Protein Language Models Enable Accurate Cryptic Ligand-Binding Pocket Prediction

David Bloore[1], Joseph Kim[1,2], Karan Kapoor[1], Eric Chen[1], Kaifu Gao[1], Mengdi Wang[2], Ming-Hong Hao[1]. [1] Ensem Therapeutics, 888 Winter St, Waltham, MA 02451.
[2]Princeton University, 41 Olden St. Princeton, New Jersey 08544

Summary

Background: Identification of ligand binding pockets is a critical step in structure-based drug design. Unlike conventional pockets, which are apparent on protein targets regardless of ligand binding status, cryptic pockets only become obvious when a ligand binds. Uncovering cryptic pockets offers untapped opportunities for drug discovery.

Previous work: Many publications have reported identification of conventional ligand binding pockets using tools like LIGSite[1], Fpocket[2], and P2Rank[3]. For cryptic pocket identification, CryptoSite[4] and PocketMiner[5] (PM) are machine learning-based methods that have reported encouraging performance. In particular, PM leveraged a deep-learning model of protein dynamic structures to achieve superior performance in predicting protein cryptic pockets.

Objective: Inspired by insights into protein structure & function and protein-ligand interactions gleaned by protein language models (PLMs)[6,7], this study aimed to investigate the application of PLMs to the prediction of cryptic ligand binding pockets.

Results: We developed a PLM-based prediction model that demonstrated high accuracy in predicting ligand binding pockets, including cryptic ones. Trained on labeled pocket data from PDBBind v2020, our Efficient Sequence-based Prediction (ESP) model achieved high accuracy with Area Under the ROC Curve (AUC) equal to 0.93, outperforming PM, which achieved AUC = 0.87 on the same cryptic pocket testing set.

Datasets

Protein pocket labels were generated for protein-ligand complex structures in PDBBind[8]. A total of 17,986 protein sequences were labeled. A positive label was assigned to a residue when it was within 6 Å of a bound ligand in a protein-ligand complex structure. Negative labels were assigned to all other residues. This resulted in a dataset with 495,482 positively labeled pocket residues and 4,759,440 negatively labeled non-pocket residues with an average protein sequence length of 292 residues. The PM test dataset included 35 protein structures comprising 563 cryptic pocket residues and 1283 residues that do not form cryptic pockets.

Cryptic ligand binding pocket

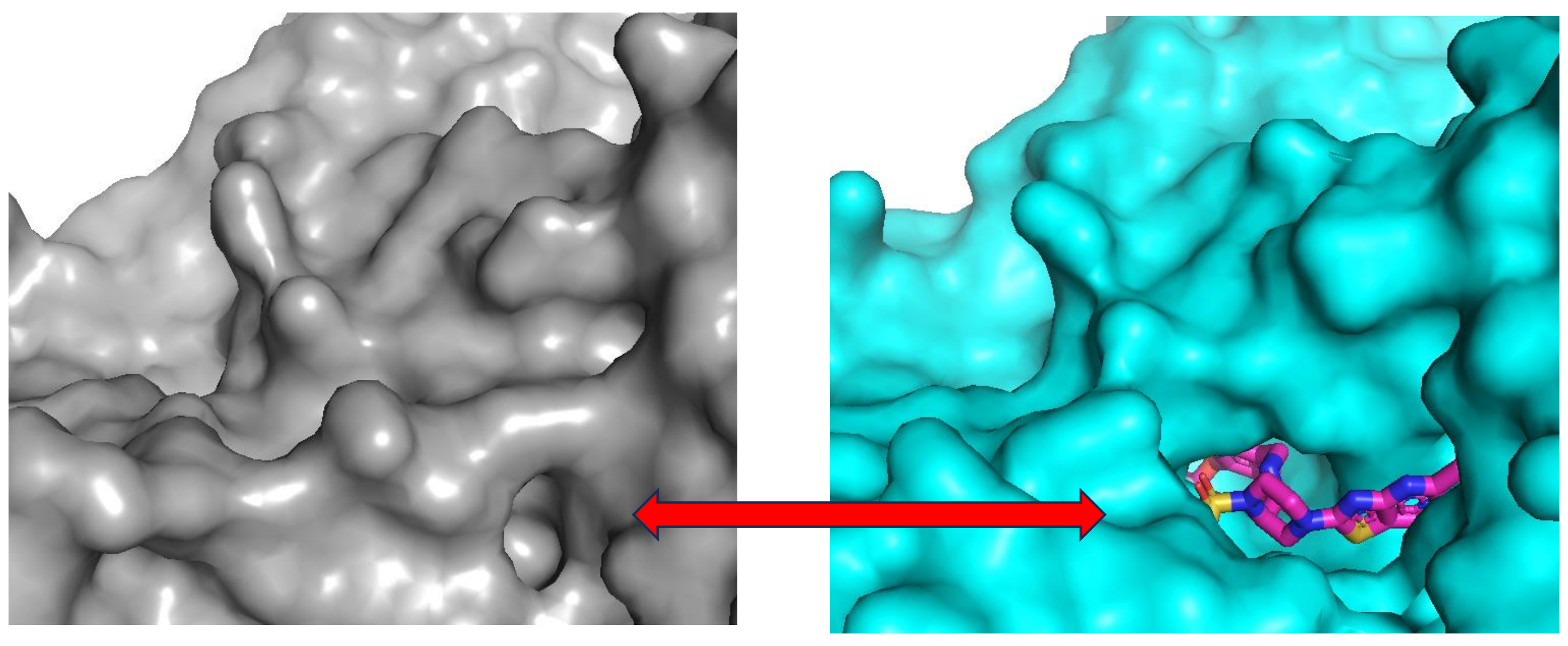


Figure 1. (A) Apo (i.e., ligand-free) structure of NS5B RNA polymerase (1QUV). (B) Ligand-bound state of the same protein crystal structure (3VPS) wherein the cryptic pocket becomes obvious.

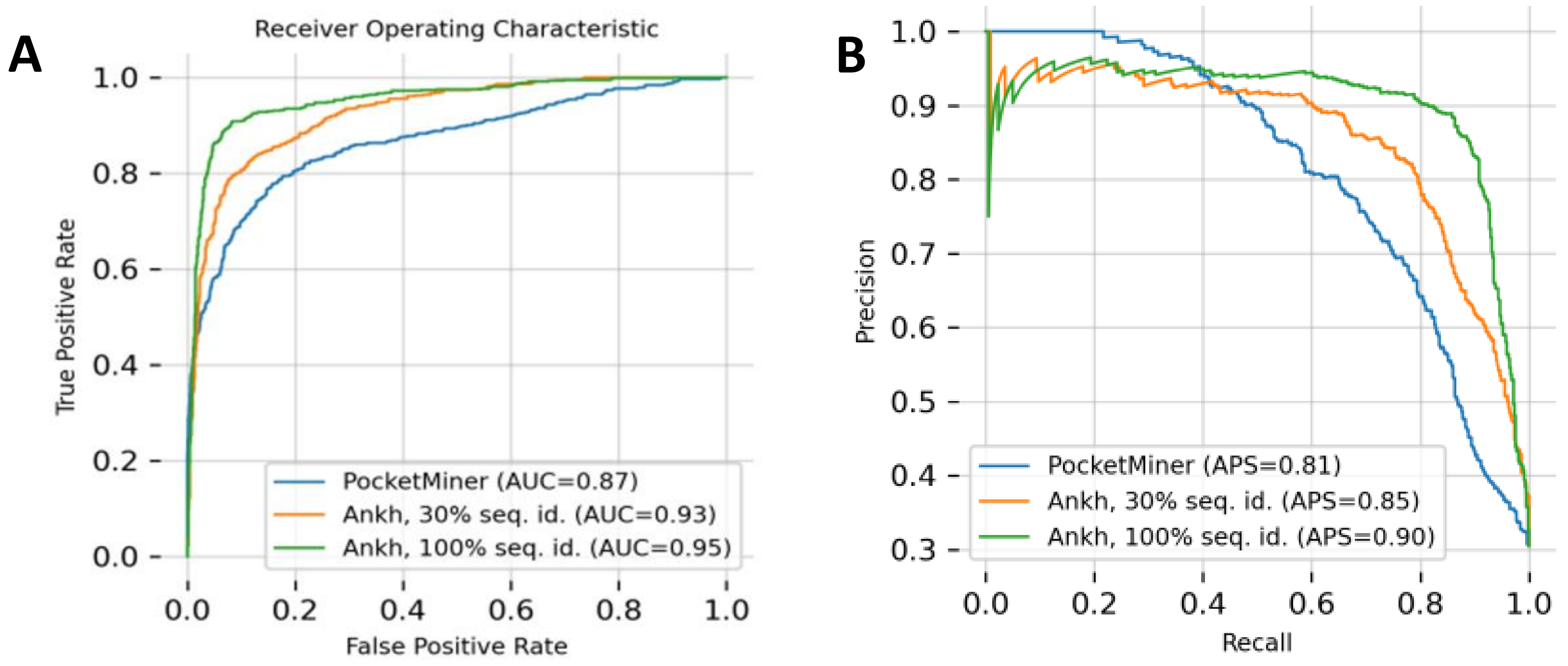


Figure 2: (A) Receiver-operator curve and (B) Precision Recall curve for the PM model and ESP model based on Ankh. With a 30% sequence identify cutoff, ESP achieved higher AUC and average precision score (APS) than PM.

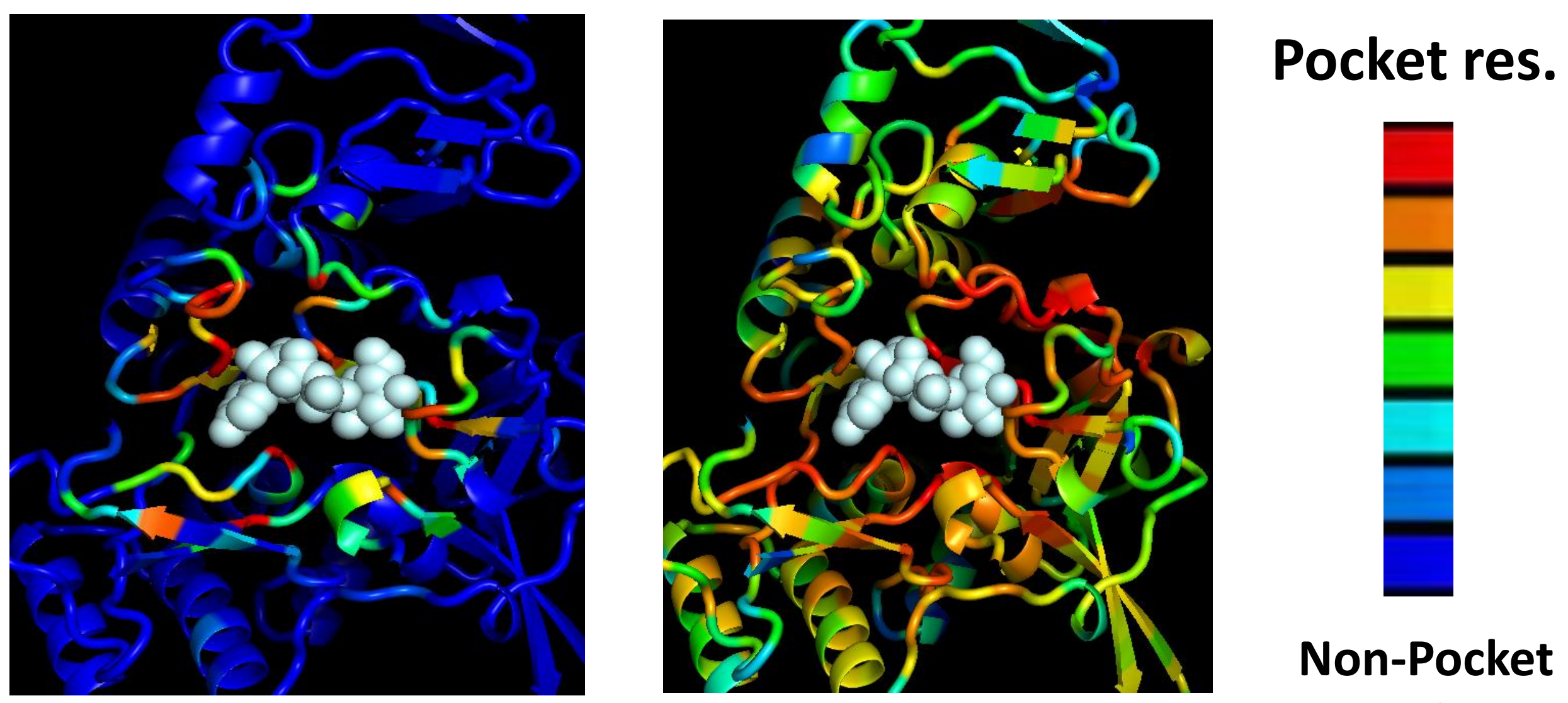


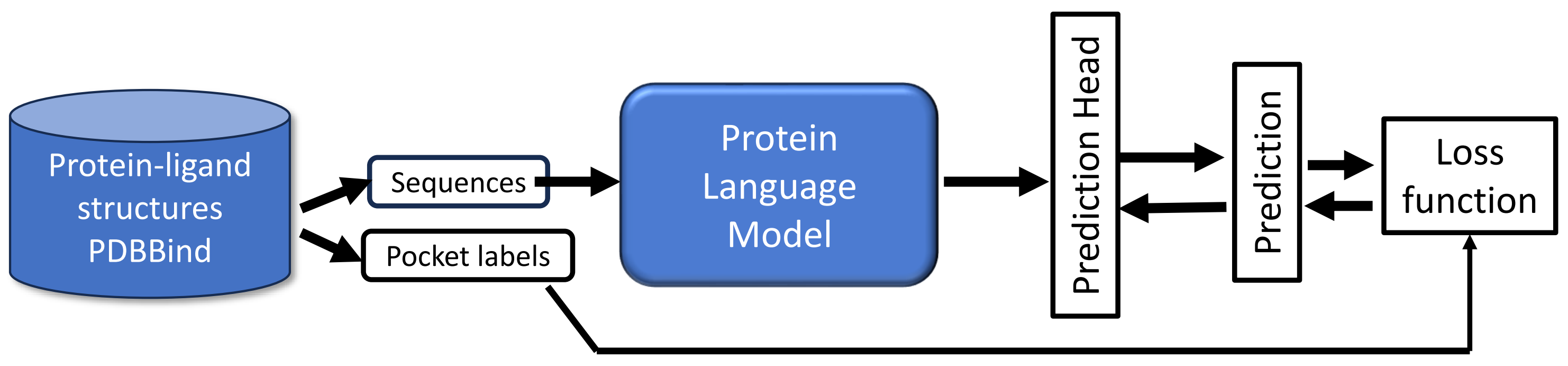
Figure 4: (A) ESP-predicted and (B) PM-predicted cryptic binding pocket for lactamase (PDBID 3GQZ) superimposed with bound ligand (crystal structure).

Conclusion

This study highlights that protein language models such as Ankh-Large and ESM-2 15B enabled more accurate prediction of cryptic ligand binding pockets on protein targets compared to earlier deep learning models such as PocketMiner. Our results suggest that the PLMs might have learned residue-level information, including the location of cryptic pockets, through unsupervised training. To further improve model accuracy, PLM-based sequence embeddings with MD-generated structural features could be investigated.

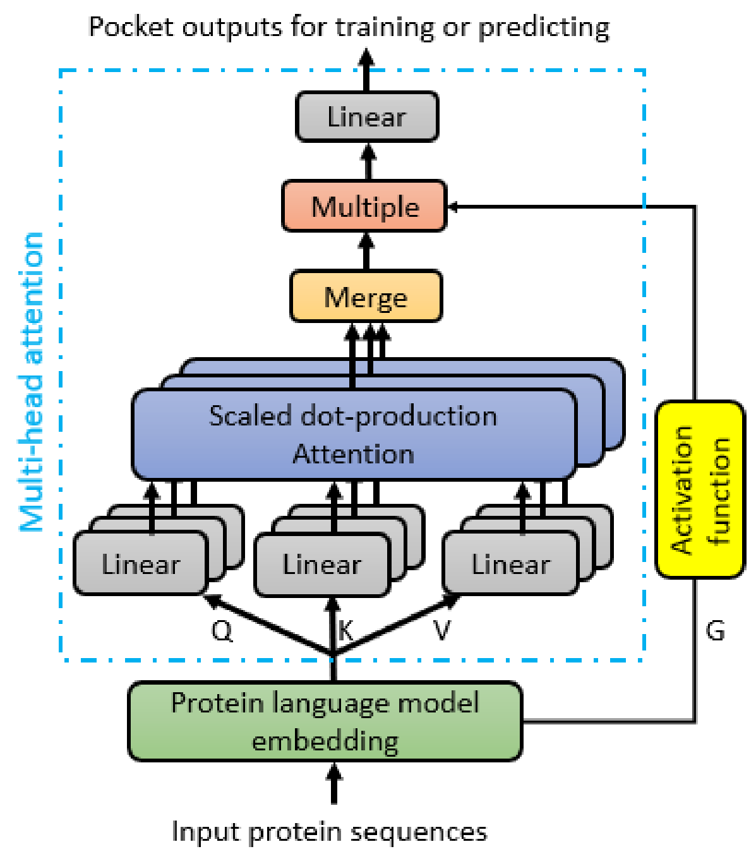
Method

Training ESP pocket-prediction model



Architecture

Multi-Head Attention (MHA) prediction head


Pocket outputs for training or predicting
Linear
Multiple
Merge
Scaled dot-production Attention
Linear
Linear
Linear
Q
K
V
G
Activation function
Multi-head attention
Protein language model embedding
Input protein sequences

Algorithms

```
Multi_head_attention:
def MHA_model(sequence, num_attn_head):
    Q, K, V, G = linear(sequence)
    reshape Q, K, V, G to (seq_length, num_attn_head, head_dim)
    O=(Q·K)·V
    merge by reshaping Q, K, V, G to (seq_length, embedding_dim)
    O=O∗activation_fn(G)
    return linear(O)
Training:
def train(training_set, num_attn_head, epoch):
    for e in range(epoch):
        predict=MHA_model(training_set, num_attn_head)
        loss=cross_entropy_loss (label, predict)
        backpropagation to update model weights
    save new MHA_model
Predicting:
def predict(sequences):
    load pretrained MHA_model
    return pretrained_MHA_model(sequences)
```

Results

Table 1: APS and AUC results by different prediction algorithms using Anhk-Large embeddings. Training samples with greater than 30% sequence identity with testing sets were removed.

PLM: Ankh-Large	PDBBind only		w/ SSP	
Architecture	APS	AUC	APS	AUC
PocketMiner (PM)	0.81	0.87	N/A	N/A
LR	0.805	0.889	0.805	0.889
MLP 16	0.805	0.890	0.801	0.890
MLP 64	**0.808**	**0.890**	0.807	0.890
MLP 256	0.806	0.890	0.807	0.890
MLP 1024	0.808	0.890	0.804	0.889
MHA 4 no CLS	0.845	0.911	0.755	0.891
MHA 8 no CLS	0.852	**0.916**	**0.853**	0.911
MHA 16 no CLS	0.820	0.897	0.841	0.908
MHA 4	0.802	0.906	0.832	0.907
MHA 8	0.821	0.908	0.849	0.897
MHA 16	**0.854**	**0.926**	0.840	0.902

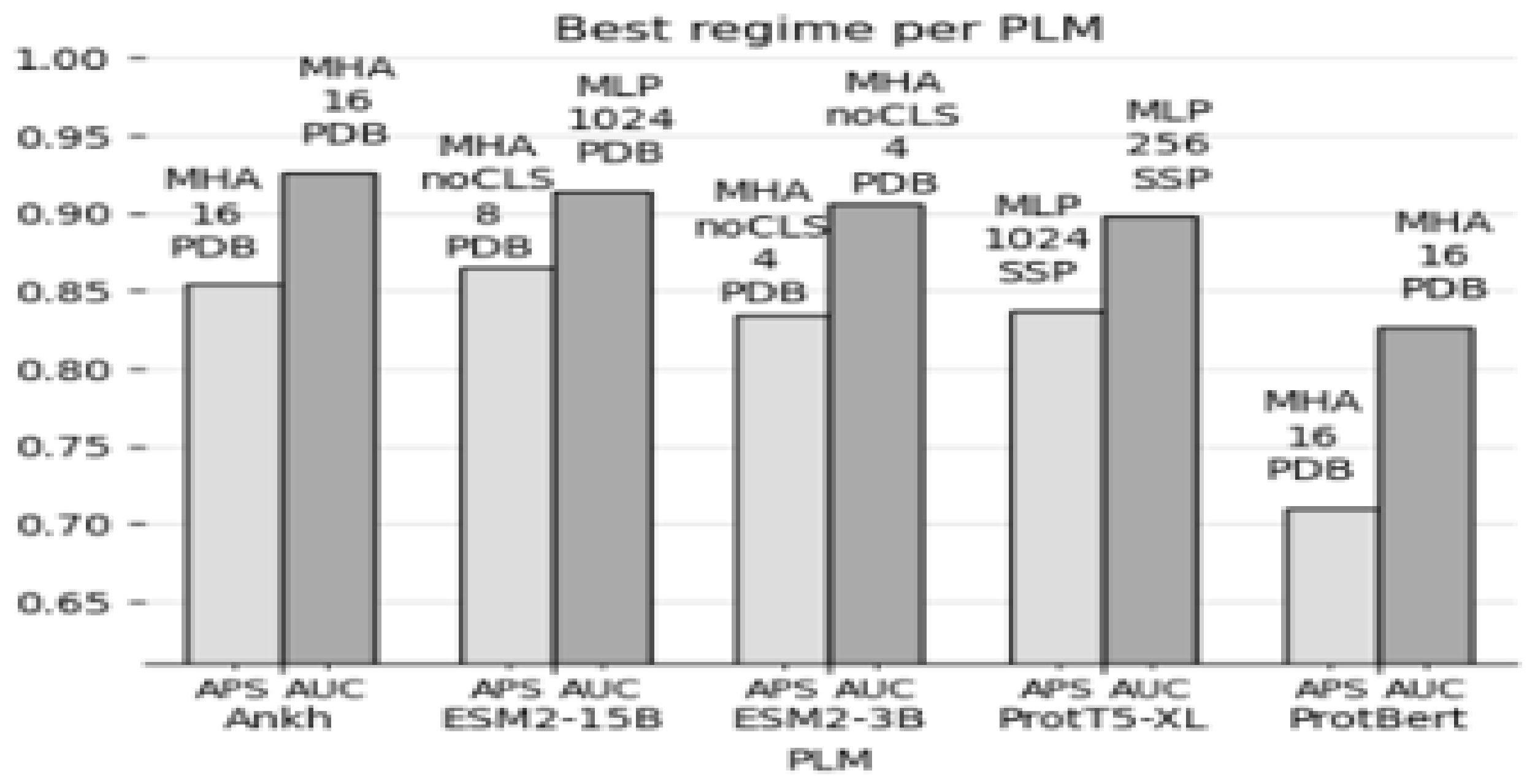
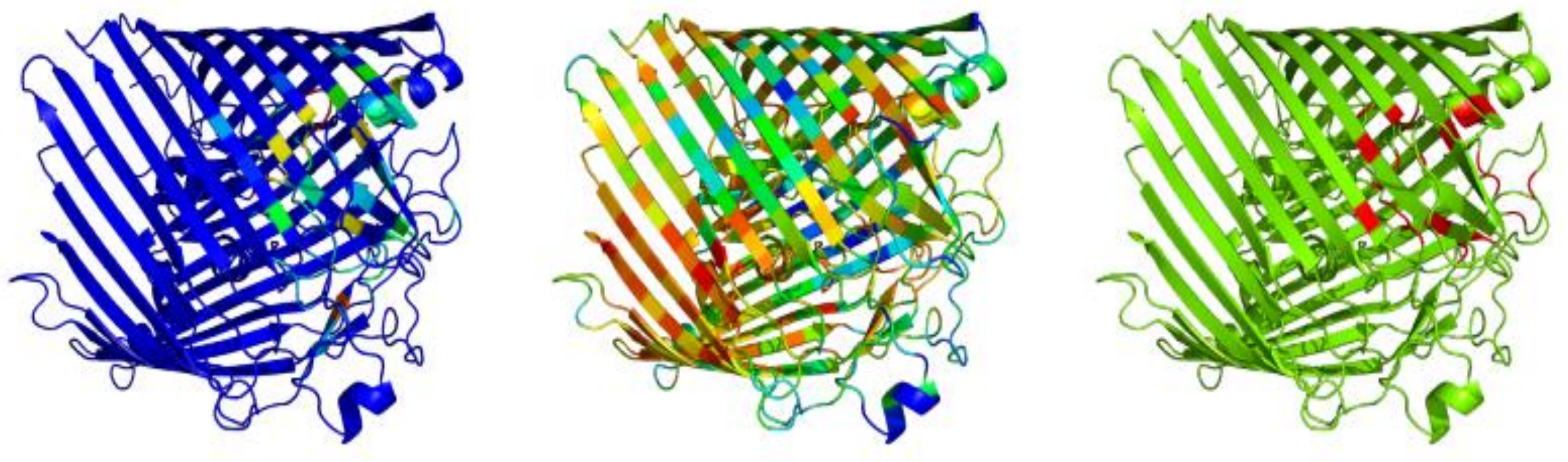


Figure 3: Comparison of the performance of ESP model prediction using different PLM embeddings, including Ankh, ESM2-15B, ESM-3B, ProtT5-Xl, and ProtBERT.



(a) 1KMO ESP inference. (b) 1KMO PM inference. (c) 1KMO PM labels.

Figure 5: Model prediction on the transporter protein FecA. (a) shows ESP inference results, (b) shows PM inference results, and (c) shows the labeled pocket residues colored in red.

References

1. Hendlich et al., J. M. Graph Model, 1997
2. Le Guilloux et al., BMC Bioinformatics 2009
3. R. Krivak et al., J. Chemoinfo., 2018
4. P. Cimermancic et al., J. Mol. Biol, 2016
5. M. Meller et al., Nat. Commu.,2023.
6. A. Rives et al., PNAS, 2019.
7. A. Elnaggar et al., arXiv, 2023.
8. M. Su et al., JCIM, 5, 895, 2019.