# Ensemble Model for Promiscuity Prediction

We have designed an ensemble model to predict enzyme promiscuity class. Our ensemble framework comprises of 47 base learners which provide their outputs to a meta learner that predicts enzyme promiscuity. Although all of our base learners are trained using the same learning algorithm (Support Vector Classifier [1]), the heterogeneity among them is governed by the 47 different feature extraction techniques (listed in Table 1) used to encode the set of enzyme sequences. Each of the 47 different feature extraction techniques creates a unique feature vector representation of enzymes which serve as input to a designated base learner. The base learner trained on the set of encoded enzyme sequences yields the predicted promiscuity class of a given enzyme sequence as an output. The outputs of the k-best base learners are passed on to the meta learner that uses a majority voting scheme to predict the enzyme promiscuity category. The k-best base learners are selected on the basis of their accuracy score on training set. The workflow of our ensemble model is presented in Figure 1.

Table 1: Feature extraction type, name, software used to deploy and literature from which it was adopted are shown.

|  |  |  |  |
| --- | --- | --- | --- |
| Type | Name | Software Used | Literature Reference |
| Kernel | Spectrum Kernel | KeBABS | [2] |
| Kernel | Mismatch Kernel | KeBABS | [3] |
| Kernel | Gappy Pair Kernel | KeBABS | [4] |
| N-gram | Kmer | Numpy | [5] |
| N-gram | GAA-kmer | Numpy | NA |
| Physicochemical | AAC | iFeature | [6] |
| Physicochemical | CKSAAP | iFeature | [7] |
| Physicochemical | TPC | iFeature | [6] |
| Physicochemical | DPC | iFeature | [8] |
| Physicochemical | DDE | iFeature | [8] |
| Physicochemical | GAAC | iFeature | [9] |
| Physicochemical | CKSAAGP | iFeature | [10] |
| Physicochemical | GTPC | iFeature | [10] |
| Physicochemical | GDPC | iFeature | [10] |
| Physicochemical | Moran | iFeature | [11] |
| Physicochemical | Geary | iFeature | [12] |
| Physicochemical | NMBroto | iFeature | [13] |
| Physicochemical | CTDC | iFeature | [14]–[16] |
| Physicochemical | CTDT | iFeature | [14]–[16] |
| Physicochemical | CTDD | iFeature | [14]–[16] |
| Physicochemical | CTriad | iFeature | [17] |
| Physicochemical | KSCTriad | iFeature | [10] |
| Physicochemical | SOCNumber | iFeature | [10] |
| Physicochemical | QSOrder | iFeature | [10] |
| Physicochemical | PAAC | iFeature | [18] |
| Physicochemical | APAAC | iFeature | [18] |
| PSSM-based | AAC-PSSM | POSSUM | [19] |
| PSSM-based | DPC-PSSM | POSSUM | [19] |
| PSSM-based | AADP-PSSM | POSSUM | [19] |
| PSSM-based | PSSM-AC | POSSUM | [20] |
| PSSM-based | PSSM-CC | POSSUM | [21] |
| PSSM-based | RPSSM | POSSUM | [22] |
| PSSM-based | Tri-gram | POSSUM | [23] |
| PSSM-based | EDP | POSSUM | [24] |
| PSSM-based | EEDP | POSSUM | [24] |
| PSSM-based | MEDP | POSSUM | [24] |
| PSSM-based | TPC-PSSM | POSSUM | [25] |
| PSSM-based | AATP | POSSUM | [25] |
| PSSM-based | k-separated bigrams | POSSUM | [26] |
| PSSM-based | D-FPSSM | POSSUM | [27] |
| PSSM-based | S-FPSSM | POSSUM | [27] |
| PSSM-based | PSE-PSSM | POSSUM | [28] |
| PSSM-based | DP-PSSM | POSSUM | [29] |
| PSSM-based | PSSM-Composition | POSSUM | [30] |
| PSSM-based | Smoothed-PSSM | POSSUM | [31] |
| PSSM-based | AB-PSSM | POSSUM | [32] |
| PSSM-based | RPM-PSSM | POSSUM | [32] |

# Ensemble Model Results

We trained our ensemble model on 2933 enzyme sequences comprising of 1248 enzymes that represent the non-promiscuous class and 1685 enzymes that represent the promiscuous class. We also created a validation dataset which had 1541 enzymes, 1124 belonging to the promiscuous class and 417 belonging to the non-promiscuous class. Our ensemble model performed modestly on the held out validation dataset achieving an accuracy score of 0.64 and f1 score of 0.76. It achieved very low precision and recall scores of 0.34 and 0.33 on the class labeled non-promiscuous (the minority class in the dataset). The performance of our ensemble model along with the 17 best performing base models which were used in the ensemble is presented in Table 2.

Table 2: Performance of Ensemble Model and 17 best-performing base learners is shown. The models were judged on four performance metrics, precision on the non-promiscuous class, recall on the non-promiscuous class, F1 score and accuracy.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model Name | Precision | Recall | F1 score | Accuracy |
| Ensemble | 0.34 | 0.33 | 0.76 | 0.64 |
| CTriad | 0.27 | 0.39 | 0.66 | 0.55 |
| Moran | 0.3 | 0.42 | 0.69 | 0.58 |
| DDE | 0.29 | 0.4 | 0.68 | 0.57 |
| KSCTriad | 0.28 | 0.4 | 0.66 | 0.55 |
| DPC | 0.27 | 0.38 | 0.68 | 0.56 |
| Geary | 0.31 | 0.43 | 0.69 | 0.58 |
| TPC | 0.32 | 0.48 | 0.69 | 0.58 |
| GTPC | 0.3 | 0.4 | 0.69 | 0.58 |
| CKSAAP | 0.34 | 0.34 | 0.7 | 0.6 |
| NMBroto | 0.3 | 0.43 | 0.68 | 0.58 |
| GAA-kmer | 0.25 | 0.25 | 0.72 | 0.59 |
| S-FPSSM | 0.33 | 0.43 | 0.72 | 0.61 |
| CTDD | 0.28 | 0.38 | 0.68 | 0.57 |
| RPM-PSSM | 0.32 | 0.42 | 0.71 | 0.6 |
| SOCNumber | 0.27 | 0.37 | 0.67 | 0.56 |
| AB-PSSM | 0.32 | 0.43 | 0.71 | 0.6 |
| PSSM-Composition | 0.32 | 0.41 | 0.71 | 0.6 |

Diagram

Description automatically generated

Figure 1: Workflow of the ensemble model. Different feature representations create separate models, and the final model output is the majority vote of the predictions made by each individual model.

# CNN Model for Promiscuity Prediction

[1] C. Cortes and V. Vapnik, “Support-Vector Networks,” *Mach. Learn.*, 1995, doi: 10.1023/A:1022627411411.

[2] C. Leslie, E. Eskin, and W. S. Noble, “The spectrum kernel: a string kernel for SVM protein classification.,” *Pac. Symp. Biocomput.*, 2002, doi: 10.1142/9789812799623\_0053.

[3] C. S. Leslie, E. Eskin, A. Cohen, J. Weston, and W. S. Noble, “Mismatch string kernels for discriminative protein classification,” *Bioinformatics*, 2004, doi: 10.1093/bioinformatics/btg431.

[4] P. Kuksa, P. H. Huang, and V. Pavlovic, “A fast, large-scale learning method for protein sequence classification,” 2008.

[5] D. Jurafsky and J. H. Martin, “Language Modeling with N- grams,” *Speech Lang. Process.*, 2016.

[6] M. Bhasin and G. P. S. Raghava, “Classification of nuclear receptors based on amino acid composition and dipeptide composition,” *J. Biol. Chem.*, 2004, doi: 10.1074/jbc.M401932200.

[7] K. Chen, L. Kurgan, and M. Rahbari, “Prediction of protein crystallization using collocation of amino acid pairs,” *Biochem. Biophys. Res. Commun.*, 2007, doi: 10.1016/j.bbrc.2007.02.040.

[8] V. Saravanan and N. Gautham, “Harnessing computational biology for exact linear B-cell epitope prediction: A novel amino acid composition-based feature descriptor,” *Omi. A J. Integr. Biol.*, 2015, doi: 10.1089/omi.2015.0095.

[9] T. Y. Lee, Z. Q. Lin, S. J. Hsieh, N. A. Bretaña, and C. T. Lu, “Exploiting maximal dependence decomposition to identify conserved motifs from a group of aligned signal sequences,” *Bioinformatics*, 2011, doi: 10.1093/bioinformatics/btr291.

[10] Z. Chen *et al.*, “IFeature: A Python package and web server for features extraction and selection from protein and peptide sequences,” *Bioinformatics*, 2018, doi: 10.1093/bioinformatics/bty140.

[11] Z. P. Feng and C. T. Zhang, “Prediction of membrane protein types based on the hydrophobic index of amino acids,” *J. Protein Chem.*, 2000, doi: 10.1023/A:1007091128394.

[12] R. R. Sokal and B. A. Thomson, “Population structure inferred by local spatial autocorrelation: An example from an Amerindian tribal population,” *Am. J. Phys. Anthropol.*, 2006, doi: 10.1002/ajpa.20250.

[13] D. S. Horne, “Prediction of protein helix content from an autocorrelation analysis of sequence hydrophobicities,” *Biopolymers*, 1988, doi: 10.1002/bip.360270308.

[14] C. Z. Cai, L. Y. Han, Z. L. Ji, X. Chen, and Y. Z. Chen, “SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence,” *Nucleic Acids Res.*, 2003, doi: 10.1093/nar/gkg600.

[15] I. Dubchak, I. Muchnik, S. R. Holbrook, and S. H. Kim, “Prediction of protein folding class using global description of amino acid sequence,” *Proc. Natl. Acad. Sci. U. S. A.*, 1995, doi: 10.1073/pnas.92.19.8700.

[16] L. Y. Han, C. Z. Cai, S. L. Lo, M. C. M. Chung, and Y. Z. Chen, “Prediction of RNA-binding proteins from primary sequence by a support vector machine approach,” *RNA*, 2004, doi: 10.1261/rna.5890304.

[17] J. Shen *et al.*, “Predicting protein-protein interactions based only on sequences information,” *Proc. Natl. Acad. Sci. U. S. A.*, 2007, doi: 10.1073/pnas.0607879104.

[18] K. C. Chou, “Prediction of protein cellular attributes using pseudo-amino acid composition,” *Proteins Struct. Funct. Genet.*, 2001, doi: 10.1002/prot.1035.

[19] T. Liu, X. Zheng, and J. Wang, “Prediction of protein structural class for low-similarity sequences using support vector machine and PSI-BLAST profile,” *Biochimie*, 2010, doi: 10.1016/j.biochi.2010.06.013.

[20] T. Liu, X. Geng, X. Zheng, R. Li, and J. Wang, “Accurate prediction of protein structural class using auto covariance transformation of PSI-BLAST profiles,” *Amino Acids*, 2012, doi: 10.1007/s00726-011-0964-5.

[21] Y. Guo, L. Yu, Z. Wen, and M. Li, “Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences,” *Nucleic Acids Res.*, 2008, doi: 10.1093/nar/gkn159.

[22] S. Ding, Y. Li, Z. Shi, and S. Yan, “A protein structural classes prediction method based on predicted secondary structure and PSI-BLAST profile,” *Biochimie*, 2014, doi: 10.1016/j.biochi.2013.09.013.

[23] P. Tao, T. Liu, X. Li, and L. Chen, “Prediction of protein structural class using tri-gram probabilities of position-specific scoring matrix and recursive feature elimination,” *Amino Acids*, 2015, doi: 10.1007/s00726-014-1878-9.

[24] L. Zhang, X. Zhao, and L. Kong, “Predict protein structural class for low-similarity sequences by evolutionary difference information into the general form of Chou[U+05F3]s pseudo amino acid composition,” *J. Theor. Biol.*, 2014, doi: 10.1016/j.jtbi.2014.04.008.

[25] S. Zhang, F. Ye, and X. Yuan, “Using principal component analysis and support vector machine to predict protein structural class for low-similarity sequences via PSSM,” *J. Biomol. Struct. Dyn.*, 2012, doi: 10.1080/07391102.2011.672627.

[26] *et al.*, “Protein Fold Recognition Using Genetic Algorithm Optimized Voting Scheme and Profile Bigram,” *J. Softw.*, 2016, doi: 10.17706/jsw.11.8.756-767.

[27] J. Zahiri, O. Yaghoubi, M. Mohammad-Noori, R. Ebrahimpour, and A. Masoudi-Nejad, “PPIevo: Protein-protein interaction prediction from PSSM based evolutionary information,” *Genomics*, 2013, doi: 10.1016/j.ygeno.2013.05.006.

[28] K. C. Chou and H. Bin Shen, “MemType-2L: A Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM,” *Biochem. Biophys. Res. Commun.*, 2007, doi: 10.1016/j.bbrc.2007.06.027.

[29] E. Y. T. Juan, W. J. Li, J. H. Jhang, and C. H. Chiu, “Predicting protein subcellular localizations for gram-negative bacteria using DP-PSSM and support vector machines,” 2009, doi: 10.1109/CISIS.2009.194.

[30] L. Zou, C. Nan, F. Hu, and J. Hancock, “Accurate prediction of bacterial type IV secreted effectors using amino acid composition and PSSM profiles,” *Bioinformatics*, 2013, doi: 10.1093/bioinformatics/btt554.

[31] C. W. Cheng, E. C. Y. Su, J. K. Hwang, T. Y. Sung, and W. L. Hsu, “Predicting RNA-binding sites of proteins using support vector machines and evolutionary information,” 2008, doi: 10.1186/1471-2105-9-S12-S6.

[32] J. C. Jeong, X. Lin, and X. W. Chen, “On position-specific scoring matrix for protein function prediction,” *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, 2011, doi: 10.1109/TCBB.2010.93.