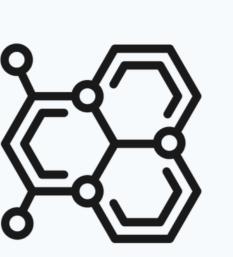




Dataset



2872

342

Training Peptides

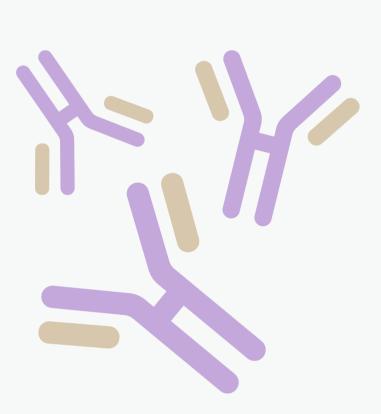
Test Peptides

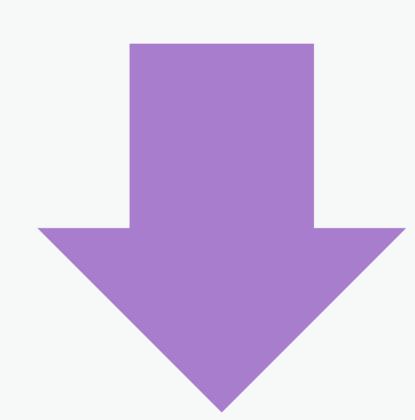
1245

1627

Training PIPs

Training non-PIPs





AGCTGAG TAGCTGAG CTAGCTGA



k-mer

Frequency counts for amino acid subsequences, capturing local motif patterns. \$(k=2, 3, 4)\$

168,400 → **35,358**

INITIAL VS. PRUNED DIMENSIONS



Physicochemical

Handcrafted descriptors including sequence length, hydrophobicity, charge, and structure propensities.

11 → **11**

INITIAL VS. PRUNED DIMENSIONS



ModIAMP

Global sequence descriptors calculated from the modIAMP Python package.

10 → **10**

INITIAL VS. PRUNED DIMENSIONS



ESM Embeddings

Contextual representations from a pretrained protein language model (CLS token).

320 → **320**

INITIAL VS. PRUNED DIMENSIONS

Final Dimensions

Retained dimensions for each model.

35,689

Model A

341

Model B

Coefficient Pruning

Filters features based on the magnitude of their logistic regression coefficients.

(Optimal threshold found to be \$\tau=0.4\$)

Zero-Variance Filtering

Removes features with no variation across the training set.

133,042

constant features removed

Initial Features

High-dimensional feature space from various descriptors.

> 168,731 Model A

> > 341

Model B



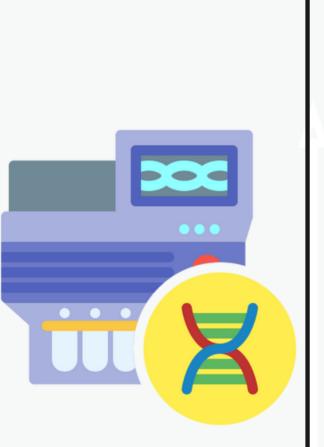
Model C: k-mer + modIAMP+ ESM

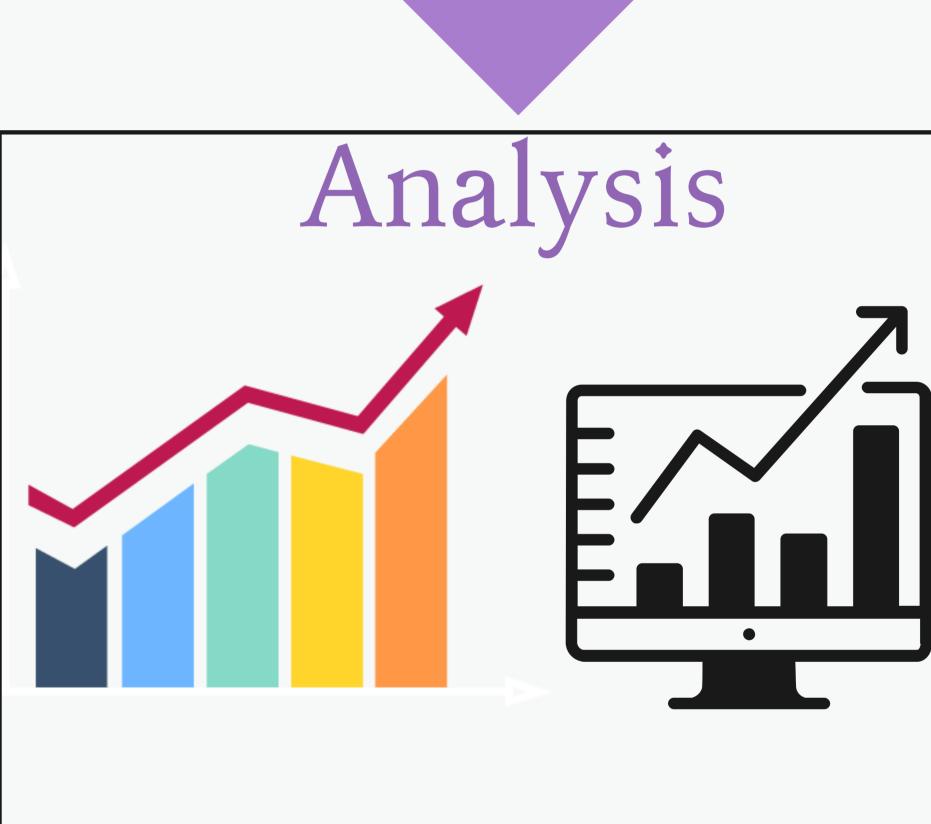
Model D: modIAMP + Physchem +

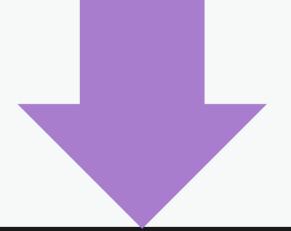
ESM

Ensemble: Average probability of

Model C & D











Supports: 1. CSV

2.FASTA

