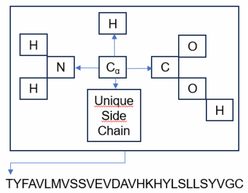
GPR-187 Deep Learning Models for Protein-Protein Binding Affinity Prediction

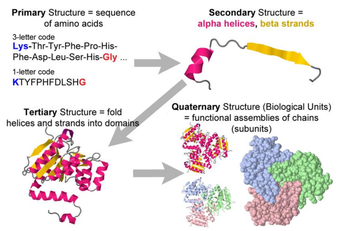
Abstract

Binding affinity (BA) prediction is important for drug discovery and protein engineering. This paper presents the development and comparative analysis of two deep learning models, a convolutional neural network (CNN) and a transformer model. The CNN model captures local sequence features effectively, while the Transformer model leverages self-attention mechanisms to learn long-range dependencies within the sequences. Protein sequences are the inputs for the models. The sequences are processed using various encoders. The predicted outputs are Gibbs free energy changes, a key indicator of binding affinity. From this study, both the CNN and transformer models can achieve the same level of accuracy under different conditions. This study emphasizes the potential of advanced deep learning architectures to enhance the predictive strengths of binding affinity models.

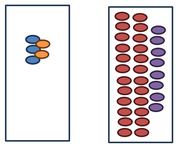
Introduction



Protein - Long chains of Amino Acids
Amino acid:
- An alpha carbon (C_α)
- An amino group (NH_2)
- A carboxyl group (COOH)
- A hydrogen atom (H)
- A side chain



Primary Structure – Sequence of Amino Acids
Secondary Structure - Alpha Helices and Beta Sheets (Portions of a chain)
Tertiary Structure - overall three-dimensional shape of the protein (One full chains)
Quaternary structure - the arrangement of multiple polypeptide chains in a protein (All chains)



Protein-Protein Interactions (PPIs) are physical contact between two or more protein molecules.
Binding affinity (BA) - the strength of interaction
Kd (dissociation constant):
- Lower value – stronger binding
- Higher value – weaker binding

Dataset

$$\Delta G = -RT * \ln(K_d)$$

- K_d is collected from PDBbind, which is based on Protein Data Bank
- ΔG is the Gibbs free energy change
- R is the ideal gas constant, which is approximately 0.0019858775 kcal * mol^{-1} * K^{-1}
- T is the temperature, which is 298.15 Kelvin
- K_d is the value determined experimentally. The unit is M (molar), or μm, nm, and so on

	A	B	C	D	E	F	G	H	I	J
1	Protein_ID	AA_Log	AC_Log	AP_Log	CC_Log	CP_Log	PP_Log	ba_val	sequence	
2	3SGB	2.484907	2.397895	3.178054	1.098612	2.302585	2.397895	-14.65	ISGGDAIYSSTGRCSLGF	
3	2TGP	2.197225	2.70805	3.218876	1.098612	2.995732	2.302585	-7.66	VDDDDKIVGGYTCGANT	
4	2PTC	2.197225	2.70805	3.258097	1.098612	2.995732	2.197225	-18.03	IVGGYTCGANTVPYQVS	
5	2SNI	3.218876	2.302585	3.433987	1.098612	2.564949	2.564949	-15.95	AQSVPYGVSQIKAPALH	
6	1ATN	3.044522	2.484907	3.367296	1.609438	2.079442	2.70805	-12.74	XDEDETTALVCDNGSGL	
7	1GLA	2.197225	2.302585	2.639057	1.386294	2.484907	1.098612	-9.2	TEKKYIVALDQGTTSSRA	
8	2PCC	3.044522	2.70805	3.258097	3.583519	3.044522	2.772589	-7.9	MITTPLVHVASVEKGRSY	

Property Features – Polar, Apolar, Charged

Models

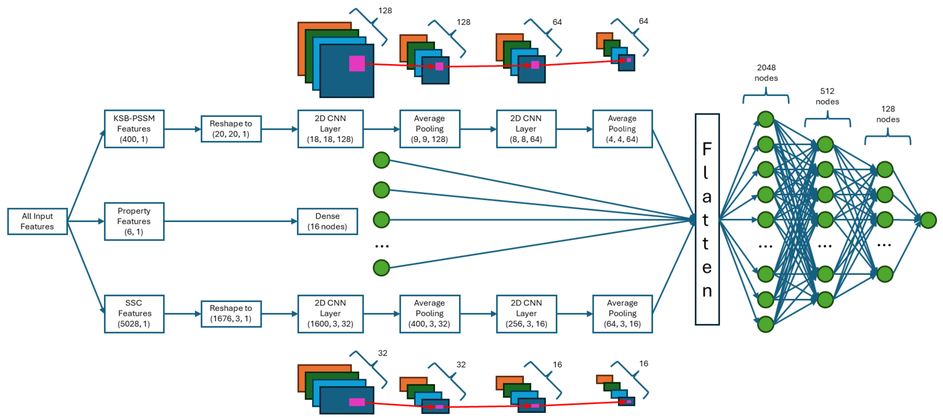


Figure 1: Convolutional Neural Network

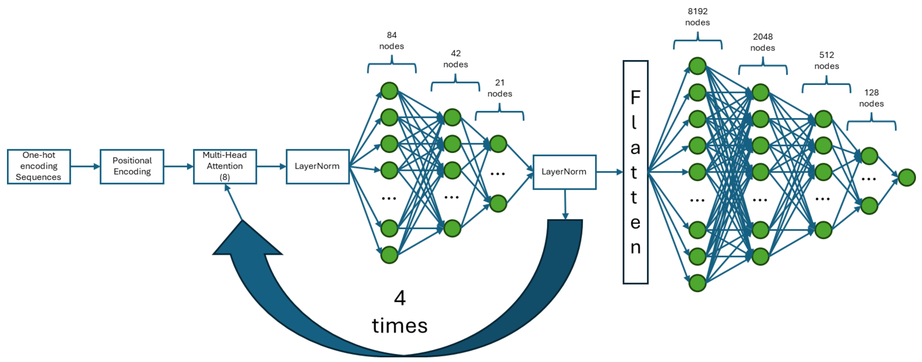


Figure 2: Transformer

Results

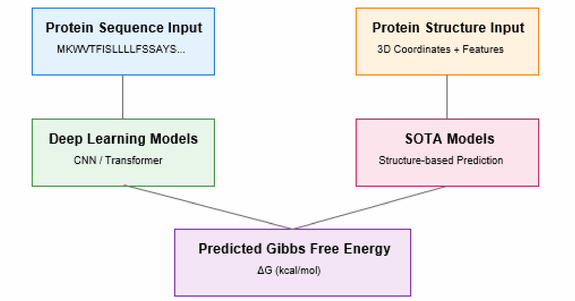
Data	Features	Models	MAE	MSE	RMSE	MAPE
Less*	SSC	CNN+DNN	1.70	4.62	/	/
Less*	one-hot	Transformer	1.84	5.27	2.29	0.2124
Less*	ProteinBERT	CNN+DNN	1.84	5.27	2.29	/
Full	KSB-PSSM	DNN	1.73	4.88	2.21	0.1895
Full	KSB-PSSM, Property	CNN+DNN	1.69	4.68	2.16	0.1952
Full	KSB-PSSM, Property, SSC	CNN+DNN	1.61	4.45	2.11	0.1936
Less*	one-hot	Transformer	1.62	4.29	2.07	0.1888
Full	one-hot	Transformer	1.79	5.30	2.30	0.21

Less*: All protein sequence lengths less than 676

Table 1: Deep Learning Models and Results

Model	Features	MAE	RMSE
Proposed CNN	Sequence and Property	1.61	2.11
Proposed Transformer	Sequence Alone	1.62	2.07
ProBAN (CNN)	Structure and Property	1.60	1.95

Table 2: Comparison with SOTA model



Comparable performance achieved using sequence-based approach

Figure 3: Sequence vs Structure based Approaches

Conclusions

The CNN model and the transformer model have their own advantages.
For the **CNN** model, it can handle full data without sacrificing performance. However, it takes much more time to preprocess the features from the protein sequences.
The **transformer** model can achieve the same level of accuracy as the CNN model with no big predictive errors for each protein. However, it requires the model to run on less data, which removes some unusually long protein sequences.

Acknowledgments

Kazi Nasif, Nisha Bagdwal, Heng Quan
Dr. Bobin Deng and **Dr. Shuteng Niu**

Contact Information

Lingtao Chen: lchen25@students.kennesaw.edu
Chloe Yixin Xie: yxie11@kennesaw.edu

Publication

This paper has been presented at IEEE **ICTAI 2024** and will be published at IEEE conference proceedings.

References

[1] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of aconvolutional neural network," in 2017 international conference onengineering and technology (ICET). Ieee, Conference Proceedings,pp. 1–6.
[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez,. Kaiser, and I. Polosukhin, "Attention is all you need," Advances inneural information processing systems, vol. 30, 2017.
[3] E. A. Bogdanova and V. N. Novoseletsky, "Proban: Neural networkalgorithm for predicting binding affinity in protein–protein complexes,"Proteins: Structure, Function, and Bioinformatics, 2024.

KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING

Author(s): Lingtao Chen
Advisors(s): Chloe Yixin Xie