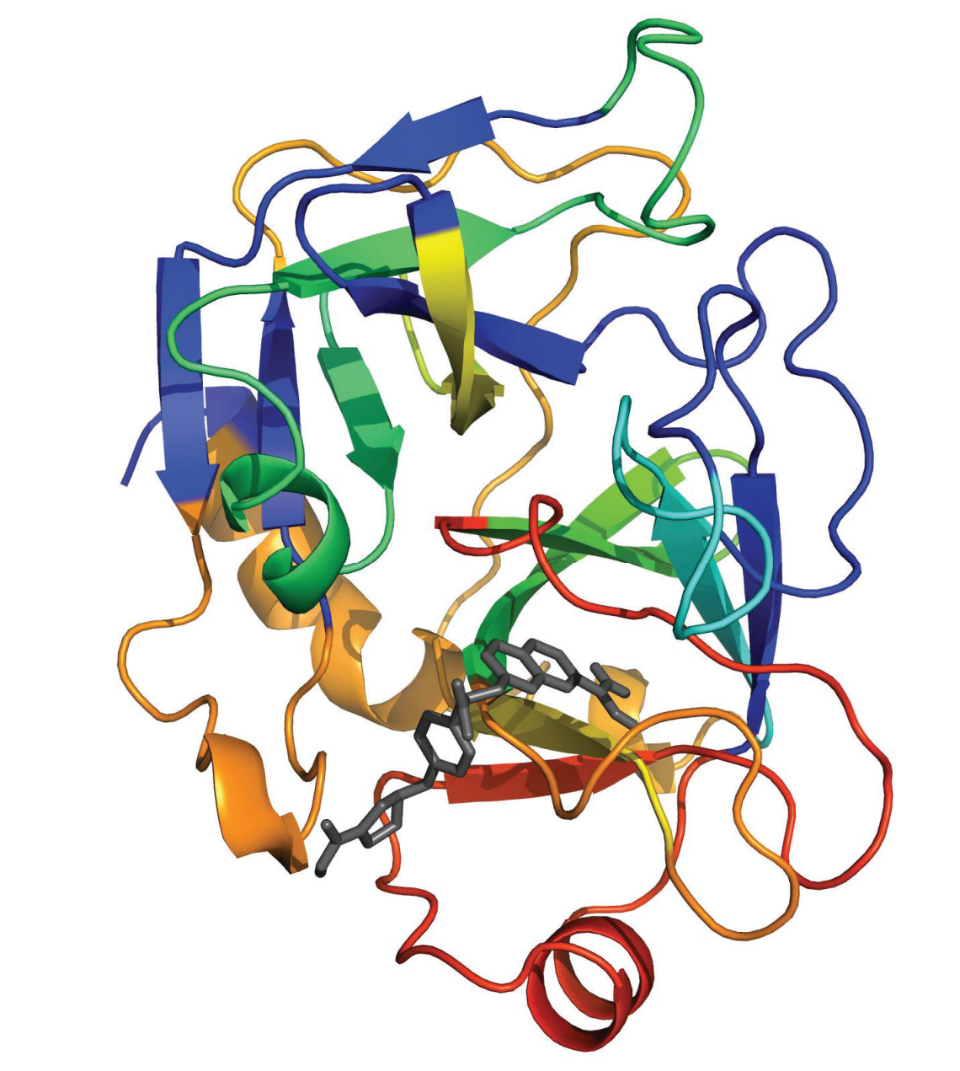
INTERPRETABILITY

How do the compound-protein pairs interact?

		Number of SSEs		Top 10% (4) SSEs predicted as binding site by joint attn.			
Target–Drug Pair	PDB ID	total	binding site	# of TP	Enrichment	Best rank	P value
Human COX2–rofecoxib	5KIR	40	6	1	1.68	4	1.1e-2
Human PTP1B–OBA	1C85	34	5	1	1.70	1	1.1e-10
Human factor Xa–DX9065	1FAX	31	4	3	5.81	2	2.2e-16

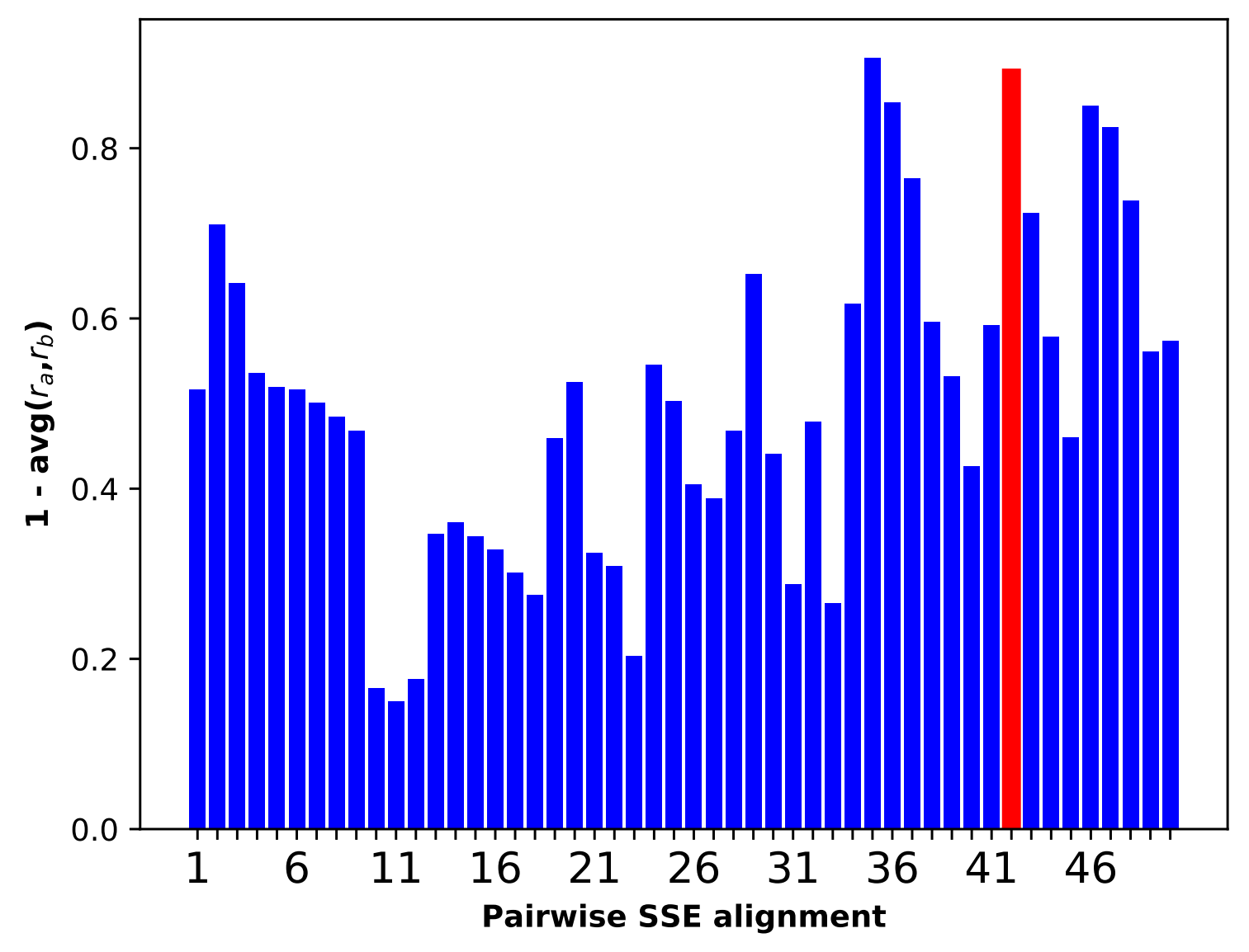
Compared to randomly ranking the SSEs, our approach can enrich binding site prediction by 1.6∼2.0 fold for the three CPIs.

Human factor Xa–DX-9065a interaction:



The binding site was correctly predicted with a high rank 2. And the SSE ranked first, a false positive, was its immediate neighbor in sequence.

How are targets selectively interacted?



- Position 192 has been identified as the source of specificity: it is a charge-neutral polar glutamine (Gln192) in Xa but a negatively-charged glutamate (Glu192) in thrombin.
- The ground-truth segment (red) was ranked the 2nd among 50 segments.

REFERENCES

[1] Mostafa Karimi, Di Wu, Zhangyang Wang, Yang Shen. "DeepAffinity: Interpretable Deep Learning of Compound-Protein Affinity through Unified Recurrent and Convolutional Neural Networks", Bioinformatics 35(18), 3329-3338.

ACKNOWLEDGEMENTS

This project is in part supported by the National Institute of General Medical Sciences of the National Institutes of Health (R35GM124952 to YS) and the Defense Advanced Research Projects Agency (FA8750-18-2-0027 to ZW). Most of the computing time is provided by the Texas A&M High Performance Research Computing