Novel Few-shots Learning Neural Network for Predicting Carbohydrate-active Enzyme (CAZyme) Affinity Towards Fructo-oligosaccharides

Shaoxun Liu
University of Southern California
Yi Kou (yikou@usc.edu)
University of Southern California https://orcid.org/0000-0002-7139-2043
Lin Chen
University of Southern California

Research article

Keywords: Machine learning, few-shots learning, neural network, CAZyme, sugar binding affinity, resistant sugar, fructo-oligosaccharides, Poisson process

Posted Date: October 19th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-92730/v1

Title Page

Novel Few-shots Learning Neural Network for Predicting Carbohydrate-active Enzyme (CAZyme) Affinity Towards Fructo-oligosaccharides

Shaoxun Liu[1, a], Yi Kou[1, a, *] and Lin Chen [a,*]

[1] Co-first Authorship

[a] Molecular and Computational Biology, Department of Biological Sciences, University of Southern California, 1050 Childs Way, Los Angeles, CA 90089, USA

*Correspondence should be addressed to Y.K. (yikou@usc.edu) or L.C. (linchen@usc.edu)

Abstract

Background: The enzymatic activity of the microbiome toward carbohydrates in the human digestive system is of enormous health significance (Zou, Y., et al., 2019; Pinard, D., et al., 2015). Predicting how carbohydrates through food intake may affect the distribution and balance of gut microbiota remains a major challenge. Understanding the enzyme-substrate specificity relationship of the carbohydrate-active enzyme (CAZyme) encoded by the vast gut microbiome will be an important step to address this question. In this study, we seek to establish an *in-silico* approach to studying the enzyme-substrate binding interaction.

Results: We focused on the key carbohydrate-active enzyme (CAZyme) and established a novel Poisson noise-based few-shots learning neural network (pFSLNN)

for predicting the binding affinity of indigestible carbohydrates. This approach achieved higher accuracy than other classic FSLNNs, and we have also formulated new algorithms for feature generation using only a few amino acid sequences. Sliding bin regression is integrated with mRMR for feature selection.

Conclusion: The resulting pFSLNN is an efficient model to predict the binding affinity between CAZyme and common oligosaccharides. This model can be potentially applied to binding affinity prediction of other protein-ligand interactions based on limited amino acid sequences.

Key words

Machine learning; few-shots learning; neural network; CAZyme; sugar binding affinity; resistant sugar; fructo-oligosaccharides; Poisson process.

Background

In recent years, increased attention has been paid to the human c microbiome and its health effect. Those microorganisms, mostly bacteria, inhabit the human gastrointestinal tract and engage in a symbiotic relationship with their host (Huttenhower, C. et al., 2012) (Conlon, M. A. and Anthony, B., 2015). The species of microorganism presented in the human body varies among individuals. Up to now, over 2000 species of the human microbiome have been discovered, and this number is predicted to increase as more human gut microbiome samples are collected (Almeida,

A. et al., 2019). Within those identified species, 997 species are proven to have statistical significance in positively influencing human health, which is named probiotic microbiome.

One conserved feature of the vast microbiome is the expression of carbohydrate-active enzymes (CAZyme). CAZymes are enzymes that perform the synthesis, recognition, and degradation (digestion) of carbohydrates. CAZymes are essential for the microbiome to break down the complex carbohydrates from various food sources, including plant cell walls and seaweeds (Huang, L. et al., 2018). Despite its prevalence, the human genome only expresses approximately 17 CAZymes (Bhattacharya, T. et al., 2015). As such, most of the food carbohydrates that are indigestible to human enzymes also referred to as dietary fibers, are digested by CAZymes encoded in the gut microbiome. Human gut microbiome CAZymes are highly diverse in sequences (Huang, L. et al., 2018). Remarkably, CAZymes encoded by the gut microbiome shows the adaptability to the carbohydrates accessible to the host (Jan-Hendrik Hehemann, et al., PNAS, 109, 19786, 2012). These observations suggest that understanding the enzyme-substrate specificity relationship of CAZyme of the gut microbiome may provide a way to use specific carbohydrates (prebiotics) to modulate population abundance and distribution of gut microbiota to promote probiotic effect. Carbohydrate-binding modules (CBM) are non-catalytic modules of CAZymes that facilitate substrate binding (Boraston, A.B. et al., 2004). This study will focus on

64 substrate binding by CBMs of CAZymes expressed by human microbiomes.

65

66 Oligosaccharides, consisting of 3-10 monosaccharides, are complex carbohydrates
67 found in a wide variety of biological systems. Oligosaccharides are abundant in
68 glycolipids and glycol proteins, where they play indispensable roles in cell recognition
69 and cell adhesion. One of the commonly seen oligosaccharides is fructo-
70 oligosaccharide (also called oligofructan), which is mainly found in fruits and
71 vegetables and has prebiotic activity as dietary fiber. Fructans can react with reactive
72 oxygen species, and this antioxidant activity can reduce potential inflammation
73 (Franco-Robles, E. and Mercedes G. L., 2015). While being a major component of the
74 human diet, Fructo-oligosaccharides cannot be digested by native human digestive
75 enzymes (Moise, A. and Maria R., 2017). They are solely digested by the human gut
76 microbiome by bacterial CAZymes (Franco-Robles, E. and Mercedes G. L., 2015).
77 However, the detailed enzyme-substrate recognition mechanisms between these
78 prebiotic oligosaccharides and microbiome encoded CAZymes remain largely
79 unexplored. To address this question, we seek to establish a high throughput and robust
80 computational approach that can be used to predict the carbohydrate substrate
81 preference by a given CAZyme of a specific bacterial species in the human gut
82 microbiome. As a first step toward this goal, here we studied the binding of four
83 model carbohydrate substrates to the active sites of CAZyme: 1-kestose, raffinose,
84 nystose, and stachyose. These are four fructo-oligosaccharides that are shown to be

digestible by the human microbiome (Hayakawa, K. et al., 1990), thus are used as CAZyme binding substrates during protein docking to analyze their binding pattern. These analyses will provide a structural basis for future exploration of the enzyme-substrate specificity relationship of CAZymes in the human gut microbiome at the level of big data.

To facilitate binding affinity prediction, we used protein models generated by I-Tasser and binding affinity predicted by Molegro Virtual Docker (MVD). Structure simulation and modeling by I-Tasser are carried out to predict the tertiary structure of a given amino acid sequence (AAS). I-Tasser is by far one of the most accurate protein structure prediction servers (MacCarthy, E. and Derrick, P., 2019) with more than 90% quality prediction accuracy and 85.1% accuracy in assigned molecular functions (Roy, A. et al., 2010). MVD is the software used for substrate-enzyme binding predictions between selected CBMs and the four aforementioned oligosaccharides. This software is also used in the docking analysis between chlorogenic acid and aldose reductase since it provides a consistent and relatively accurate score for binding models with different binding energy (Naeem, S. et al., 2013). MolDock algorithm used by this software provides protein cavity and substrate binding location predictions with around 87% accuracy and position deviation within 2Å (Thomsen, R. and Mikael H. Christensen., 2006). MVD provides possible binding locations, binding energy scoring (rerank score), as well as cavity related fragment sequences and substrate binding residuals. MVD is

by far the most optimum docking software considering accuracy, information output, and runtime.

The simulation of binding interactions between the CAZyme CBM and substrate oligosaccharide is relatively accurate but highly time-consuming. An average docking process for one CAZyme on the I-Tasser server is approximately 30 hours. The advent of machine learning provides an efficient approach for this time cost issue. However, the difficulty for these enzymes and substrates to be in full data set simulation using any machine learning models is its lack of ample sample space. Heavy data training is, therefore, often impossible due to the scarcity of available enzyme/substrate structures. To overcome this limitation, we sought to apply few-shots learning (FSL) ideas and develop our version of protein sequence-based Poisson augmentation few-shots learning network.

This study aims to establish a method of predicting enzyme-substrate (protein-ligand) binding affinity across an unlimited number of proteins in a given sample ensemble, based on a small sample (~50) of enzyme-substrate docking results. Few-shot learning is used to generate a neural network that is capable of differentiating the distinctive classes under various goals, for example, classifying pictures of different animal species will small data set (Richard, Z. et al., 2017) (Li, Z. et al., 2017). This property is especially important since the high variability of AAS in proteins allows testing samples

to consist of rather various sequences from training samples. A neural network optimization algorithm of finding the loss of each round of neural network generation is a core feature of few-shot learning (Garcia, V. and Joan B., 2017). The loss algorithm applied is based on a prototypical neural network with adjustment of using accuracy rate instead of Euclidean distance; since among the various neural networks available, a prototypical neural network is the most reliable means of approach in this situation thanks to its outstanding performance in the small sample space in practices (Pan, Y. et al., 2019), which often outputs the prediction accuracy that has surpassed human recognition (He, K. et al., 2016). By integrating with few-shot learning algorithms, the prototypical neural network achieved an approximated 70% accuracy in 5-way 5-shot image classification (Richard, Z. et al., 2017).

In this study, the whole set CAZyme CBMs of probiotic human microbiomes are obtained from CAZy-database (Lombard, V. et al., 2014). The over 4000 proteins are clustered based on K-nearest neighbors according to the primary structure. This study provides the novel idea of selecting anchor protein as bases for feature generation, including cavity site and protein binding site similarity calculated through fuzzy search according to anchor protein binding site fragment sequences.

Aim to establish an improved few-shots learning model, we bring in the data augmentation through Poisson noise, since it represents the distribution of amino acid

in 1D. Previous research shows that the site substitution mutation of proteins can be described by the Poisson-correction method (Sadygov, R. G., 2018). Especially when the substitution rate is independent among sites, Poisson -correction can best describe the scenario (Grishin, N. V., 1995). In this study, since the site-dependence of substitution is unknown, site-independent substitution will be assumed. In addition, we mapped the data into several higher dimensions. We have also compared the Poisson data augmentation with Gaussian, random, and salt-and-pepper noises.

The major significance of this study is several folds: firstly, this study takes the first step towards understanding enzymatic function at the scale of the gut microbiome, which is a timely topic attracting much attention. Secondly, the study establishes a generalized method pipeline for future similar few shots learning in biology and is the first to try FSL and noise augmentation on proteins. Since the enzyme-substrate binding predictions are based on primary structure instead of the tertiary structure used in most other studies, the time of protein simulation can be reduced. Thirdly, the study sets up the first example for future studies of protein-substrate interactions to be performed with minimal data input and limited computational power with reasonable accuracy.

Results

1 Prototypical neural network for Few-shot learning

This experiment aims to provide a method of prediction of protein-ligand interaction based on a small amount of labeled data, since acquiring the labeled training set using I-Tasser and Molgro is the most time-consuming.

Using a small training set, though timelier, provides less accurate prediction results using traditional machine learning algorithms. To achieve better performance, we adopted and modified the prototypical neural network for few-shot learning. This model applies to our data set in two aspects. First is that most features of the data set resemble distance from a specific anchor data, which renders each data point inherent distance to a calculated prototype. This feature generating technique inherently implies the protein evolution tree, where proteins with similar functions from similar organisms closely resemble ligand binding site structure with each other. Second is that since the training set is small, multiple epochs of neural network formation are best to run to exploit the random selection of the starting point of linear regression so that the neural network which has the best performance during cross-validation can be selected.

Applying prototypical neural network algorithms increases the F1 score of prediction for 18%, comparing the next best machine learning model SVM (table 1). Data augmentation techniques are also applied to the data set. Such a technique has been used in image recognition in previous studies of artificial intelligence industries.

Previous studies on using data augmentation in FSL incorporated gaussian noise and linear transformations, and the prediction accuracy on EMNIST data and Face recognition reached accuracies of 80.25% and 58.46% using 25 samples (Antoniou, A. et al., 2017). Poisson noise augmentation was applied to the data set, and in optimum configurations, it increased the F1 score of prediction by 8.67%. The justification and discussion were stated in the following section (table 1).

2 Effect of Poisson augmentation

The augmenting input sequence increases the sample size, which better supports the neural net formation when experimental data scare. The natural mutation of AAS sequences is independent between each site of amino acid. The mutation rate of each AA remains constant, disregarding the sequence. Assuming the probability of occurrence of a specific event in a small interval of a sequence is equal to the macroscopic intensity, such mutation rate can be described by a Poisson distribution, where the value of lambda represents the mutation rate. This mutation rate consists of both total mutation rate, the possibility that an amino acid site will mutate versus will not mutate, and the amino acid-specific mutation rate, the probability of which amino acid the site will mutate. The amino acid-specific mutation state was generated by summarizing the occurrence frequency of each amino acid in the entire sample set. This likely represents the relative abundance of each amino acid. Such a method was

compared with an evenly distributed model, in which the former model has better prediction results. A range of different total mutation rates was tested, and 10% gave the best result. The increase of the total mutation rate exhibits a possible trade-off between overfitting and information perseverance.

3 Feature importance

Minimum Redundancy Maximum Relevance (mRMR) algorithm was used to calculate the effectiveness and redundancy of the features. The mRMR score of each feature was calculated, and the features were rearranged accordingly. Without loss of generality, a sliding bin of 10-feature was used to slide over the rearranged features. Features in the bin were the only input features for the FSL algorithm. F1 scores of those trails were obtained and shown in figure 2.

Features ranking 1-40 shows a decreasing trend as expected. A second peak appeared at feature group 51-60, suggesting that there are features, though redundant, are capable of providing more substantial information. The redundancy of features can be explained by either caused by combinations between features or the underlying scaling mechanism of the mRMR algorithm. The features were further rearranged according to the resulting F1 values to validate the claim above. Not to lose generality, a sliding bin of size 20 was used to generate three groups, with 1-10 and 51-60 in group 1, 11-20 and

41-50 in group 2, and 21-30 and 61-70 in group 3. The resulting F1 values are shown in figure 2.

51-60 contain all 9 Sugar binding Alignment scores and no Cavity Fragment Alignment scores. This can be explained by the possible redundancy of Sugar Binding Alignment as they are basically, as mentioned in the method section, Cavity Fragment Alignment with a sugar-binding coefficient matrix applied to the AA exchange matrix. This suggests that the modification of the sugar-binding coefficient does not have an apparent effect on the prediction ability of cavity fragments.

The decreasing trend shown supports the hypothesis. This suggests that the most important features in this pNN neural network come from group 1 of the rearranged features. Note that the F1 score of intervals 1-10 and group 1 both exceeds the F1 score of all features suggesting that contradicting features downplays the predicting capability of the model.

Table 2 listed out the types of features that are selected from the 71-feature to the top 20 features generating the highest F1 value. High percentages of α-Helix prediction score and Cavity Fragment Alignment score was selected. The identified important features retain intrinsic structural significance. α-Helix prediction score consists of individual residue count (SSSH_01) and long strand count (SSSH_02). A higher score

253 of either feature indicates the more abundant in α-helix (figure 4a, 4b). Cavity
254 alignment and Sugar binding alignment scores show the possibility of occurrence of a
255 similar cavity pattern between the sample protein and a given anchor protein (figure 4f).
256 The whole sequence of sugar-binding score shows gives the average affinity of residues
257 in the sample protein to a specific ligand (figure 4g).
258
259 Protein samples with a high correlation in values of important features are likely to
260 preserve similar structural identities and functionality. For example, P19 is a sample
261 protein that has resulted in a perfect score aligning with anchor protein P5 (CFAL_05,
262 SBAK_05), and such a relationship is validated by AAS alignment using super
263 algorithm resulting in an RMSD of 0.389 (figure 4e). Whole sequence alignment shows
264 the level of similarity in AAS between a sample and an anchor protein. P19, though
265 possesses a drastically different sequence compared with P5, has a relatively high whole
266 sequence alignment score (WSAL_05). Thus, a high structural similarity between P5
267 and P19 is identified (figure 4c, 4d). Subsequent research indicates that P5 and P19 are
268 glycogen debranching proteins in B. glumae and A. veronii (Lim, JaeYun, et al., 2009;
269 Yang, Honghui, et al., 1996). Since B. glumae and A. veronii are of different classes, it
270 is highly possible that P5 and P19 are enzymatic proteins that underwent convergent
271 evolution. The discovery of the Glxg proteins provides evidence that this pFSLNN has
272 the potential to identify proteins of similar function regardless of phylogenic origin.
273 Using the same important features for assessment, we also identified P10 and P52 to

contain similar residues in their binding cavity (figure 5). Those two proteins also come from different bacteria species. Despite having drastically different AAS, the two proteins showed similar key residues at their binding cavities. Such ability of this model can serve to accomplish the goal of classifying the human microbiome basing on enzymatic functionalities instead of 16s RNA, and be beneficial to the research of the functionalities of probiotics.

Discussion

This study focuses on the novel FSL with Poisson augmentation on data sets. This idea can be used in other fields such as genomic prediction, where datasets are few. The FSL model was built upon techniques in the generation of feature matrices, which can be applied to prediction models on interactions with unknown causal features but has symbolic labeled subjects as anchors. Another important finding is that certain features, including cavity fragment similarity and α-helix pattern, are important for the prediction of binding affinity for resistant sugars. Moreover, the method of evaluating features by sliding bin regression can be applied to other FSL learning models.

It should be noted that there are certain limitations for this study: a) Although it facilitates the acquirement of virtual calculated binding affinity data when the dataset is small and especially when the mass structure simulation is not an option, the

generalization of this specific Poisson augmented FSL pNN directly to others are relatively not easy. b) It still remains as a question of whether certain features are causally representative of a special kind enzyme as predicted and selected by this model, and likewise, if an apparent unknown transferable feature set can be obtained among similar protein/enzyme species. c) Feature encoding combining vector embedding method and traditional ways in FSL needs to be further explored for the sake of prediction confidence and power. d) There are certain other FSL modes and patterns which could further enhance the prediction score if Poisson augmentation is added. Other methods using deep learning networks (Thapa, N. et al., 2020) may achieve better results for this kind of study, but in terms of time cost and coding vector embedding, it may not be very well suited to FSL framework, especially with pNN, but is definitely worthy of future investigation. e) Specifically, for CAZymes, more structural features involving side chain interactions with certain sugar structure types can be further explored. Depending on the interacting group characteristics, evaluation scores can be reranked towards certain preferences such as H-bond/aromatic stacking, and the results can be regionally optimized and cross-validated globally (McCartney, L. et al., 2004). Still, this study is only a small step towards understanding the CAZyme features among thousands of probiotic types. The intriguing world of probiotic bacteria and their CAZyme relationships, together with the charming world of FSL modeling, is definitely worth substantial future works to be devoted to.

Conclusion

The study focused on the binding of 4 typical resistant sugars with key carbohydrate-active enzymes (CAZyme) and established a novel Poisson noise-based few-shots learning neural network (pFSLNN) for predicting the binding affinity of indigestible carbohydrates. This approach achieved higher F1 scores than other classic FSLNNs using Poisson noise augmentation, which has never been applied in the FSL fields before. The Poisson augmentation is found to be optimal at a 10% noise level. During the pFSLNN establishment, we have also formulated several new algorithms for generating feature matrix depending on a few linear amino acid sequences, such as sliding window fuzzy search and two-dimension threshold optimization. We have also evaluated feature importance by novel sliding window method. Several discoveries concerning the binding pattern of the resistant sugars have been made during the pFSLNN prediction: 1) Different proteins share relatively similar binding cavities and patterns concerning the same sugar substrate, with the same interaction residues and 3D structures around the sugar. 2) The overall structures can be quite similar even across different 16S classes with vastly distinctive sequences, which suggests that some key residues and fragment parts far from the cavity are enough to reestablish the similar same binding mode and the whole protein structures. These results suggest a new binding function-based relationship between CAZymes and resistant sugars from the structure perspective endowed by pFSLNN prediction.

Methods

1 Data collection

1.1 Sample preparation

A list of probiotic human microbiomes that contains 997 species was adopted from previous research (Forster, Samuel C., et al., 2016)). Each species was searched on the CAZy database (Alisdair B. et al., 2004) for its expressed CAZymes, and all CBM's AAS were downloaded. This pool of CBM contains 3749 molecules. To acquire AAS with distinct features, the pool of AAS was first grouped using K-mean cluster analysis. Without loss of generality, h = 500 was selected as the cutoff line, and 10 groups were generated. 6 AAS were randomly chosen from each group to make up the sample set of 60 AAS. Random selection after clustering can ensure AAS with different general characteristics is evenly represented in the sample set.

The 60 AAS samples were uploaded to the I-Tasser server (https://zhanglab.ccmb.med.umich.edu/I-TASSER/) for protein structure modeling and simulation. Substrate oligosaccharide models of 1-kestose (440080), raffinose (439242), nystose (166775), and stachyose (439531) were obtained from PubChem databank (Berman, H.M. et al., 2000). Molgro Virtual Docker was used to detect carbohydrate-binding cavities and protein-ligand binding positions. The cavity was searched for each protein, and the binding position search was performed within a 15 Å radius around the center of the cavity after considering the sugar substrate sizes in this study. 10 binding simulations

were performed for each protein-ligand pair. The binding position with the lowest Rerank score was recorded. For each AAS-oligosaccharide pair, ones with Rerank score below -100 were labeled as 1, representing binding, and others were labeled as 0, representing non-binding. Each AAS thus has four labels.

1.2 Anchor AAS selection

Assuming each AAS in different groups is distinct, one AAS from each group (10 in total) were selected as anchor AAS. Those AAS were not used as testing samples in the following few-shot learning process. For those 10 AAS, residues that are within 6Å (Biro, J. C., 2006) of the cavity site were recorded as cavity related fragments with connected residues in the same fragment. Fragments of less than three residues were neglected. Sugar-binding fragments were also recorded based on the binding position of each oligosaccharide. Those fragments were searched for in each AAS.

The key concept of the feature generation pipeline is to obtain the binding pattern of 10 anchor AAS. The higher similarity in secondary structure, cavity fragments, and sugar-binding fragments between a tested AAS and an anchor AAS suggest a higher possibility for the two proteins to have the same protein-ligand binding pattern. Anchor AAS always has the maximum available score when compared to its secondary structure, cavity fragment, and sugar-binding fragments; thus, they were taken as feature standards by the prototypical neural network and remained in training set for

each round of learning.

1.3 Feature value matrices preparation

According to the secondary structure sequence returned from I-Tasser, the frequency of each AA appearing as each general secondary structure type (Helix, Sheet, Coil) was recorded. These data were used to predict secondary structure.

The AA exchange matrix was adopted from Lev Y. Yampolsky and Arlin Stoltzfus's research on The Exchangeability of Amino Acids in Proteins (Yampolsky, L.Y. and Arlin S., 2005), this matrix was used to assign similarity scores when performing fuzzy search between the cavity fragments and AAS. The sugar-binding coefficient was then applied to the AA exchange matrix to generate a sugar-binding exchange matrix.

2 Neural network Preparation

2.1 Feature generation pipeline

For each AAS sample, a total of 71 features (6 from secondary structure score, 10 from binding cavity alignment, 10 from whole sequence alignment, 40 sugar-binding fragment alignment, 4 from sugar-binding whole sequence alignment, 1 from sample AAS length) were generated according to the AAS and the matrixes mentioned above: 6 features were generated for secondary structure score, including the estimated number of promoting AA and estimated number of long consecutive representing each of the

three general secondary structure types. Those parameters of secondary structure give hints to the overall shape of the protein, as more helix promoting AA with less helix strand suggests helix strands being longer towards a rod shape. 10 features were generated from cavity fragment alignment. Fuzzy search algorithm (Algorithm 2) was applied to cavity fragments generated from anchor proteins on each sample AAS to search for the longest succeeding fragment chain. A higher score against either anchor protein implies a higher possibility for a similar cavity to form. 10 features of the whole sequence alignment score implied the possibility of the whole sequence to present similar interactions between the anchor protein and sample AAS. 40 features were generated from the sugar-binding AA exchange matrix. The same algorithm was applied, but the AA exchange matrix has been modified according to the frequency of each AA binding with a respective oligosaccharide. 4 features of whole sequence sugar-binding scores were generated using a fuzzy search algorithm with an interaction coefficient modification to the AA exchange matrix. 1 feature of AAS length was added.

Using fragment and whole sequence similarity as a feature instead of direct and simple AAS has three advantages. Firstly, the median length of sample AAS is approximately 530, introducing 530 features in building a neural network is unrealistically time-consuming. Secondly, the properties of AA cannot be linearly represented due to 3D intramolecular structures. This means that the feature matrix for each AA would be indefinite and hard to be quantified in only 1 dimension. Thirdly, since the sample AAS

has a different length, a convolutional neural network that was to be applied would be increasing its time cost. Since the aim of this study is to complete mass prediction in the shortest time with only a limited sample size, applying a fixed number of features that describe binding patterns would be optimum.

2.2 prototypical Neural Network (pNN) formation

For the 60 sample AAS, each AAS was 1:10 augmented by Poisson noise (detailed description is in the data augmentation section below). A matrix of 660 AAS samples, each with 71 features and 4 labels, was generated after the feature generation pipeline. AAS samples augmented from the same AAS sample, including the original AAS, were defined to possess the same root. The set of AAS samples was denoted D. The set of 110 anchor AAS were denoted D^{anchor}, where $D^{anchor} \in D$. 110 samples of 10 different roots from D - D^{anchor} was randomly selected as the training set, denoted as D^{train}, the remaining is the testing set, denoted D^{test}. The ratio between the training and testing set is 5:1.

For each epoch of FSL training, D^{train} was divided into supporting set S and query set Q, where the number of samples in S and the number of samples in Q has a ratio of 4:1. 440 AAS samples of 40 different roots, including D^{Anchor} were used to compute the prototype from S. The training algorithm of the Few-Shot learning model is the same as Jake Snell, Kevin Swersky, and Richard Zemel's (Richard, Z. et al., 2017). D^{Anchor}

was always included in the supporting set as they provided the guidelines of each feature. Anchor proteins contributed the most to the class generation as they have the most distinct feature values. The neural network generated that performs the classification is denoted NN. Q, consisting of the leftover 110 AAS, is used to compute Loss-J of NN. Loss-J was modified to be the number of incorrect predictions of this neural network on the validation set. As epochs proceed to 100, the NN with the least loss-J was selected to be returned as the best neural network (bNN). The prediction accuracy of bNN for D^{test} was recorded.

3 Feature generation

3.1 Generation of secondary structure score

-Feature structure

Each amino acid sequence (AAS) input returned 6 values that consist of a secondary structure score, denoted as Mij ($i \in [1,3]$, $i \in [1,2]$). Those values included estimation in the number of Amino acids (AA) promoting each classic secondary structure (Helixes, plated Sheets, and random Coils) that were denoted as H, S, and C, as well as the estimation of the number of long, consecutive strands (≥ 5) of AA promoting the same secondary structure. To be convenient as a demo, we defined a fragment of AAS, showing the consecutive occurrence of the same classic secondary structure over 5 times as a secondary structure strand.

The function that generates secondary structure scores, denoted GenSS(), was based on the secondary structure promotion matrix (sspM) and secondary structure promotion threshold matrix (sspT). sspM recorded the frequency of each AA appearing in each classic secondary structure after normalization according to the average and standard deviation of the training set. The normalization of this matrix allowed the values to be in a statistical range. sspT contained three sets of thresholds, which are the promotion bar, demotion bar, and the tolerance number. The two bars characterized AA into Promoting, indifferent, and demoting for each of the three secondary structure types. The promotion bar is the lower bound of the sspM value of a given AA being characterized to start or succeed a secondary structure strand. The demotion bar is the higher bound of the sspM value of a given AA being characterized to prohibit or terminate a secondary structure strand. AA with sspM value between the two thresholds is considered indifferent to a classic secondary structure type. The tolerance number is the minimum number of AA in a secondary structure strand to view the next indifferent AA as a successor of the ongoing secondary structure strand.

-Generation process

The data input that generates those matrixes consisted the AAS training set, denoted A^{trian}_{ij} ($i \in [1,50]$, $j \in [1, length(A^{trian}_{i})]$), where i and j mean the j^{th} AA from the i^{th} AAS of the set, and the secondary structure sequence (SSS) returned by I-Tasser, denoted S^{trian}_{ij} ($i \in [1,50]$, $j \in [1,length(S^{trian}_{i})]$).

The data collection for $m_{xy} \in$ sspM ($x \in [1,20]$, $y \in [1,3]$), where m_{xy} represents the x^{th} AA, or the AA "x", and the y^{th} classic secondary structure, is shown below. m_{xy} finds the portion between an AA in the helix (sheet, coil) and the total amount of the AA type, normalized by the average and standard deviation of all three classic secondary structure types.

$$m_{xy} = \frac{-\overline{m_y} + \frac{\sum_{i,j \in (A_{ij}^{Train}=x)} f\left(A_{ij}^{Train}\right)}{\sum_{i,j \in (A_{ij}^{Train}=x)} 1}}{\sigma_{m_y}}$$

$$f\left(A_{ij}^{Train}\right) = \begin{cases} 1, A_{ij}^{Train} = S_{ij}^{Train} \\ 0, A_{ij}^{Train} \neq S_{ij}^{Train} \end{cases}$$

The threshold matrix $t_{xy} \in$ sspT ($x \in [1,3]$, $y \in [1,3]$), where t_{xy} represents the x = 1 (higher), x = 2 (lower), x = 3 (tolerance) threshold for the y^{th} classic secondary structure, was optimized through linear regression of minimizing difference between GenSS() output value of A^{trian}_{ij} and S^{trian}_{ij}. A demo run of GenSS() using the values from a given sspT was denoted by $GenSS^t()$. The sspT was optimized when the difference between the estimated value and real value (running SSS in the same algorithms gives the real value) is minimized.

$$\frac{d \sum \frac{\left|GenSS^t\left(A^{train}\right) - GenSS^t\left(S^{train}\right)\right|}{GenSS^t(S^{train})}}{dt} = 0 \quad (3)$$

GenSS() has two parts: the first is counting the number of each secondary structure promoting AA. For the c^{th} AAS, there are:

$$M_{i1} = \sum_{j \in A_{cj}^{Train}} f\left(A_{cj}^{Train}\right)$$

$$f\left(A_{cj}^{Train}\right) = \begin{cases} 1, m_{A_{cj}^{Train}\ i} \geq \ \mathrm{t}_{1i} \\ 0, m_{A_{cj}^{Train}\ i} < \ \mathrm{t}_{1i} \end{cases}$$

The generation of M_{i2} is further illustrated in Figure 8.

3.2 Generation of fragment binding and whole sequence binding score

-AA exchange fuzzy search

Fuzzy search is a searching algorithm based on the sliding-window idea with a penalty of the difference applied to each distinctive element in the window (Vernica, R. and Chen L., 2009). The advantage of the sliding-window algorithm against the Smith-Waterman algorithm is that the Smith-Waterman algorithm aims to find local alignments between the two strands, which neither must include the other, while the sliding-window algorithm ensures to find consecutive and including alignments. In addition, the Smith-Waterman algorithm aims to find the aligning strand while the aim of the fuzzy search is to return the alignment score for each site.

The substitution matrix in this alignment was the AA exchange matrix, denoted $aaEX_{ij}$

(i∈[1,20], j∈[1, 20]), where i is the substituting AA, and j is the substituted AA. The score for the same AA substitution is 1000, and the higher score indicated better substitution efficiency. One AAS of sequence and fragment were inputs for one round of fuzzy search. The fragment was being searched throughout the sequence. The sequence is denoted S, where S_m is the m^{th} AA of the sequence. The fragment is denoted F, where F_n is the n^{th} AA of the fragment. The Fuzzy search returns a vector of length m, denoted R_m. Each value of R represented the alignment score between S and F at the given position. The value was the average of aaEX values substituting each AA from the sequence for the AA from the fragment.

$$R_i = \frac{\sum_{j=1}^{length(F)} aaEX_{S_{i+j-1}F_j}}{length(F)}$$

-Longest increasing fragment

A vector of cavity fragments was obtained from each anchor protein. Each fragment was labeled by its order in the anchor protein sequence. A new vector V of length m was first filled with placeholder values. For each R_{xm}, which x indicates the x^{th} fragment from the anchor, looping through R_x, each i (i∈[1,m]) that has R_{xi} larger than threshold value T, V_i is labeled x.

$$V_i = f(R_{xi}) \qquad f(R_{xi}) = \begin{cases} 0, R_{xi} < T \\ x, R_{xi} \geq T \end{cases}$$

A second vector W of the same length as V stored the fuzzy search alignment value of

each corresponding position. A longest increasing subsequence searching algorithm (Aldous, D. and Persi D., 1999) was applied to V with the weight of each position modified to its corresponding value in W. The returning cavity fragment alignment score was the total weight of the longest increasing subsequence in V that was divided by the total number of AA in all the fragments.

-Two-dimensional optimization of threshold T

Threshold T was a crucial parameter in the search for the longest cavity strand. This parameter was decided to increase the standard deviation of each column vector of feature matrix while maintaining the minimum difference of such standard deviation across the 10 features. We denote each column vector of the 10 cavity fragment alignment of all samples as F_i ($i \in [1,10]$), where i indicates the cavity fragment alignment vector with the i^{th} anchor protein. The standard deviation of F_i using t as threshold T is denoted sdF_i^t. The standard deviation of sdF_i^t for $i \in [1,10]$, is denoted $sd(sdF^t)$. t that fulfills equation 8 was chosen as T. T value is 608 in the experimental run.

$$\frac{d\frac{\sum_{i=1}^{10} sdF_i^t}{sd(sdF^t)}}{dt} = 0$$

-Whole sequence fuzzy search

For the whole sequence alignment score, the shorter sequence between the anchor

sequence and the sample sequence was viewed as a fragment. The same algorithm in the previous section is applied with the substituting and substituted AA assigned according to the compared length between the two AA. AA was substituted from the sample AAS to the anchor AAS. The fuzzy search alignment value was returned as the whole sequence alignment score.

-Sugar binding matrix

4 sugar-specific binding matrix was multiplied with a coefficient to aaEX to form sugar-binding AA exchange matrix denote $sbEX^{i}_{mn}$ ($i \in [1,4]$, $m,n \in [1,20]$), where i represents 1-kestose, nystose, raffinose, and stachyose, n, and m represent substituting AA m with AA n. 4 vectors of AA-sugar interaction frequency, denoted sbV_j ($j \in [1,20]$), where j represents AA. For $i \in [1,4]$, the AA-sugar affinity matrix was generated by counting AA residuals that appear within 5Å (Sharma, R. et al., 2008) of the sugar-binding site. This method is similar to the method introduced in the work by Misaki Banno when performing AA-sugar affinity prediction (Banno, M. et al., 2017). sbM^{i}_{mn} was a matrix that contains the ratio between sbV_n and sbV_m. A larger ratio represents a higher affinity of substituted AA. $sbEX^{i}$ was generated by applying a sbM^{i} filter, multiplied by a factor F, on aaEX.

$$sbEX_{mn}^{i} = aaEX \times \left(1 + F \times \left(sbEX_{mn}^{i} - 1\right)\right) \quad (9)$$

F was also optimized using equation 8, replacing T by F.

3.3 Whole sequence sugar-binding

4 whole sequence sugar-binding score, one for each oligosaccharide ligand, was generated. AA-sugar interaction matrix was obtained using the same method as above. The difference is that AA in the whole sequence was accounted for instead of AA residuals that appear within 5Å of the sugar-binding site. And the average of AA-sugar interaction score for all AA in the protein was calculated as the whole sequence sugar-binding score.

3.4 Poisson augmentation

Poisson augmentation simulates the mutation of AAS to increase the sample size. The usage of Poisson distribution relied on the assumptions 1) AA mutation chance is independent of AA site; 2) the effect of minor mutations will not affect sugar-binding efficiency. This augmentation process required an AA frequency matrix, denoted $aaFM_i$ ($i \in [1,20]$), where i represents AA, and a mutation chance at each site C. The

probability mass function of Poisson distribution is given by function (10). The value of k = 1 and the value of λ was calculated by equation (11).

$$p = \frac{\lambda^k e^{-\lambda}}{k!}$$

$$\lambda_i = C \times \frac{aaFM_i}{\sum aaFM}$$

20 λ values formed an accumulated interval sequence. After an AAS was input to the Poisson augmentation function, a random number from 0 to 1 was generated for each site of AA. The interval in which the random number falls into determines if the AA at this given site would mutate and which AA it would mutate to.

The Poisson augmentation function was run on each sample AAS for 10 times to generate a set of augmented AAS of the same root. Test trials of C = 5%, 10%, and 20% were performed, and C = 10% gave the best prediction results. Augmented AAS ran the same feature generation pipeline.

List of abbreviations

CAZyme	Carbohydrate-active Enzyme
FSL	Few-Shots Learning
FSLNNs	Few-Shots Learning Neural Network

pNN	prototypical Neural Network
pFSLNN	Poisson noise-based Few-Shots Learning Neural Network
mRMR	minimum Redundancy Maximum Relevance
CBM	Carbohydrate-Binding Modules
MVD	Molegro Virtual Docker
AA	Amino Acid
AAS	Amino Acid Sequence
SVM	Support Vector Machines
sspM	secondary structure promotion Matrix
sspT	secondary structure promotion Threshold matrix
SSS	Secondary Structure Sequence
aaEX	amino acid EXchange matrix

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

Data available on GitHub: https://github.com/ShaoxunLiu/CAZyme_FSL

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

S. L and Y. K designed the model and implemented the study. S. L, Y. K and L. C drafted the manuscript. Y. K and L. C supervised the whole project. All authors have read and approved the manuscript.

Funding

No funding was obtained for this study

Acknowledgements

Great thanks to Professor Richard Zemel, Nicolas Terrapon, Alisdair Boraston, Dukka KC, and Bruce Tidor for their kind and informative discussion.

Reference

Aldous, D. and Persi D. (1999) Longest increasing subsequences: from patience sorting to the Baik-Deift-Johansson theorem. Bulletin of the American Mathematical Society 36.4: 413-432.

667 Alisdair B. et al. (2004) Carbohydrate-binding modules: fine-tuning polysaccharide
668 recognition. Biochemical journal 382.3: 769-781.
669 Almeida, A. et al. (2019) A new genomic blueprint of the human gut microbiota. Nature
670 568.7753: 499-504.
671 Antoniou, A. et al. (2017) Data augmentation generative adversarial networks. arXiv
672 preprint arXiv:1711.04340.
673 Berman, H.M. et al. (2000) The Protein Data Bank Nucleic Acids Research, 28: 235-
674 242.
675 Bhattacharya, T. et al. (2015) Global profiling of carbohydrate active enzymes in human
676 gut microbiome. PloS one 10.11: e0142038.
677 Biro, J. C. (2006) Amino acid size, charge, hydropathy indices and matrices for protein
678 structure analysis. Theoretical Biology and Medical Modelling 3.1: 15.
679 Bonk, B. M. et al. (2019) Machine Learning Identifies Chemical Characteristics That
680 Promote Enzyme Catalysis. Journal of the American Chemical Society 141.9: 4108-
681 4118.
682 Banno, M. et al. (2017) Development of a sugar-binding residue prediction system from
683 protein sequences using support vector machine. Computational biology and chemistry
684 66: 36-43.
685 Boraston, A.B. et al. (2004) Carbohydrate-binding modules: fine-tuning polysaccharide
686 recognition. Biochemical journal 382.3: 769-781.
687 Conlon, M. A. and Anthony, B. (2015) The impact of diet and lifestyle on gut

688 microbiota and human health. Nutrients 7.1: 17-44.

689 Forster, Samuel C., et al. (2016) "HPMCD: the database of human microbial
690 communities from metagenomic datasets and microbial reference genomes." Nucleic
691 acids research 44.D1: D604-D609.

692 Franco-Robles, E. and Mercedes G. L. (2015) Implication of fructans in health:
693 immunomodulatory and antioxidant mechanisms. The Scientific World Journal 2015.

694 Garcia, V. and Joan B. (2017) Few-shot learning with graph neural networks. arXiv
695 preprint arXiv:1711.04043.

696 Grishin, N. V. (1995) Estimation of the number of amino acid substitutions per site
697 when the substitution rate varies among sites. Journal of molecular evolution 41.5: 675-
698 679.

699 Han, X. et al. (2018) Fewrel: A large-scale supervised few-shot relation classification
700 dataset with state-of-the-art evaluation. arXiv preprint arXiv:1810.10147.

701 Hayakawa, K. et al. (1990) Effects of soybean oligosaccharides on human faecal flora.
702 Microbial Ecology in Health and Disease 3.6: 293-303.

703 He, K. et al. (2016) Deep residual learning for image recognition. Proceedings of the
704 IEEE conference on computer vision and pattern recognition. 770-778

705 Huang, L. et al. (2018) dbCAN-seq: a database of carbohydrate-active enzyme
706 (CAZyme) sequence and annotation. Nucleic acids research 46.D1: D516-D521.

707 Huttenhower, C. et al. (2012) Structure, function and diversity of the healthy human
708 microbiome. Nature 486.7402: 207.

709 Li, Z. et al. (2017) Meta-sgd: Learning to learn quickly for few-shot learning. arXiv
710 preprint arXiv:1707.09835.
711 Lim, JaeYun, et al. (2009) Complete genome sequence of Burkholderia glumae BGR1.
712 Journal of bacteriology 191.11: 3758-3759.
713 Lombard, V. et al. (2014) The carbohydrate-active enzymes database (CAZy) in 2013.
714 Nucleic acids research 42.D1: D490-D495.
715 MacCarthy, E. and Derrick, P. (2019) Advances in protein super-secondary structure
716 prediction and application to protein structure prediction. Protein Supersecondary
717 Structures. Humana Press, New York, NY. 15-45.
718 McCartney, L. et al. (2004) Glycoside hydrolase carbohydrate-binding modules as
719 molecular probes for the analysis of plant cell wall polymers. Analytical biochemistry
720 326.1: 49-54.
721 Moise, A. and Maria R. (2017) The Gut Microbiome: Exploring the Connection
722 Between Microbes, Diet, and Health. ABC-CLIO.
723 Montserrat, D. et al. (2017) Training object detection and recognition CNN models
724 using data augmentation. Electronic Imaging 2017.10: 27-36.
725 Naeem, S. et al. (2013) Docking studies of chlorogenic acid against aldose redutcase
726 by using molgro virtual docker software. Journal of Applied Pharmaceutical Science
727 3.1:13.
728 Ndeh, D. et al. (2017) Complex pectin metabolism by gut bacteria reveals novel
729 catalytic functions. Nature 544.7648: 65-70.

730 Pan, Y. et al. (2019) Transferrable prototypical networks for unsupervised domain
731 adaptation. Proceedings of the IEEE Conference on Computer Vision and Pattern
732 Recognition. 2239-2247
733 Pinard, D., et al. (2015) "Comparative analysis of plant carbohydrate active enZymes
734 and their role in xylogenesis." BMC genomics 16.1: 1-13.
735 Richard, Z. et al. (2017) Prototypical networks for few-shot learning. Advances in
736 neural information processing systems. 4077-4087.
737 Roy, A. et al. (2010) I-TASSER: a unified platform for automated protein structure and
738 function prediction. Nature protocols 5.4: 725-738.
739 Sadygov, R. G. (2018) Poisson model to generate isotope distribution for biomolecules.
740 Journal of proteome research 17.1: 751-758.
741 Sharma, R. et al. (2008) The interaction of carbohydrates and amino acids with aromatic
742 systems studied by density functional and semi-empirical molecular orbital calculations
743 with dispersion corrections. Physical Chemistry Chemical Physics 10.19: 2767-2774.
744 Thapa, N. et al. (2020) DeepSuccinylSite: a deep learning based approach for protein
745 succinylation site prediction. BMC bioinformatics 21: 1-10.
746 Thomsen, R. and Mikael H. Christensen. (2006) MolDock: a new technique for high-
747 accuracy molecular docking. Journal of medicinal chemistry 49.11: 3315-3321.
748 Yang, Honghui, et al. (1996) Coordinate genetic regulation of glycogen catabolism and
749 biosynthesis in Escherichia coli via the CsrA gene product. Journal of bacteriology
750 178.4: 1012-1017.

751 Vernica, R. and Chen L. (2009) Efficient top-k algorithms for fuzzy search in string
752 collections. Proceedings of the First International Workshop on Keyword Search on
753 Structured Data. 9-14.
754 Yampolsky, L.Y. and Arlin S. (2005) The exchangeability of amino acids in proteins.
755 Genetics 170.4: 1459-1472.
756 Zou, Y., et al. (2019) "1,520 reference genomes from cultivated human gut bacteria
757 enable functional microbiome analyses." Nature biotechnology 37.2: 179-185.
758
759 **Figure legends**
760 Figure 1: Fructo-Oligosaccharides. Name, CAS number, 3D structure, and 2D structure
761 of the four fructo-oligosaccharides that are used as substrates in this study.
762 Figure 2: F1 scores of FSL models trained using each 10-feature groups arranged
763 according to mRMR scores.
764 Figure3: F1 score of FSL models inputting 20 features from three groups of mRMR
765 intervals.
766 Figure 4: Structural presentation of sample proteins. a) Structure of P9 with alpha
767 helixes in red. SSSH_01 = 12, SSSH_02 = 150. b) Structure of P6 with alpha helixes
768 in red, SSSH_01 = 1, SSSH_02 = 35. c) Structure of anchor protein P5 with alpha
769 helixes in blue and beta sheets in yellow; binding ligand1-kestose is shown in red. d)
770 Structure of protein P19 with alpha helixes in blue and beta sheets in yellow; binding
771 ligand1-kestose is shown in red. e) Structural alignment of cavity fragments of P5 (blue)

772 and P19 (gray) using align function in pymol with alignment RMSD = 0.389; residue
773 labels are shown in corresponding colors. f) Sample AAS alignment of P5 (GlgX [B.
774 glumae]) and P19 (GlgX [A. veronii]) around the two cavity fragments; aligned AA of
775 a given secondary structure or belongs to a cavity fragment is shown in the color scheme.
776 g) Sample AAS alignment of non-cavity parts of P5 and P19, with AA of a given
777 secondary structure shown in the color scheme.
778 Figure 5: Predicted interaction of two CAZyme CBMs (P10 and P52) with 1-kestose.
779 P10 and P52 are highly correlated according to group 1 features, while having no
780 significant similarity in sequence. Similar amino acid residues are found in the radius
781 of the interaction of both P10 and P52, including SER, ALA, ARG, and LEU. This
782 suggests that the feature matrix may obtain inherent biological meaning. a) Interaction
783 model of P10 viewing from beta-D-fructofuranose. b) Interaction model of P10 viewing
784 from beta-D-fructofuranosyl residue. c) Interaction model of P10 viewing from alpha-
785 D-glucopyranosyl residue. d) Interaction model of P52 viewing from beta-D-
786 fructofuranose. e) Interaction model of P52 viewing frombeta-D-fructofuranosyl
787 residue. f) Interaction model of P52 viewing from talpha-D-glucopyranosyl residue.
788 Figure 6: Flow chart for feature generation procedure. Arrows show derivation.
789 Figure 7: Flow chart of feature matrix generation and pFSLNN process.
790 Figure 8: This figure illustrates the bases of the algorithm estimating the number of
791 residuals of each general secondary structure type. Without loss of generality, α-Helix
792 (H) is shown as an example. A cursor scans over the input AAS to find the first H

promoting AA. A secondary structure strand of H begins. When the strand length is more than or equal to 5, the strand is considered a long strand. As the cursor proceeds, 3 types of successor (promoting, demoting, indifferent) AA can result in 5 cases. Initial sspTH = (-0.5, -1.6, 2).

Tables

Table 1: Accuracy and F1 score for experimented models

Method	Augmented	Accuracy	F1 score
Random Forest	N	67.70%	44.36%
SVM	N	72.50%	52.13%
FSL unaugmented	N	66.55%	70.13%
FSL Poisson aug (5%)	Y	67.00%	76.92%
FSL Poisson aug (10%)	Y	69.15%	78.67%
FSL Poisson aug (20%)	Y	66.55%	77.24%
FSL Poisson aug (30%)	Y	65.75%	75.38%
FSL Poisson aug (40%)	Y	66.15%	68.47%
FSL Salt & Pepper aug (10%)	Y	65.00%	67.73%
FSL Random mutation (10%)	Y	67.45%	73.22%
FSL Gaussian aug (10%)	Y	65.50%	65.87%

Table 2: Percentage of selected feature types in the top 20 features

Feature Type	Total	Top 20	Percent Selected
Cavity Fragment Alignment	10	4	40%
Whole sequence Sugar Binding	4	3	75%
Whole sequence Alignment	10	2	20%
α-Helix prediction	2	2	100%
Sugar binding Alignment	40	9	22.5%

Figures

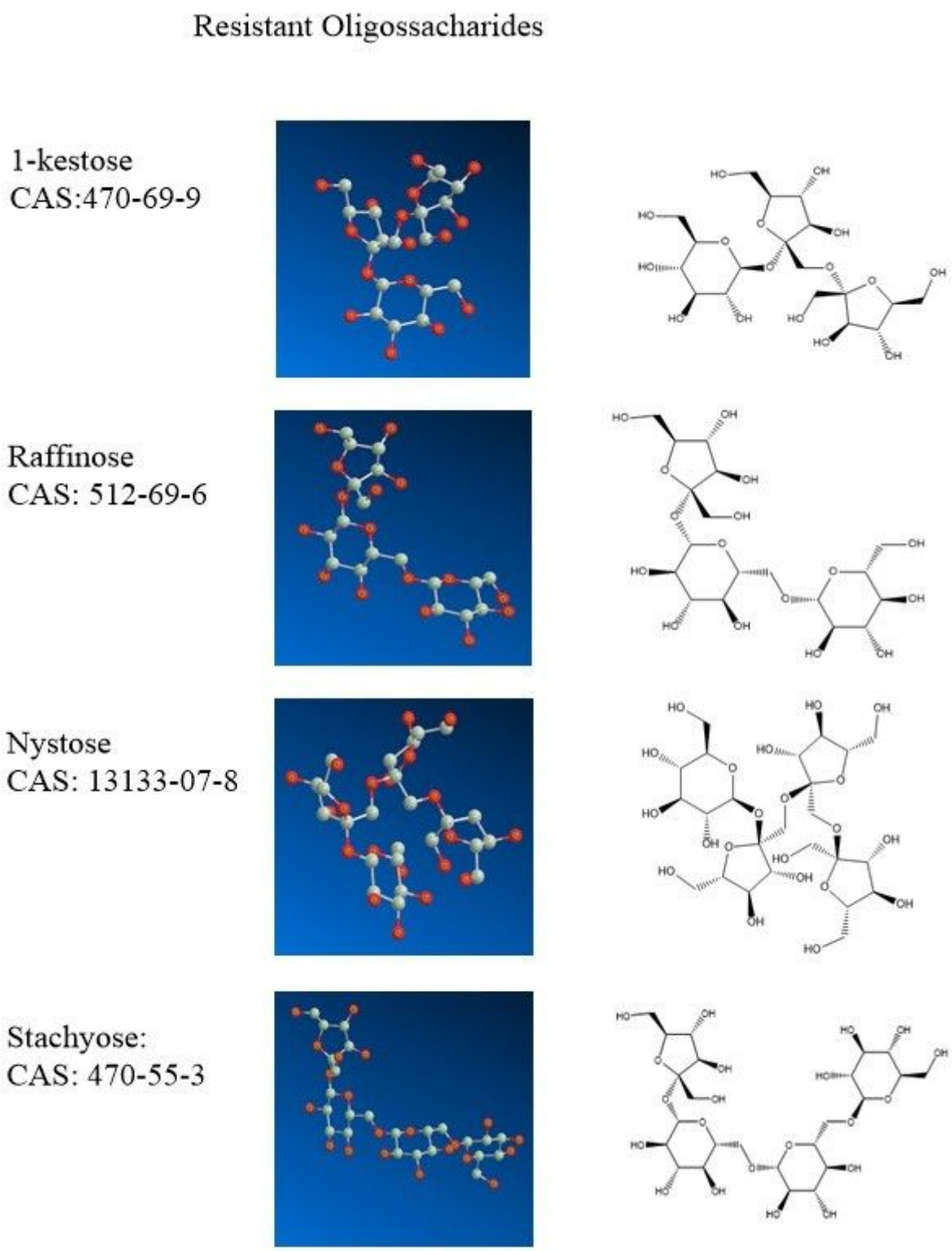

Figure 1

Fructo-Oligosaccharides. Name, CAS number, 3D structure, and 2D structure of the four fructo-oligosaccharides that are used as substrates in this study.

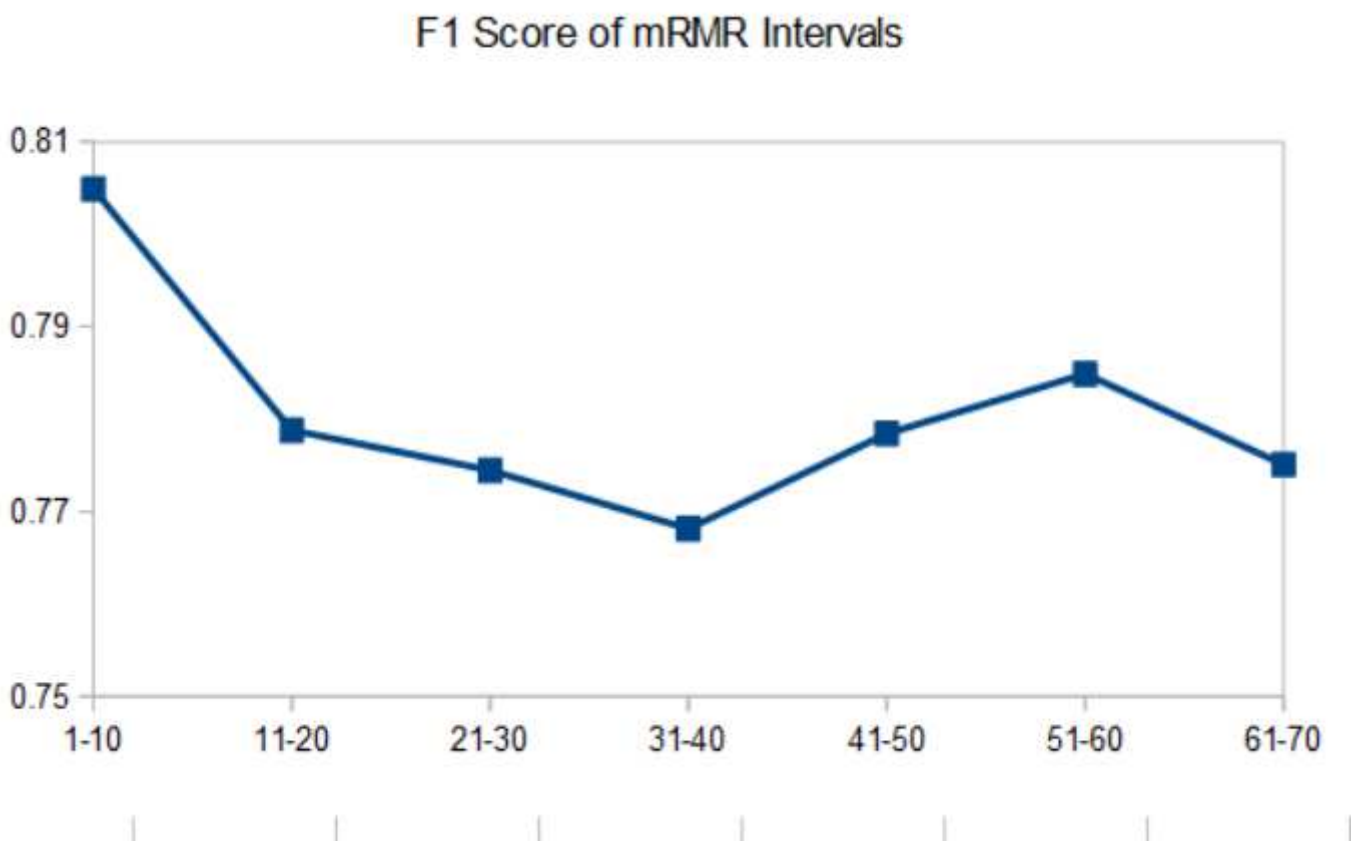

Figure 2

F1 scores of FSL models trained using each 10-feature groups arranged according to mRMR scores.

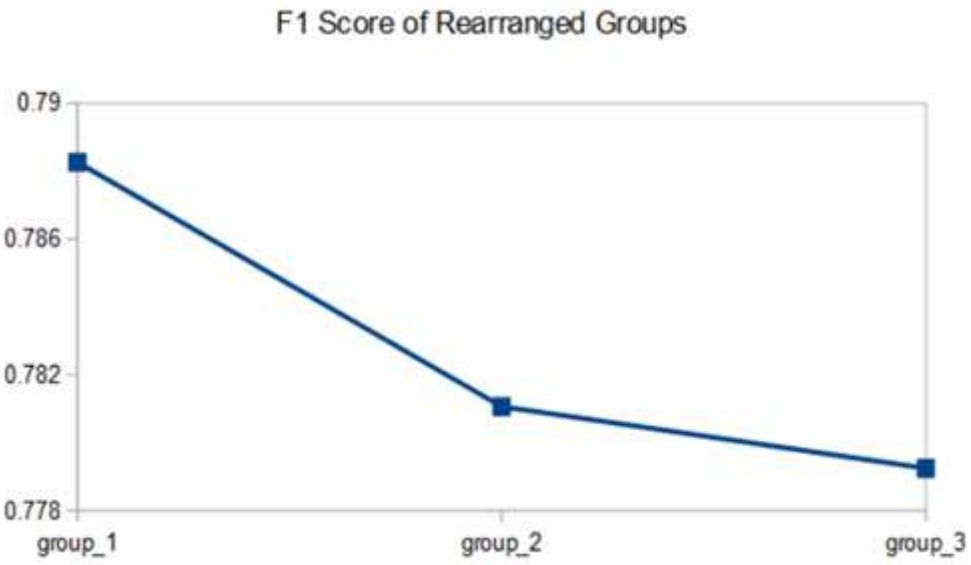

Figure 3

F1 score of FSL models inputting 20 features from three groups of mRMR intervals.

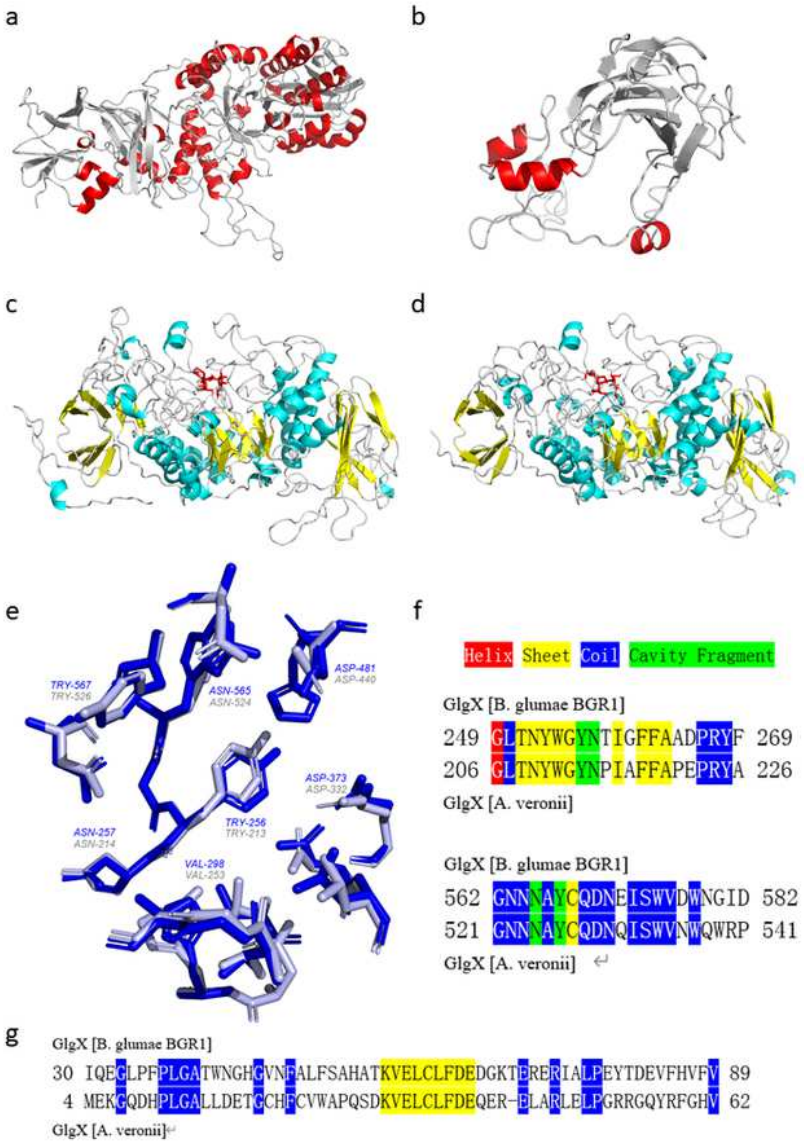

Figure 4

Structural presentation of sample proteins. a) Structure of P9 with alpha helixes in red. SSSH_01 = 12, SSSH_02 = 150. b) Structure of P6 with alpha helixes in red, SSSH_01 = 1, SSSH_02 = 35. c) Structure of anchor protein P5 with alpha helixes in blue and beta sheets in yellow; binding ligand1-kestose is shown in red. d) Structure of protein P19 with alpha helixes in blue and beta sheets in yellow; binding ligand1-kestose is shown in red. e) Structural alignment of cavity fragments of P5 (blue) and P19 (gray) using

align function in pymol with alignment RMSD = 0.389; residue labels are shown in corresponding colors. f) Sample AAS alignment of P5 (GlgX [B. glumae]) and P19 (GlgX [A. veronii]) around the two cavity fragments; aligned AA of a given secondary structure or belongs to a cavity fragment is shown in the color scheme. g) Sample AAS alignment of non-cavity parts of P5 and P19, with AA of a given secondary structure shown in the color scheme.

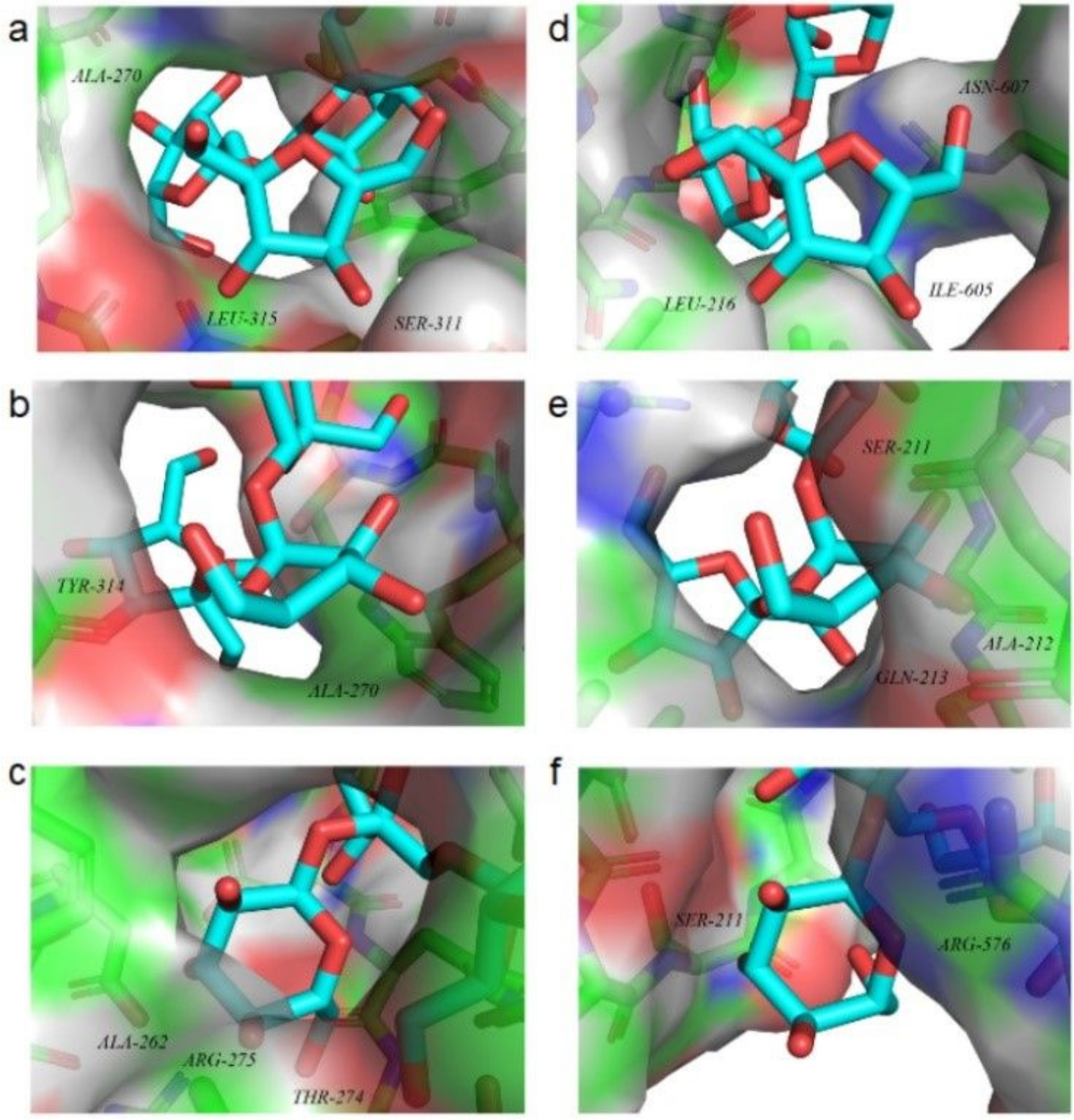

Figure 5

Predicted interaction of two CAZyme CBMs (P10 and P52) with 1-kestose. P10 and P52 are highly correlated according to group 1 features, while having no significant similarity in sequence. Similar amino acid residues are found in the radius of the interaction of both P10 and P52, including SER, ALA, ARG, and LEU. This suggests that the feature matrix may obtain inherent biological meaning. a) Interaction model of P10 viewing from beta-D-fructofuranose. b) Interaction model of P10 viewing from beta-D-fructofuranosyl residue. c) Interaction model of P10 viewing from alpha-D-glucopyranosyl residue. d) Interaction model of P52 viewing from beta-D-fructofuranose. e) Interaction model of P52 viewing frombeta-D-fructofuranosyl residue. f) Interaction model of P52 viewing from talpha-D-glucopyranosyl residue.

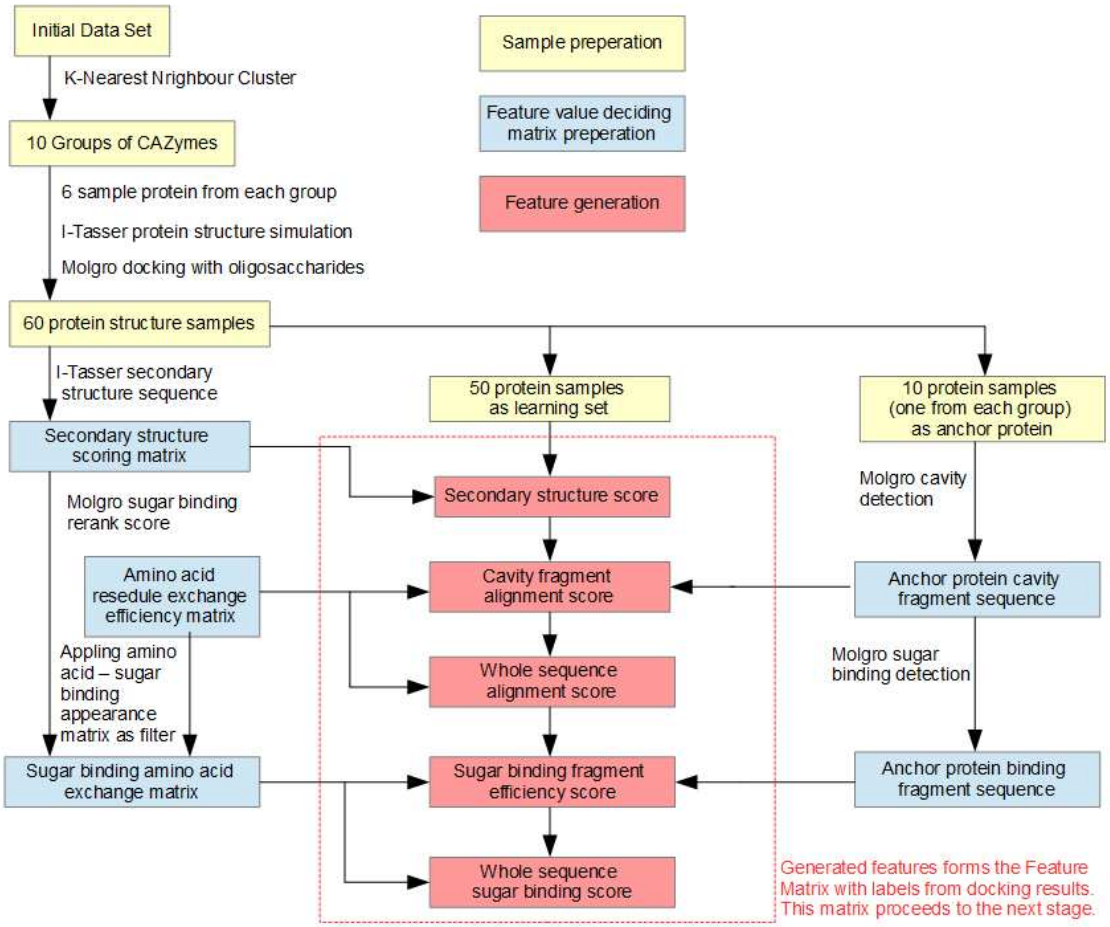

Figure 6

Flow chart for feature generation procedure. Arrows show derivation.

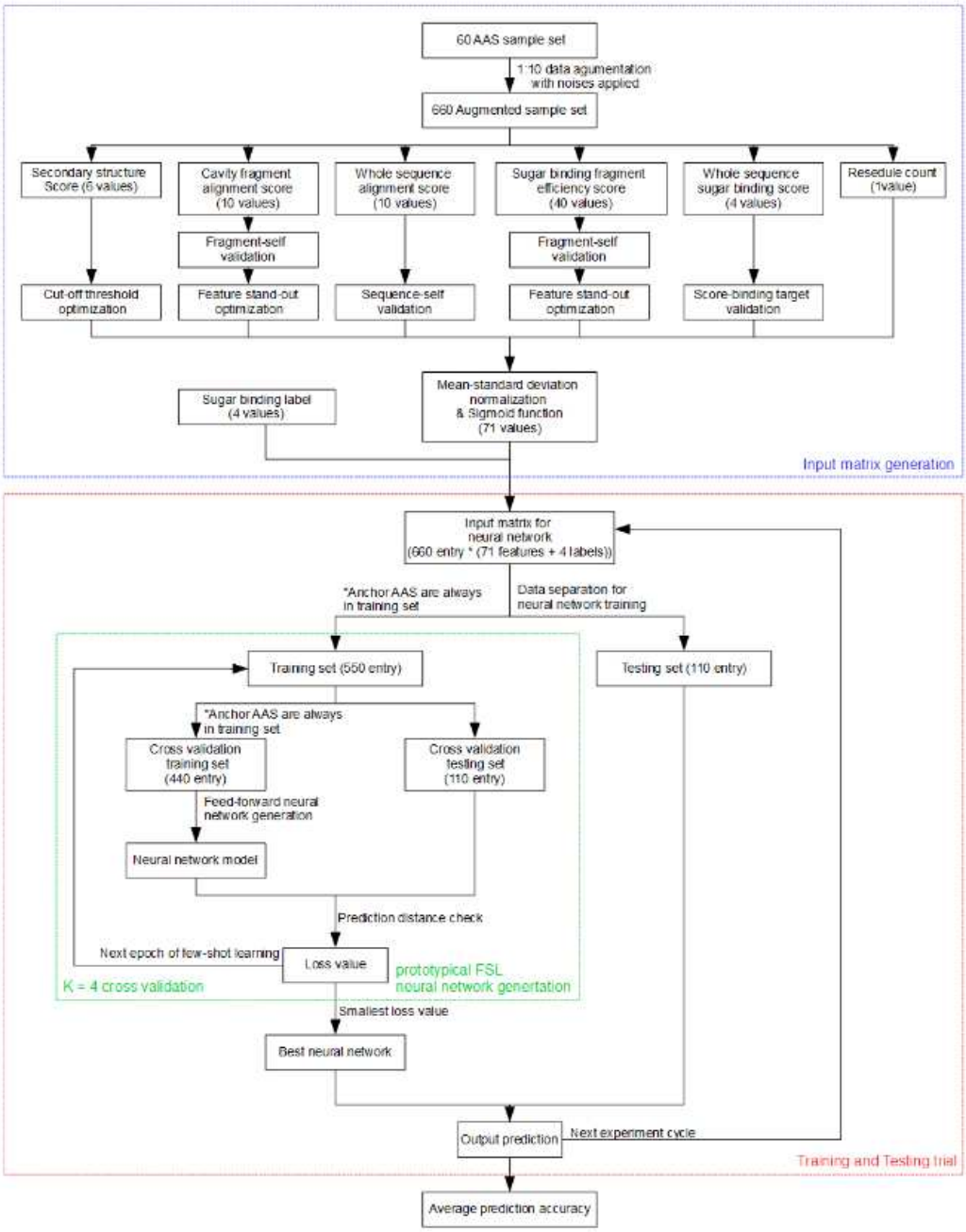

Figure 7

Flow chart of feature matrix generation and pFSLNN process

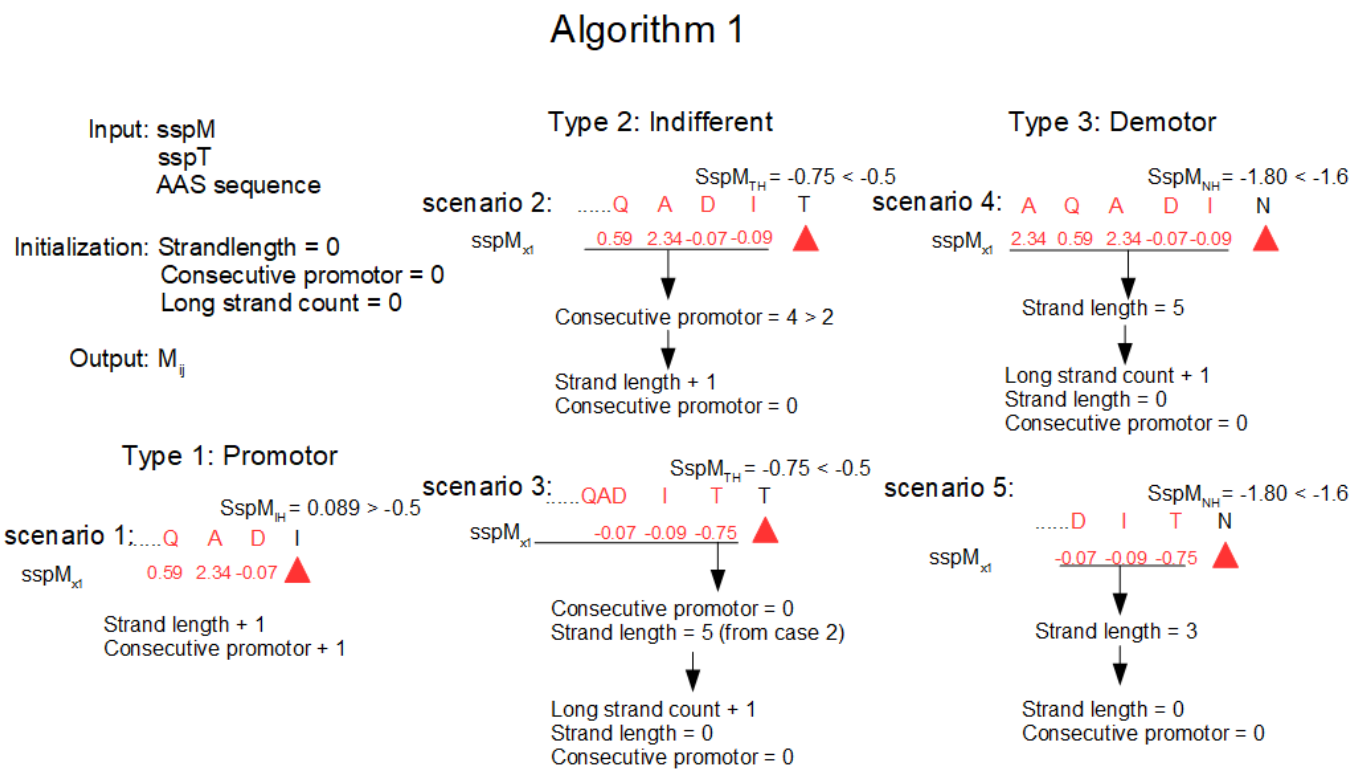

Figure 8

This figure illustrates the bases of the algorithm estimating the number of residuals of each general secondary structure type. Without loss of generality, α-Helix (H) is shown as an example. A cursor scans over the input AAS to find the first H promoting AA. A secondary structure strand of H begins. When the strand length is more than or equal to 5, the strand is considered a long strand. As the cursor proceeds, 3 types of successor (promoting, demoting, indifferent) AA can result in 5 cases. Initial sspTH = (-0.5, -1.6, 2).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplimental1.bmp
- Supplimental2.png