DBP-Finder: Enhanced Identification of DNA-Binding Proteins Using Fine-Tuned Protein Language Models

Alexander Gavrilenko[1], Elizaveta Shaburova[1], Denis Antonets[1], Yury Vyatkin[1], Vasily Ramensky[123]

1 Institute for Artificial Intelligence, Lomonosov Moscow State University, Moscow, Russia; 2 National Medical Research Center for Therapy and Preventive Medicine of the Ministry of Healthcare of Russian Federation, Moscow, Russia; 3 Department of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia

Objective

This study aims to identify DNA-binding proteins (DBPs) using transfer learning on pretrained protein language models (PLMs) (http://dx.doi.org/10.1101/2024.02.05.578959). By leveraging PLMs for informative sequence representations, we will develop a predictive algorithm for DBP identification, addressing the need for scalable, automated prediction methods amidst limited experimental annotations and increasing genomic data.

Methods

Training set construction

We sourced high-quality, manually annotated, non-redundant protein sequences from UniProtKB/Swiss-Prot. DBPs were selected using the DNA-binding GO term, while Non-DBPs excluded nucleic acid-binding GO terms per QuickGO (https://doi.org/10.1093/bioinformatics/btp536).
Sequences shorter than 50 or longer than 1,024 amino acids and those with undefined amino acids “X” were excluded. The training set balanced 34,936 DBPs with 34,936 Non-DBPs.

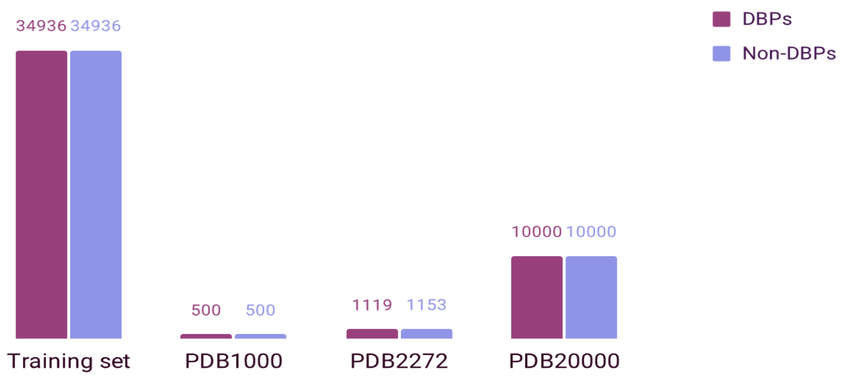
Testing sets

Three test datasets were used as benchmarks to facilitate direct comparison with previous studies:

- **PDB2272** dataset from Du, Diao, Liu, & Li (2019) (https://doi.org/10.1021/acs.jproteome.9b00226)
- **PDB20000** and **PDB1000** datasets from Ma (2019) (http://dx.doi.org/10.17504/protocols.io.2rdgd26)

Protein sequences in these datasets originated from UniProtKB and had GO annotations assigned by UniProt.

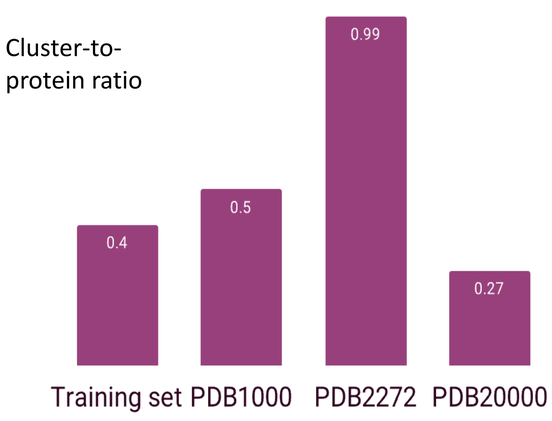
Counts of DBPs and Non-DBPs in training and testing sets

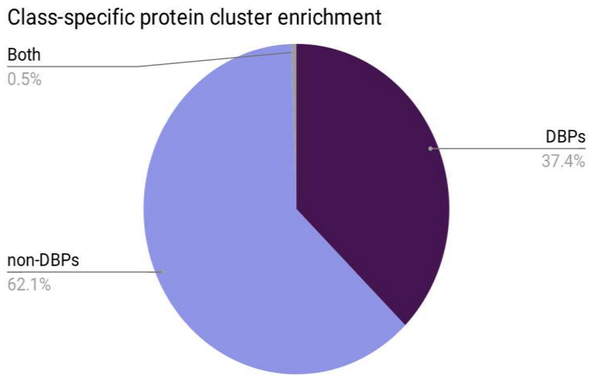


Clustering

We used MMseqs2 (https://doi.org/10.1038/nbt.3988) with a 50% sequence identity threshold to filter out homologous sequences, preventing data leakage and ensuring fair model evaluation. We also calculated the cluster-to-protein ratio and counted DBPs and Non-DBPs within each cluster.

Cluster-to-Protein Ratio: Analyzing Protein Clustering Distribution





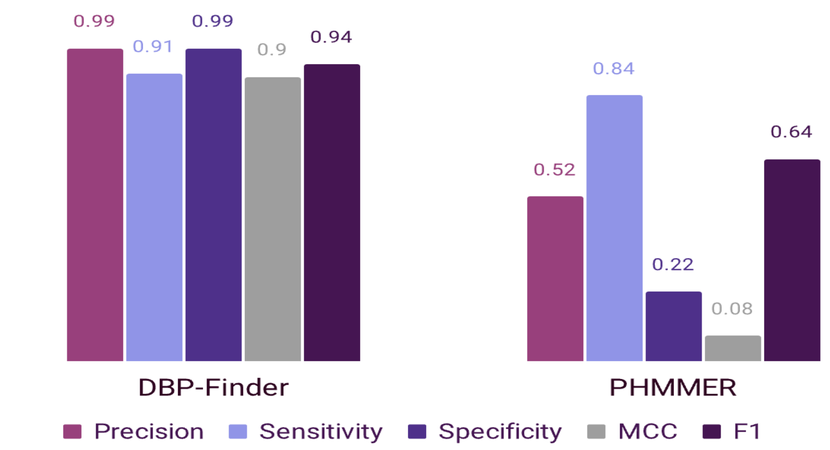
Clusters are predominantly enriched with either DBPs or Non-DBPs, indicating good data quality and supporting the idea that protein amino acid sequences determine DNA-binding function.

Results

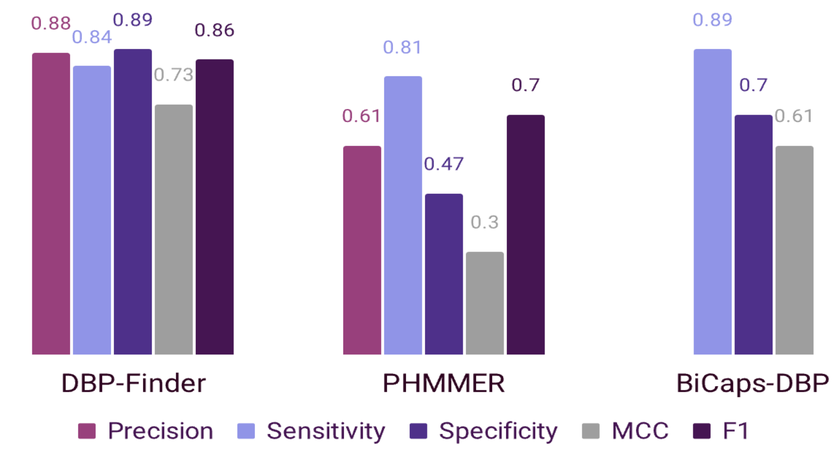
We trained an Ankh model (http://dx.doi.org/10.1101/2023.01.16.524265) with 13 million parameters, using a transformer backbone for protein sequence representation and a classification head. Training used the AdamW optimizer, batch size of 64, and an initial learning rate of 2e-4 over 9 epochs, retaining the model with the lowest validation loss. Accuracy and stability were enhanced by training an ensemble of five models.

Performance comparison

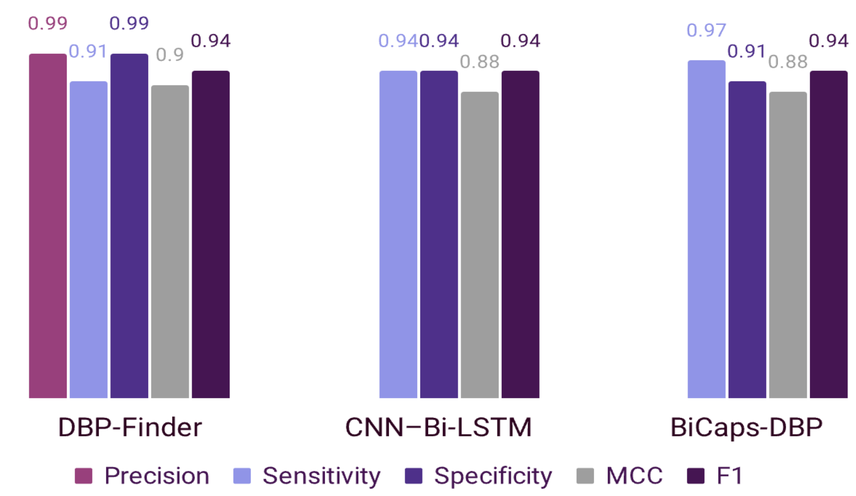
PDB1000 Testing Set:
PHMMER (https://doi.org/10.6019/tol.hmmer-w.2018.00001.1): HMM-based method for sequence similarity searches.



PDB2272 Testing Set:
BiCaps-DBP (https://doi.org/10.1016/j.compbiomed.2023.107241): a method with a three-layer architecture: encoding layer for one-hot encoding, Bi-LSTM layer for contextual features, and 1D-CapsNet layer for feature correlation and classification.



PDB20000 Testing Set:
CNN–Bi-LSTM (https://doi.org/10.1371/journal.pone.0225317): uses one-hot encodings, includes layers for amino acid numbers, continuous vectors, convolutions with max pooling, and Bi-LSTM for contextual features.



This work is an output of a research project implemented as part of the Research Program at the MSU Institute for Artificial Intelligence.