Predictive modelling of directed evolution for de-novo design of solid binding peptides

Saransh Shreepal Jain

A thesis

submitted in partial fulfilment of the

requirements for the degree of

Master of Science

University of Washington

2021

Committee:

René Overney

Mehmet Sarikaya

Program Authorized to Offer Degree:

Department of Chemical Engineering, College of Engineering

University of Washington

Abstract

Predictive modelling of directed evolution for de-novo design of solid binding peptides

Saransh Shreepal Jain

Chair of the Supervisory Committee:

René Overney

Department of Chemical Engineering

Genetically Engineered Polypeptides for Inorganics are the solid binding polypeptides designed to exploit their molecular specificity and binding affinity towards certain inorganic material surfaces. These solid binding polypeptides are selected using combinatorial methods such as phage display. These selections can then be optimized using directed evolution. Directed evolution involves the application of the molecular insights gained from the previous methods to evolve the activities of extant peptides and proteins. In the current thesis, we have developed an statistical approach to identify quantitative amino acid properties relevant to the binding behaviours of solid binding peptides using biopanning experimentation data for determining directed evolution trends. We have also trained machine learning models for the modelling of the binding behaviours of 12 amino acid length MoS_2 binding peptides, and for the de-novo design of sequences. In doing so, we have

developed a simple and efficient methodology for the predictive modelling of directed evolution for de-novo design of solid binding peptides. The protocols developed are expected to impact the technological applications on the peptide-single layer solid based bio/nano soft interfaces such as biosensors, bioelectronics, and potentially also medical applications.

Table of content

Chapter 1

Introduction

Proteins are large and complex molecules which are essential for the existence of life as we know it. They play a major role in almost all the diverse biological processes. This includes enzymatic reactions, structural enhancement, cell motility amongst others. This versatility in application is because of their polymeric structure of varying lengths consisting of 22 proteinogenic Amino acids (20 standard amino acids and additional 2 incorporated by special translation mechanisms) as monomer units. This results in the proteins folding in specific three-dimensional structure or conformation which gives the protein its activity. [1-3] This activity can be traced to the specific interactions between these proteins and their ligands. The molecules which display the same kind of binding site interactions can thus be considered for similar applications. Polypeptides, with less than 20-30 Amino Acid residues can be used to sufficiently mimic proteins, with the added advantage of possible modifications. These modifications can be either addition of non-proteinogenic amino acids or modifications to the very backbone of these peptides. [3-5]

Genetically Engineered Polypeptides for Inorganics (GEPI) are such short Amino Acid sequences which display a binding affinity towards solid inorganic materials. [6,7] They are also known as Solid binding peptides (SBPs) and Inorganic Binding Peptides (IBPs). [8,9] The uniqueness of these peptides is in the phenomenon of molecular recognition. That is, they can be selected to have an exceptional binding affinity towards a certain material, but none whatsoever towards others. This is done using phage display or cell surface display techniques. [10,11] These selections can be further optimised using directed evolution. Directed Evolution uses molecular insights gained from previous methods and applies them to bias the protein or peptide sequences towards selection of favourable mutations. [12] In

synthetic biology, natural evolution can prove destructive by adding a high degree of randomness into the system. [13] Directed Evolution, however, mimics natural evolution in a laboratory setting where the factors affecting it can be controlled or "directed". [14]

The problem with these approaches is that while directed evolution makes it possible to select highly functional peptides, the space of possible peptide sequences is too large for an exhaustive search. It is severely limited by the library used initially. It is only possible to find the local optima, instead of the best possible sequence for a given function. The (specifically) functional peptides are extremely scarce in this local peptide space. Furthermore, increasing the functionality makes the applicable sequences exponentially scarcer. [15,16] It should also be noted that directed evolution discards the information gained from sequences that do not exhibit the desired functionality.

It is possible to circumvent this problem by using the predictive machine learning models. Thus, it would be desirable to use these methods in the de-novo design of peptides. There have been some notable developments of machine learning algorithms for selection of functional proteins. [17] Recently, Müller et al successfully used Neural Networks to predict antimicrobial peptides with 82% of pseudo-probability for the training set and 65% for randomly sampled sequences with similar amino acid distribution. [18] In the current work the authors have attempted to identify properties that are displayed by amino acid residues in the GEPI which bind to a Molybdenum disulfide substrate, a two-dimensional atomically thick semiconductor solid. This addresses the outstanding problem of loss of interpretability. The machine learning models developed to accurately map the relationship between the position specific amino acid property descriptor values and the COAM values also help us better understand the influence of the properties on the binding phenomenon displayed. This allows us to predictively generate new peptides with the desired binding affinity to MoS2.

(See fig 1)

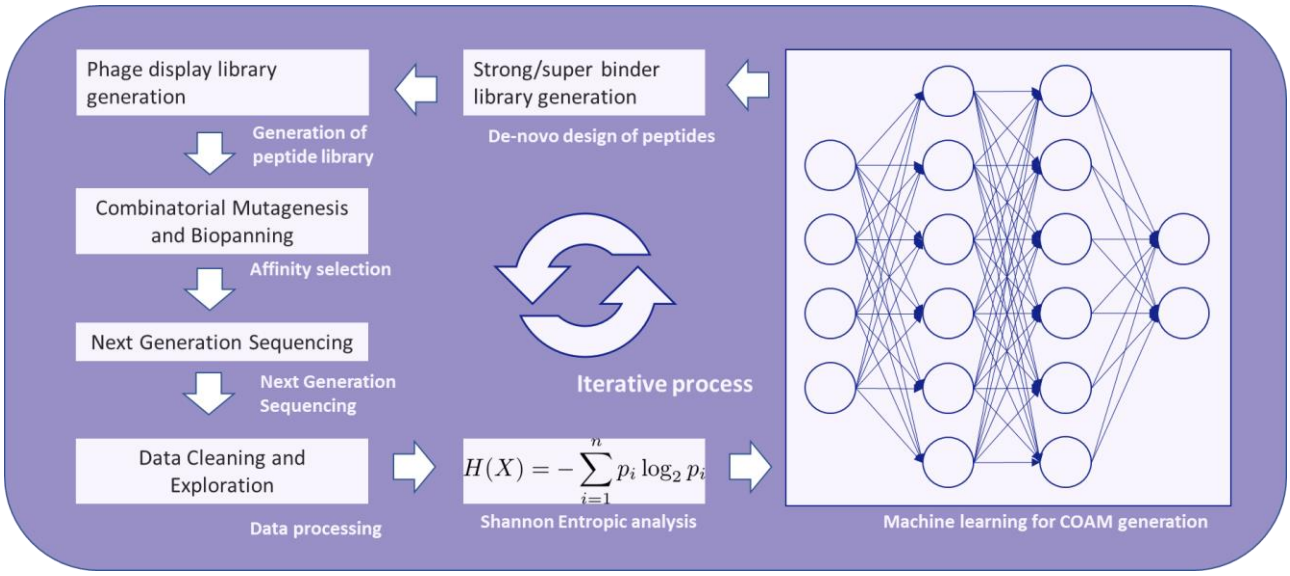


Figure 1 De-novo design of solid binding peptides. It is an iterative methodology to refine peptide libraries and increase the density of functional peptides

The GEPI-MoS_2 interaction described in the current thesis paves way for the possibility of a high-throughput design of Self-Assembled Peptide/Single Layered Atomic Material (SAP/SLAM) systems. Self-Assembly of peptides is a process in which peptides aggregate to form highly ordered 3-dimensional structures without external influences. The structures are maintained in a stable low-energy state by the non-covalent forces, namely Hydrogen bonding, hydrophobic interactions, electrostatic interactions, and van der Waals forces.[105-107] Self-assembly of biological molecules on semiconductors is fundamental to the bottom up approach of bio-electronic integration. Using such self-assembled peptides, such bioelectronic interfaces can be formed on highly controlled Single Layered Atomic Materials. [108-111] These interfaces display some of the simplest tuneable bio-nano interactions in nature.

Chapter 2

Background/Biopanning

Introduction:

Molybdenum disulfide (MoS2) is a transition metal dichalcogenide, which are materials with the structure of MX2, where M is a Transition metal (Ti, Zr, Hf, V, Nb, Ta, Mo, W, Tc, Re,Co, Rh, Ir, Ni, Pd, Pt), and X is a Chalcogen (S, Se, Te). They demonstrate semiconducting, superconducting, or metallic electronic properties.[112] Single layer MoS2 exhibits properties similar to Graphene, with the added advantage of a direct band gap of 1.8 eV and abundance in the form of molybdenite.[113] As a result of the band gap, 2D MoS2 can be used for the next generation switching and optoelectronic devices.[114] such as light emitting diodes, ultrasensitive transistors, and flexible solar cells. [104]

Previously in the GEMSEC lab, a preliminary high-throughput peptide-selection protocol was developed in which solid binding peptides were selected against a two-dimensional atomically thick molybdenum disulfide substrate generating a preliminary data-pool with about 2 million unique peptides. [100] The protocol and its results are summarized in the current chapter.

Methods:

1. **Combinatorial Mutagenesis and Biopanning:** A dodecamer Phage (M13 bacteriophage) Display library is used to select MoS2 binding peptides. For such a screening, 5 mg of MoS2 flakes are dispersed in 1 mL of potassium phosphate/sodium carbonate buffer (PC, 55 mM KH2PO4, 45 mM Na2CO3, and 200 mM NaCl, pH 7.4), containing 0.02% Tween 20 detergent. 10 μL of the library is incubated in the mixture for 3 hours at room temperature and washed before overnight incubation. The

flakes are then washed successively with the PC buffer, and 0.1%, 0.2%, and 0.5% of the Tween 20/80 detergent to screen the weakly bound phages into Wash 1, Wash2, and Wash3. The remaining phages are then washed from the surface using an elution buffer consisting of 0.2 M Glycine-HCl (pH 2.2) into an Eluate. Each phage pool is then neutralized and purified via PEG/NaCl precipitation and resuspended in de-ionized water. [100]

2. **Washes:** The experiments were designed such that the peptides selected had an increasing binding affinity towards MoS_2 surface from wash 1 to wash 2 to wash 3. These are designed to completely disrupt these bonds. They could potentially display an evolution towards the selection of strong binding peptides. This binding affinity only addresses one or more specific binding mechanism explored by the detergent in the biopanning experiment. [22-24] Thus, the eluate includes only the binding peptides which employ all the other kinds of mechanisms.
3. **DNA Isolation and Next-Generation Sequencing:** The single-stranded DNA (ssDNA) of the M13 bacteriophage is isolated from wash and eluate phage pools. The sequencing library is prepared by amplifying the 36 bp peptide-coding variable region using Q5 polymerase (New England Biolabs). Sequencing adapters containing p5 and p7 index sequences are attached using a second PCR. The purified PCR products from each phage pool are loaded in duplicates on the sequencer plate and sequenced nusing Next Generation Sequencing (NGS). NGS is a high-throughput approach to DNA sequencing. It uses miniaturized and parallelized platforms for sequencing of 1 million to 43 billion short reads (50-400 bases each) per instrument run. After combining with the other replicates, the data was translated into amino acid sequences to generate peptide datasets. [100]

Results:

1. **DNA Isolation and Next-Generation Sequencing:** The NGSThe Next-Generation Sequencing (NGS) platform yielded about 288 million DNA sequences in total. Upon processing more than 2 million unique peptides were identified with varying copy numbers that survived in the successively stringent washes. [100] The peptides are generated in form CSV (Comma Separated Value) files that contain the sequence data from the clusters that pass filter on a flow cell. Three csv were generated and used for the analysis.

Chapter 3

Shannon Entropic Analysis

Introduction:

The biopanning dataset of peptides distributed through the washes holds a treasure trove of information about the macro-behaviours of peptides when interacting with the materials and the solvents. It was used to extract information about the binding phenomenon to gain knowledge about the binding mechanism(s) disrupted in the experiments. It is desired to discern the influence of properties on the manipulate peptide sequences for a preferred binding effect (sticking or antifouling).

Methods:

1. **Dataset:**

1.1. Sets of washes: Based on the increasing binding affinity with respect to a binding mechanism, the sets of washes were defined as follows:

1. Wash 1 only: The peptide sequences which are exclusive to wash 1.
2. Wash 1+2: The peptide sequences exclusively common to both washes 1 and 2. The sequences present only in wash 1 or only

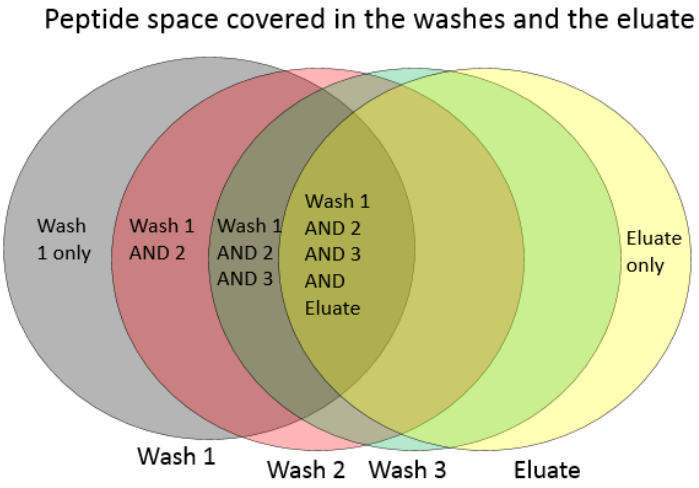


Figure 2 Peptides space and washes considered. The sections from 'Wash 1 only' to 'Wash 1 AND 2 AND 3' include the pools representing the continuous peptide space with increasing binding affinity with respect to the mechanism explored. 'Eluate only' specifically included peptides which display all the other mechanisms, and 'Wash 1 AND 2 AND 3 AND Eluate' includes peptides which could be disrupted by both kinds of washing.

in wash 2 were not included. Similarly, sequence present in wash 3 or eluate were also discarded.

3. Wash 1+2+3: The peptide sequences exclusively common to washes 1, 2, and 3. The sequences present only in wash 1, only in wash 2, or only in wash 3 were not included. Similarly, sequence present in eluate were also discarded.
4. All Washes: The peptide sequences exclusively common to washes 1, 2, 3 and the eluate. The sequences present only in wash 1, only in wash 2, only in wash 3, or only in the eluate were not included.
5. Eluate only: The peptide sequences which are exclusive to the eluate.

(See fig 2)

1.2. Data Cleaning: The files were cleaned to remove repetition of the peptide sequences. Any sequence which was not exclusive to the 'set' of washes being considered was removed. (More on this in Section 2.2) Similarly, any repeating sequences within a set were removed such that a sequence only appears once.

2. Entropic Analysis:

2.1. Shannon's Information Entropy: Information entropy (see fig 3) is the average rate at which information is produced by a stochastic source of data.

$$H = -\sum P_i \log_2 P_i$$

Figure 3 Information Entropy

Using Entropy, it is possible to understand for the existing sequences, the propensities for exhibition for certain properties, given the abundance or scarcity of an amino acid at a given position. [21]

Entropies were calculated at each position in the polypeptide sequences with respect to Amino Acids and their properties, for each set.

2.2.Property descriptors: 527 structural properties (See Table B1, Appendix B) were quantified with respect to their influence at each of the positions. Properties such as charge, aromaticity, hydrophobicity are of particular interest. This data had been compiled from research articles studying amino acids and amino acid residues in peptides and proteins. At a glance, many of these properties displayed similar values and trends with respect to Amino acid. Upon investigation, many of these properties showed a high Pearson correlation (see fig 4). Pearson Correlation helps identify the linear correlation between any datasets. It was desired to eliminate highly correlated properties as they are assumed to describe the same phenomenon. Eliminating the extra datasets increases the efficient packing of peptide data. Thus, the sets of property descriptors with more than 90% correlation were identified. The 90% threshold was chosen arbitrarily. If multiple property descriptors were similar, it was assumed that they are describing the same underlying phenomenon. The properties which showed the correlations were disposed such that only one of the properties representing the assumed description remained. Thus, the list was reduced to 292 unique property descriptions (Available in Github).

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

Figure 4 Pearson Correlation

2.3. Identification of relevant properties:

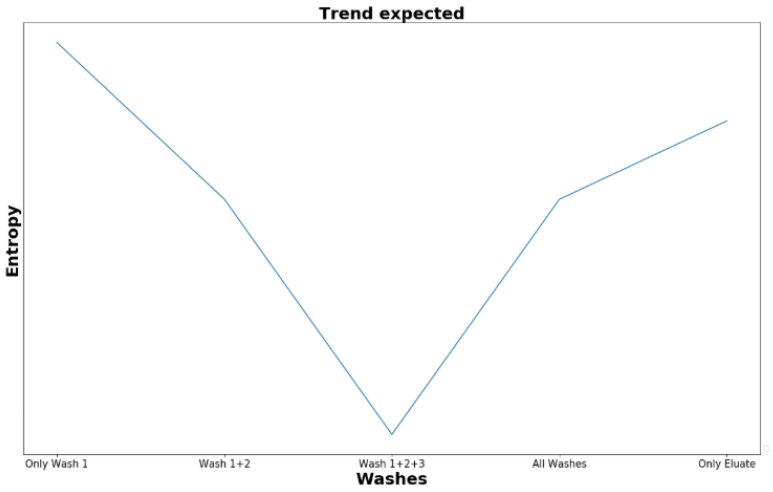


Figure 5 Trend that shows an evolution (from Only Wash 1 to Wash 1+2+3) Only the Entropies displayed by 'Wash 1+2+3' is of importance. It must be lower than all the other values. The relationship between other values is inconsequential to the current study, but not irrelevant in the information displayed about the distribution of influence of the binding mechanisms.

Based on the insights gained from the Shannon's Information Entropy for the properties, it is possible to identify the properties at the positions which display a relationship with the binding affinity. These relationships may not necessarily be of a causal nature. The experiments are designed such that for properties which affect the binding, the entropy would show a specific trend for the sets. The entropy for the 'Only Eluate' represents the peptides binding to the MoS_2 substrate using all the other binding mechanisms, and as a result must be higher than the 'Wash 1+2+3' and also for the 'Only Wash 1' set which consists of all the weak and the non-binders, without any particular specificity. The entropy should be the lowest for the third set (wash 1+2+ 3) as it indicates an evolution or selection of peptides which display the property. (See fig 5)

Results and Discussion:

Identification of relevant properties: Based on the entropic analysis for all the sequences, about 196 out of the 527 property parameters were identified which showed any correlation for any of the positions. [24-101] These properties were chosen based on the threshold difference of the entropy for 'Wash 1+2+3' with respect to the other entropies. The threshold was arbitrarily set as 0.02 to indicate a difference of about 10%.

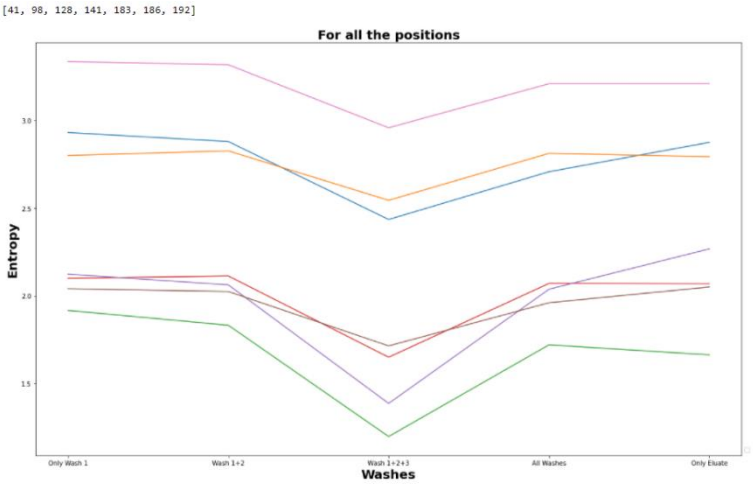


Figure 6 Property Entropy graph for all the positions. The property numbers of the properties which were found to be related to GEPI binding to MoS2 are mentioned on the top left corner. For more on the properties see Table 1.

About 7 of these properties show a strong correlation, enough to display the trend for the entire peptides with a threshold of 0.24 (See table 1, See figure 6) The influence of different properties seems to be localized to different position in the 12 Amino acid length sequences. It must be noted that the property parameters identified may be causal or correlational. The graphs and the list of the properties for the entropic analyses can be found in the Appendix A and B.

Table 1 Properties which display a high correlation to the binding mechanism(s) of GEPI to MoS_2.

Property index	Property	Reference
83	Consensus normalized hydrophobicity scale	[51]
262	Weights for alpha-helix at the window position of 6	[82]
282	Information measure for middle helix	[83]
345	Weights for beta-sheet at the window position of -3	[82]
351	Weights for beta-sheet at the window position of 3	[82]
357	Beta-sheet propensity derived from designed sequences	[85]

Chapter 4

Predictive Modelling

Introduction:

Peptide space denotes a permutational space of natural amino acid residues. The space is vast and functionally sparse, making it practically impossible to generate all the sequences required to understand global peptide behaviors and interactions with other materials. The biopanning data generated previously is large yet limited as compared to the expanse of the peptide space. Having identified the relevant properties using Shannon Entropic Analysis, we can leverage the data generated and deep learning to identify the underlying directed evolution trends from the large yet locally limited peptide binding datasets. Using machine learning tools can help us design peptides residing in a much larger expanse of the global peptide space at a fraction of the computational and monetary costs.

Methods:

1. **Dataset:**
    - 1.1. **Center of Abundance Mass:** Using centre of abundance mass allows for a better quantification of the binding mechanisms and strengths as displayed by individual peptides. Centre of Abundance mass is the weighted average of the distribution of any given peptide within the washes and eluate. It is the sum of the product of a peptide population in the wash and wash index (0: Wash 1, 1 Wash2, 2: Wash 3, 3: Eluate) divided by the total population of the peptide. The Center of Abundance Mass metric allows for a quantitative representation of peptides included in more than one washes. It follows a Gaussian distribution, with numerous outliers. [1]
    - 1.2. **Property Descriptors:** 527 quantitative structural properties were used to quantify the relative influence of different amino acid residues at each of the

positions. Multiple sets and forms of these 527 properties were used to determine the effectiveness of the property descriptors.

Table 2 Sets of property descriptors

Number of Properties	Description
527	All quantitative structural properties
196	In the previous article, Shannon Entropy was used to identify properties which showed correlation to the binding phenomenon displayed in the biopanning experiments. Out of the 527 properties, only 196 showed correlation to the binding mechanism.
292	The property data has been compiled from multiple research articles studying amino acids and amino acid residues in peptides and proteins. The sets of property descriptors with more than 90% correlation were assumed to be describing the same underlying phenomenon and were disposed such that only one of the properties representing the assumed description remained. Out of the 527 properties, only 292 were found to display less than 90% correlation.
111	Only the properties which showed correlation to the binding mechanism and less than 90% correlation with each other. This is an intersection of the 196 and 292 properties identified above.
20	Most of the 527 properties were found to be correlated with each other. Thus, to describe the unique trends underlying the property descriptors, 20 principle components were calculated using Principle Component Analysis. The components are orthogonal, and thus do not have any correlation amongst each other.

1.3.Input datasets: Each peptide sequence is first converted into a two dimensional matrix [Position x Amino acid property descriptor] where each element represents a normalized (on the scales of 0 to 1) property descriptor value for the amino acid residue at any given position within the peptide. This matrix is then reshaped into a one-dimensional feature. Multiple features are stacked together to create a two-dimensional input matrix [Number of peptides x Position-specific amino acid residue property].

1.4. Output/Target dataset: The center of abundance mass values are normalized on the scale of 0 to 1. The output is a one-dimensional vector [Number of peptides] consisting of center of abundance mass values for each peptide.

2. Data Processing:

2.1. Training and Test set: The datasets were divided into a training set and a test set. Approximately 10% (16569 peptides) of the dataset is saved as the test set, and the rest 90% (150061) is used as the training set. This is done to test the generalisability of the models for unseen data, to evaluate the bias, and reduce the overfitting of the training dataset.

2.2. Classification: The dataset is divided into multiple classes to separate the peptides based on their binding characteristics. The COAM follows a Gaussian distribution with a mean of 1.1288 and a standard deviation of 0.2877. The peptides with COAM lower than two standard deviations below the mean are classified as weak, whereas the peptides with COAM higher than two standard deviations above the mean are classified as strong. The peptides with COAM within the two standard deviations around the mean are classified as medium binders. The outlier peptides found exclusively in the eluate pool are classified as super binders. These are the peptides with COAM of 3.0.

2.3.Data imbalance problem: The center of abundance mass metric used to quantify peptide binding strength follows a Gaussian distribution (see figure 7). Approximately 95% of the peptides by design fall under 'medium' classification. About 2.5% of the peptides fall under 'weak' and 'strong' classifications each. The 'super' classification covers less than 0.002% of the peptides. This imbalance in the classification of the datapoints leads to bias in the machine learning algorithms. It is possible to address the data imbalance by improving the ratio of the peptides in the classifications.

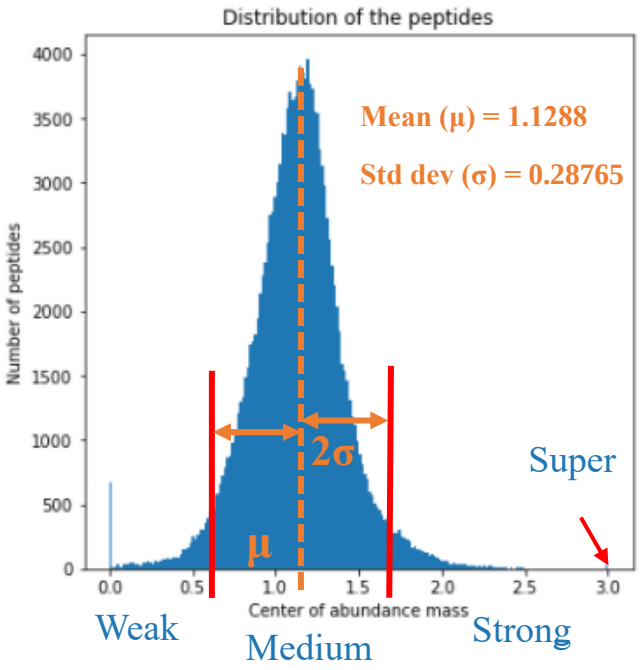


Figure 7 Distribution of Center of Abundance mass displayed by the peptides. The distribution is Gaussian with two sets of outliers (at COAM 0 and 3)

2.4.Augmentation: In this case we have increased the number of weak, strong, and super binders, by multiplying these datasets and reduced the medium binders by randomly removing half of the dataset. A small noise is randomly introduced into the input of the multiplied dataset to avoid memorization of values. It must be noted that the noise values added are gaussian in distribution to mitigate an accretion of errors.

Furthermore, the entire dataset was duplicated with reverted sequence data. It is assumed that the 'macro' binding behaviour does not change significantly with the direction of peptide. (See fig 8)

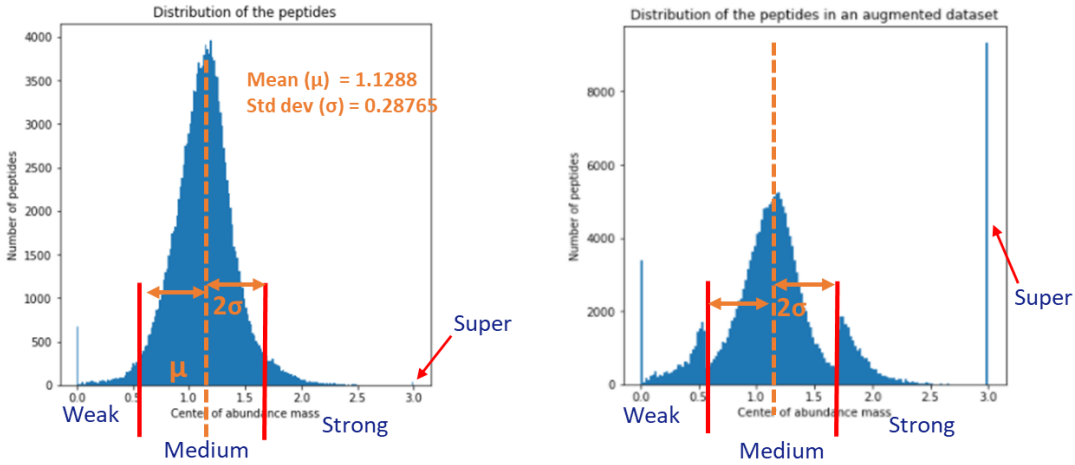


Figure 8 Typical data augmentation. Left: Distribution of peptides before augmentation, Right: Distribution of peptides after augmentation.

2.5.Normalization: The property descriptors mined from various articles had values as high as 707 and as low as -205. It also had values from the order of 10^(2) to the order of 10^(-3). With input values existing on different scales, the associated parameters will also exist on different scales. It becomes difficult for machine learning algorithms to iteratively achieve values closer to accurate parameters. Standardization of the values allows us to retain variance information for the amino acids in the case of any property, while reducing the disparity of the scale between different properties. Using normalization for multiple linear regression also increases interpretability by estimating the influence of any given property and position on the binding behaviours.

2.6.K-fold cross validation: K-fold cross validation is a resampling procedure used to evaluate the efficiency of a model on unseen data samples. In this procedure, the parameter k refers to the number of groups into which the dataset is split randomly. In every iteration, one of the groups is held back as a validation set, and the model is trained with the rest of the groups. It is then tested against the validation set to evaluate the efficiency of the model. In the current models, 5-fold

validation is employed because of limited weak, strong, and super datasets. The training set is thus split into training and validation sets with a 4:1 split.

3. Predictive modelling:

3.1. Multiple Linear Regression:

Multiple Linear Regression is a supervised machine learning algorithm which models a linear relationship between multiple explanatory variables and a response or dependent variable. [101]

Architecture: In the current project, Multiple Linear Regression models are used with slight modifications. An activation filter (ReLU) is used to keep the COAM values more than or equal to zero. (fig 9) The loss function used was Mean Squared Error. MSE is average of the squares of the difference between the predicted output values and the actual/target output values. Multiple models were created with only selective datasets to only distinguish between two classes at a time. Models trained are displayed in Table 3.

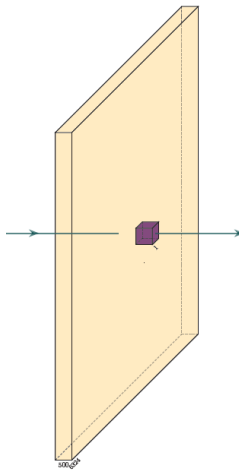


Figure 9 Model Architecture. A Multiple Linear regression followed by a ReLU filter.

Table 3 Multiple Linear Regression Models trained to discern different classes of peptides

MLR model	Description
All peptides model	A multiple linear regression model trained using all peptide datasets. It is used to predict weak, medium, strong, and super binders.
SS model	A multiple linear regression model trained using super and strong binding peptide datasets. It is used to predict super binders.
SM model	A multiple linear regression model trained using strong and medium binding peptide datasets. It is used to predict strong binders.
MW model	A multiple linear regression model trained using medium and weak binding peptide datasets. It is used to predict weak binders.

3.2.Directed Acyclic Graph: Directed Acyclic graph (DAG) is a directed graph with no cycles. In the DAG, the Multiple Linear Regression models were arranged in a topological order. The classifications are made in the hierarchy is given a preference in the order of the SS model first, then the SM model, and finally the MW model. The SS model is used to predict 'super' binding peptides. The peptides which are not classified as 'super' are passed through the SM model which filters out the 'strong' binding peptides. Finally, the remaining peptides are passed through the MW model which classifies 'weak' binding peptides. The rest are automatically labelled as 'medium' binding peptides. (See fig 10) We created multiple Directed Acyclic Graphs using the property sets described earlier. (See Table 2)

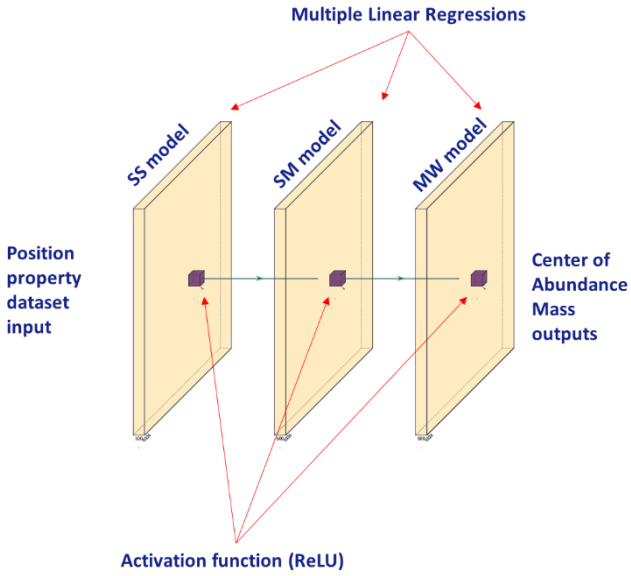


Figure 10 Directed Acyclic Graphs. A hierarchy of three MLR models wo consecutively filter out peptides with decreasing predicted binding affinity towards MoS_2

4. **Comparison with Shannon Entropic Analysis:** The weightage parameters for models created were screened to identify the properties with highest relative influence on any individual position. In MLR, the weightage parameters correspond to the influence of the feature vector on the predicted output. The thresholds were adjusted to identify the top three properties in each of the models. These properties were then compared with the properties identified from Shannon Entropic Analysis for cross-validation.

5. **Recommender System:** A simple system is used to determine the strongest binding peptides. An iterative method is used to identify two or more the best performing amino acids at each given position. This is done by keeping the amino acid residues at all the other positions constant and iterating over only a single position at a time. This is repeated for all the positions. Using these sets of amino acids, thousands of sequences can be generated. This is only possible because the effects of neighbouring or local residues are not considered.

Results and Discussion:

1. Predictive modelling:

1.1. Model loss: The Directed Acyclic Graph models were tested against test peptide datasets and the losses are shown in table 3. Even as fewer properties are used, the loss increases only moderately suggesting a higher efficiency in the training of the models. It must be noted that the average 24.5% increase in loss for 20 property component DAG may be attributed to orthogonality. Unlike the previous ones, the features for this dataset are independent. As the binding related phenomena may only represented once in the dataset, it becomes easier for the model to be stuck in local optima, making the 20 property component comparatively inaccurate for COAM predictions.

Table 4 Loss of the Multiple Linear Regression models in the different Directed Acyclic Graphs

	Loss in SS model	Loss in SM model	Loss in MW model
527 DAG	0.0141	0.0147	0.0147
292 DAG	0.0142	0.0147	0.0179
196 DAG	0.0147	0.0148	0.0147
111 DAG	0.0156	0.009	0.0162
20 DAG	0.0196	0.0169	0.0181

1.2. Class predictions: When tested for peptide classification, the Directed Acyclic Graph display a high true positive rate for the unseen test datasets. The density of strong and super binding peptides is increased to up to 12.36% in the predicted peptides, as compared to 2.74% in the original dataset. Similarly, the density of weak binding peptides is increased from 2.48% in the original dataset to up to 7.81% in the predictions.

There are only 3 super binding peptides in the test dataset which hinders our ability to test the classification of super binding peptides effectively. However, two of the three peptides have been classified correctly in some of the models (one peptide in four models and the other in one model), suggesting that the models are able to learn some correlations between peptides in the training set. Despite the higher information entropy in the eluate, this provides support for further investigation into the correlations between super peptides.

1.3. **Weightages for positions and properties:** While the 20 property DAG performs worse on the accuracy as compared to the others, it provides a much higher potential for interpretability. The orthogonal nature of the property components makes it possible to understand the influence of each phenomenon more accurately.

The positions of importance are identified using the weightage parameter. Two of the top three properties in the WM model were the same as the two of the top three properties in the MS model. None of the top three properties for SS model were found in the WM and MS models. This is consistent with the idea that the peptides within the washes 1 to 3 display the same binding phenomenon, whereas the peptides in the eluate pool are substantially different from the others.

2. **Comparison with Shannon Entropic Analysis:** The weightage parameters for models created were screened to identify the properties with highest relative influence on any individual position. In MLR, the weightage parameters correspond to the influence of the feature vector on the predicted output as shown in fig 11. The thresholds were adjusted to identify the top three properties in each of the models. These properties were then compared with the properties identified from Shannon Entropic Analysis for cross-validation. Top three property components displayed by the SS models and the MW

models, and two out of the top three property components displayed by the SM models showed relevance to the binding mechanism.

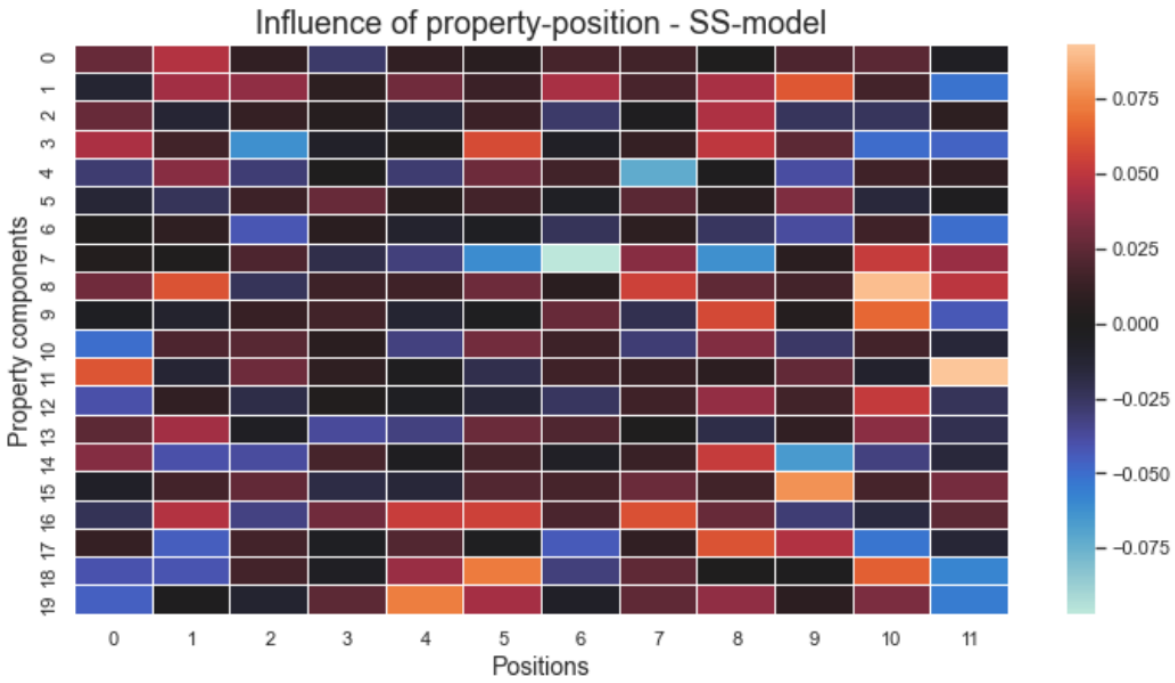


Figure 11 Influence of a property and a position on the COAM values based on SS model

3. **Recommender systems:** Up to 12000 possible super binding peptides were generated using the recommender systems. The peptides are not experimentally validated. Based on the statistics from class predictions, it can be estimated that there may be more than 36 super binders and 344 strong binders in the set of generated peptides. It must be noted that the current recommender systems do not take residue interactions into account. Thus, the influence of any amino acid properties is strictly local and additive on the COAM value.

Chapter 5

Conclusion and Outlook:

We have developed a simple and efficient methodology to statistically identify the phenomenon relevant to the binding behaviours of solid binding peptides, and to predictively model the directed evolution for the selection of solid binding peptides using structural property descriptors. Using the models, hundreds of super binding peptide candidates were generated for binding with Molybdenum Disufide. The methodology can be used to design peptides against practically any material.

The training dataset has 31 super binders as compared to about 150030 other peptides. Similarly, the test dataset has only 3 super binders as compared to about 35656 other peptides. This is a limited dataset and cannot be leveraged robustly to validate a generalized MoS2 super binding peptide generator. However, the models were successfully able to identify 2 out of the 3 super binders from the test dataset. These identifications present evidence of possible generalisation of super binding trends. It enables us to look further into the binding interactions of peptides in eluate despite the higher Shannon entropy in the pool.

The Multiple Linear Regression based Directed Acyclic Graph model offers a simplistic approach to the COAM prediction. It does not allow for the accommodation of residue interactions within the peptide, but only accounts to their hyper local effect on the position they occupy. However, if used in tandem with techniques such as the Generalised Similarity Metric [102] or a metric such as the co-evolvability [103]**,** the models created can have much higher strong and super binder prediction accuracy.

Acknowledgements:

This research was carried out at GEMSEC, Genetically Engineered Materials Science and Engineering Center, University of Washington. I thank Siddharth Rath for careful mentoring throughout the MS thesis, and Tyler Jorgenson for helpful discussions.

Data Availability

Code Availability at GitHub: https://github.com/jainsaransh/GEPI-designer, which contains all of the software developed, cleaned and curated data relevant to this work.

References:

[1] Gutteridge A, Thornton JM. Understanding nature's catalytic toolkit. Trends in Biochemical Sciences. 2005 Nov;30(11):622-629. DOI: 10.1016/j.tibs.2005.09.006.

[2] Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walters P (2002). "The Shape and Structure of Proteins". Molecular Biology of the Cell; Fourth Edition. New York and London: Garland Science. ISBN 978-0-8153-3218-3.

[3] Ambrogelly, A., Palioura, S. & Söll, D. Natural expansion of the genetic code. Nat Chem Biol 3, 29–35 (2007). https://doi.org/10.1038/nchembio847

[4] Groß A, Hashimoto C, Sticht H and Eichler J (2016) Synthetic Peptides as Protein Mimics. Front. Bioeng. Biotechnol. 3:211. doi: 10.3389/fbioe.2015.00211

[5] Jason B. Hedges and Katherine S. Ryan, Biosynthetic Pathways to Nonproteinogenic α-Amino Acids Chemical Reviews Article ASAP DOI: 10.1021/acs.chemrev.9b00408

[6] Sarikaya et al. (2007) Genetically engineered polypeptides for inorganics: A utility in biological materials science and engineering. Mat Sci and Engg C 27:558–564

[7] Tamerler et al. (2010) GEPI-Based Biological Routes to Technology. Biopolymers (Peptide Science) 94:78-94

[8] Andrew Care, Peter L. Bergquist, Anwar Sunna, Solid-binding peptides: smart tools for nanobiotechnology. Trends in Biotechnology, Volume 33, Issue 5,2015, Pages 259-268, ISSN 0167-7799, doi:10.1016/j.tibtech.2015.02.005

[9] Ersin Emre Oren, Candan Tamerler, Deniz Sahin, Marketa Hnilova, Urartu Ozgur Safak Seker, Mehmet Sarikaya, Ram Samudrala, A novel knowledge-based approach to design inorganic-binding peptides, Bioinformatics, Volume 23, Issue 21, 1 November 2007, Pages 2816–2822, https://doi.org/10.1093/bioinformatics/btm436

[10] Sang Yup Lee, Jong Hyun Choi, Zhaohui Xu, Microbial cell-surface display, Trends in Biotechnology, Volume 21, Issue 1, 2003, 45-52, https://doi.org/10.1016/S0167-7799(02)00006-9.

[11] Tamerler, C., Khatayevich, D., Gungormus, M., Kacar, T., Oren, E.E., Hnilova, M. and Sarikaya, M. (2010), Molecular biomimetics: GEPI-based biological routes to technology. Biopolymers, 94: 78-94. doi:10.1002/bip.21368

[12] Jäckel C, Kast P, Hilvert D (2008) Protein design by directed evolution. Annu Rev Biophys 37:153-173.

[13] Michael J Dougherty, Frances H Arnold, Directed evolution: new parts and optimized function, Current Opinion in Biotechnology, Volume 20, Issue 4, 2009, Pages 486-491, https://doi.org/10.1016/j.copbio.2009.08.005.

[14] Arnold FH (2018) Directed evolution: bringing new chemistry to life. Angew. Chem. Int. Ed. 57:4143-4148.

[15] Mandecki, W. The game of chess and searches in protein sequence space. Trends Biotech 16, 200–202 (1998).

[16] Smith, J. M. Natural selection and the concept of a protein space. Nature 225, 563 (1970).

[17] Yang, K.K., Wu, Z. & Arnold, F.H. Machine-learning-guided directed evolution for protein engineering. Nat Methods 16, 687–694 (2019). https://doi.org/10.1038/s41592-019-0496-6

[18] Alex T. Müller, Jan A. Hiss, and Gisbert Schneider (2018). “Recurrent Neural Network Model for Constructive Peptide Design”. Journal of Chemical Information and Modeling. 58 (2), 472-479. DOI: 10.1021/acs.jcim.7b00414

[19] Rumelhart, D., Hinton, G. & Williams, R. Learning representations by back-propagating errors. Nature 323, 533–536 (1986). https://doi.org/10.1038/323533a0

[20] Jaeger, H. (2001), "The "echo state" approach to analysing and training recurrent neural networks-with an erratum note," Bonn, Germany: German National Research Center for Information Technology GMD Technical Report, 148
[21] "A Mathematical Theory of Communication", Claude E. Shannon, Bell Telephone System Technical Publications, 1948. http://cm.bell-labs.com/cm/ms/what/shannday/paper.html
[22] Verkhivker GM, Bouzida D, Gehlhaar DK, Rejto PA, Freer ST, Rose PW. Complexity and simplicity of ligand-macromolecule interactions: the energy landscape perspective. Curr Opin Struct Biol. 2002 Apr;12(2):197-203. Review. PubMed PMID: 11959497.
[23] Tsai CJ, Kumar S, Ma B, Nussinov R. Folding funnels, binding funnels, and protein function. Protein Sci. 1999 Jun;8(6):1181-90. Review. PubMed PMID:10386868; PubMed Central PMCID: PMC2144348.
[24] Mechanisms of Protein Assembly: Lessons from Minimalist Models, Yaakov Levy and and José N. Onuchic* Accounts of Chemical Research 2006 39 (2), 135-142 DOI: 10.1021/ar040204a
[25] J Theor Biol. 1966 Nov;12(2):157-95. Relations between chemical structure and biological activity in peptides. Sneath PH. DOI: 10.1016/0022-5193(66)90112-3 PMID: 4291386 [Indexed for MEDLINE]
[26] Principal property values for six non-natural amino acids and their application to a structure-activity relationship for oxytocin peptide analogues by WOLD, S ; ERIKSSON, L ; HELLBERG, S; et al. 1987, Vol 65, Num 8, pp 1814-1820 Article
[27] Grantham R. Amino acid difference formula to help explain protein evolution. Science. 1974 Sep 6;185(4154):862-4. PubMed PMID: 4843792.
[28] Nakashima, H., Nishikawa, K. and Ooi, T. (1990), Distinct character in hydrophobicity of amino acid compositions of mitochondrial proteins. Proteins, 8: 173-178. doi:10.1002/prot.340080207
[29] Henry B. Bull, Keith Breese, Surface tension of amino acid solutions: A hydrophobicity scale of the amino acid residues, Archives of Biochemistry and Biophysics, Volume 161, Issue 2, 1974, Pages 665-670, ISSN 0003-9861, https://doi.org/10.1016/0003-9861(74)90352-X.
[30] Marvin Charton, Barbara I. Charton, The structural dependence of amino acid hydrophobicity parameters, Journal of Theoretical Biology, Volume 99, Issue 4, 1982, Pages 629-644, ISSN 0022-5193,https://doi.org/10.1016/0022-5193(82)90191-6.
[31] Eisenberg, D., Wilcox, W. and McLachlan, A.D. (1986), Hydrophobicity and amphiphilicity in protein structure. J. Cell. Biochem., 31: 11-17. doi:10.1002/jcb.240310103
[32] FAUCHÈRE, J.-L., CHARTON, M., KIER, L.B., VERLOOP, A. and PLISKA, V. (1988), Amino acid side chain parameters for correlation studies in biology and pharmacology. International Journal of Peptide and Protein Research, 32: 269-278. doi:10.1111/j.1399-3011.1988.tb01261.x
[33] "Prediction of protein antigenic determinants from amino acid sequences" T P Hopp, K R Woods. Proceedings of the National Academy of Sciences Jun 1981, 78 (6) 3824-3828; DOI: 10.1073/pnas.78.6.3824
[34] JANIN, J. Surface and inside volumes in globular proteins. Nature 277, 491–492 (1979). https://doi.org/10.1038/277491a0
[35] E Q Lawson, A J Sadler, D Harmatz, D T Brandau, R Micanovic, R D MacElroy and C R Middaugh A simple experimental model for hydrophobic interactions in proteins. J. Biol. Chem. 1984, 259:2910-2912.

[36] Ken Nishikawa and Tatsuo Ooi. Predition of the surface interior diagram of globular proteins by an empirical method. lnt. J. Peptide Protein Res. 16,1980, 19-32
[37] Nishikawa K, Ooi T. Radial locations of amino acid residues in a globular protein: correlation with the sequence. J Biochem. 1986 Oct;100(4):1043-7. PubMed PMID: 3818558.
[38] Nozaki Y, Tanford C. The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydrophobicity scale. J Biol Chem. 1971 Apr 10;246(7):2211-7. PubMed PMID: 5555568
[39] Optimization of Amino Acid Parameters for Correspondence of Sequence to Tertiary Structures of Proteins. Oobatake, Motohisa; Kubota, Yasushi; Ooi, Tatsuo. Bulletin of the Institute for Chemical Research, Kyoto University (1985), 63(2): 82-94
[40] Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution Anna Radzicka and Richard Wolfenden. Biochemistry 1988 27 (5), 1664-1670 DOI: 10.1021/bi00405a042
[41] Simon, 1976 (Could not find the article)
[42] V. Veljkovic, I. Cosic, Dimitrijevic and D. Lalovic, "Is it Possible to Analyze DNA and Protein Sequences by the Methods of Digital Signal Processing?," in IEEE Transactions on Biomedical Engineering, vol. BME-32, no. 5, pp. 337-341, May 1985.doi: 10.1109/TBME.1985.325549
[43] Zaslavsky etal., 1982,
[44] B.Yu. Zaslavsky, L.M. Miheeva, N.M. Mestechkina, S.V. Rogozhin,
[45] Physico-chemical factors governing partition behaviour of solutes and particles in aqueous polymeric biphasic systems.: II. Effect of ionic composition on the hydration properties of the phases, Journal of Chromatography A, Volume 253, 1982, Pages 149-158, ISSN 0021-9673, https://doi.org/10.1016/S0021-9673(01)88374-6.
[46] Zimmerman JM, Eliezer N, Simha R. The characterization of amino acid sequences in proteins by statistical methods. J Theor Biol. 1968 Nov;21(2):170-201. PubMed PMID: 5700434.Mitaku et al., 2002
[47] I. Cosic, "Macromolecular bioactivity: is it resonant interaction between macromolecules?-theory and applications," in IEEE Transactions on Biomedical Engineering, vol. 41, no. 12, pp. 1101-1114, Dec. 1994.
[48] Guy H. R. (1985). Amino acid side-chain partition energies and distribution of residues in soluble proteins. Biophysical journal, 47(1), 61–70. https://doi.org/10.1016/S0006-3495(85)83877-7
[49] Miyazawa, S. and Jernigan, R.L. (1999), Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. Proteins, 34: 49-68. doi:10.1002/(SICI)1097-0134(19990101)34:1
[50] Hilda Cid, Marta Bunster, Mauricio Canales, Felipe Gazitúa, Hydrophobicity and structural classes in proteins, Protein Engineering, Design and Selection, Volume 5, Issue 5, July 1992, Pages 373–375, https://doi.org/10.1093/protein/5.5.373
[51] Ann. Rev. Biochem. 1984. 53: 595-623 Three dimensional structure of membrane and surface proteins. David Eisenberg
[52] Marvin Charton, Barbara I. Charton, The dependence of the Chou-Fasman parameters on amino acid side chain structure, Journal of Theoretical Biology, Volume 102, Issue 1, 1983, Pages 121-134, ISSN 0022-5193, https://doi.org/10.1016/0022-5193(83)90265-5.
[53] Fauchere, J. & Pliska, Vladimir. (1983). Hydrophobic parameters II of amino acid side-chains from the partitioning of N-acetyl-amino acid amides. Eur. J. Med. Chem.. 18.

[54] Goldsack DE, Chalifoux RC. Contribution of the free energy of mixing of hydrophobic side chains to the stability of the tertiary structure of proteins. J Theor Biol. 1973 Jun;39(3):645-51. PubMed PMID: 4354159.
[55] Jones DD. Amino acid properties and side-chain orientation in proteins: a cross correlation appraoch. J Theor Biol. 1975 Mar;50(1):167-83. PubMed PMID: 1127956.
[56] Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 1982 May 5;157(1):105-32. PubMed PMID: 7108955.
[57] Levitt, 1976 (Article not found)
[58] Meek JL. Prediction of peptide retention times in high-pressure liquid chromatography on the basis of amino acid composition. Proc Natl Acad Sci U S A. 1980 Mar;77(3):1632-6. PubMed PMID: 6929513; PubMed Central PMCID: PMC348551.
[59] James L. Meek, Zvani L. Rossetti, Factors affecting retention and resolution of peptides in high-performance liquid chromatography, Journal of Chromatography A, Volume 211, Issue 1, 1981, Pages 15-28, ISSN 0021-9673, https://doi.org/10.1016/S0021-9673(00)81169-3.
[60] Parker JM, Guo D, Hodges RS. New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. Biochemistry. 1986 Sep 23;25(19):5425-32. PubMed PMID: 2430611.
[61] M Prabhakaran; The distribution of physical, chemical and conformational properties in signal and nascent peptides. Biochem J 1 August 1990; 269 (3): 691–696. doi: https://doi.org/10.1042/bj2690691
[62] Roseman MA. Hydrophilicity of polar amino acid side-chains is markedly reduced by flanking peptide bonds. J Mol Biol. 1988 Apr 5;200(3):513-22. PubMed PMID: 3398047.
[63] Sweet RM, Eisenberg D. Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. J Mol Biol. 1983 Dec 25;171(4):479-88. PubMed PMID: 6663622.
[64] Ponnuswamy PK. Hydrophobic characteristics of folded proteins. Prog Biophys Mol Biol. 1993;59(1):57-103. Review. PubMed PMID: 8419986.
[65] Physicochemical Basis of Amino Acid Hydrophobicity Scales: Evaluation of Four New Scales of Amino Acid Hydrophobicity Coefficients Derived from RP-HPLC of Peptides Mathew C. J. Wilce, Marie-Isabel. Aguilar, and Milton T. W. Hearn Analytical Chemistry 1995 67 (7), 1210-1219 DOI: 10.1021/ac00103a012
[66] Dacheng Guo, Colin T. Mant, Ashok K. Taneja, J.M.Robert Parker, Robert S. Rodges, Prediction of peptide retention times in reversed-phase high-performance liquid chromatography I. Determination of retention coefficients of amino acid residues of model synthetic peptides, Journal of Chromatography A, Volume 359, 1986, Pages 499-518, ISSN 0021-9673, https://doi.org/10.1016/0021-9673(86)80102-9.
[67] Davor Juretić, Bono Lučić, Damir Zucić, Nenad Trinajstić, Protein transmembrane structure: recognition and prediction by using hydrophobicity scales through preference functions, Theoretical and Computational Chemistry, Elsevier, Volume 5, 1998, Pages 405-445, ISSN 1380-7323, ISBN 9780444826602, https://doi.org/10.1016/S1380-7323(98)80015-0.
[68] Kidera, A., Konishi, Y., Oka, M. et al. Statistical analysis of the physical properties of the 20 naturally occurring amino acids. J Protein Chem 4, 23–55 (1985). https://doi.org/10.1007/BF01025492

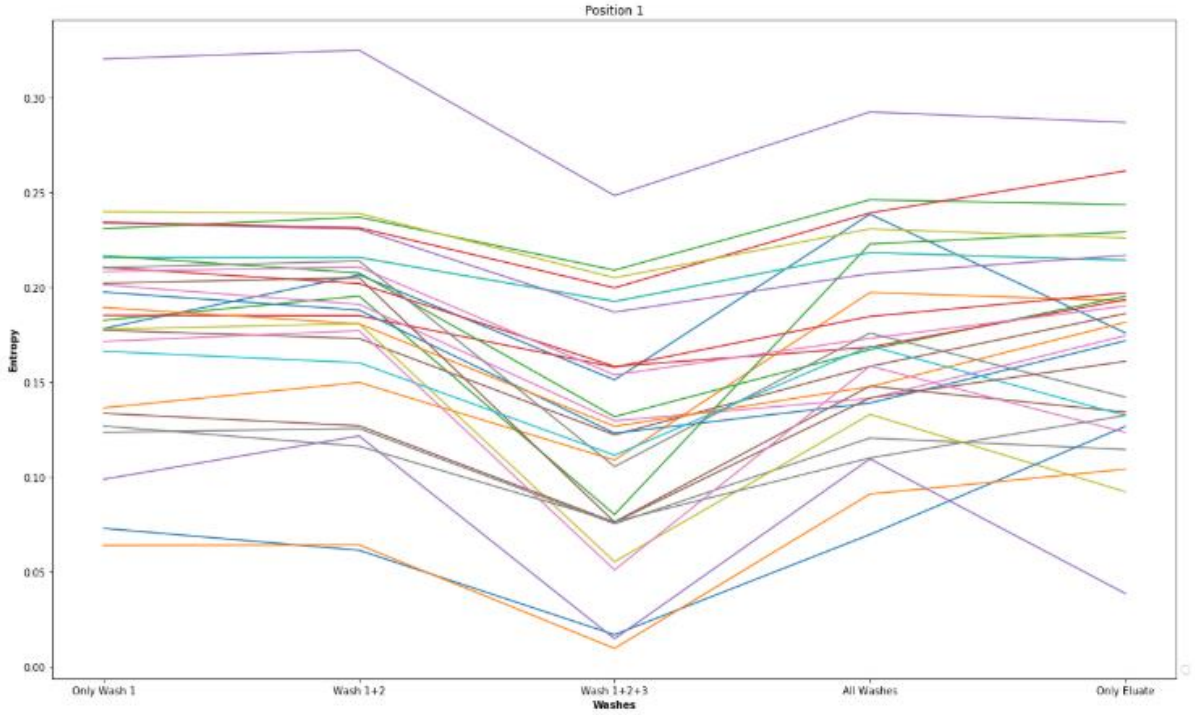
[69] Jacobs RE, White SH. The nature of the hydrophobic binding of small peptides at the bilayer interface: implications for the insertion of transbilayer helices. Biochemistry. 1989 Apr 18;28(8):3421-37. PubMed PMID: 2742845.
[70] Casari G, Sippl MJ. Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds. J Mol Biol. 1992 Apr 5;224(3):725-32. Review. PubMed PMID: 1569551.
[71] Cornette JL, Cease KB, Margalit H, Spouge JL, Berzofsky JA, DeLisi C. Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. J Mol Biol. 1987 Jun 5;195(3):659-85. PubMed PMID: 3656427.
[72] Protein conformational prediction Fasman, Gerald D. Trends in Biochemical Sciences, 1989 Volume 14, Issue 7, 295 - 299
[73] Barry Robson, David J. Osguthorpe, Refined models for computer simulation of protein folding: Applications to the study of conserved secondary structure and flexible hinge points during the folding of pancreatic trypsin inhibitor, Journal of Molecular Biology, Volume 132, Issue 1, 1979, Pages 19-51, ISSN 0022-2836, https://doi.org/10.1016/0022-2836(79)90494-7.
[74] Venanzi TJ. Hydrophobicity parameters and the bitter taste of L-amino acids. J Theor Biol. 1984 Dec 7;111(3):447-50. PubMed PMID: 6521488.
[75] Affinities of amino acid side chains for solvent water R. Wolfenden, L. Andersson, P. M. Cullis, and C. C. B. Southgate Biochemistry 1981 20 (4), 849-855 DOI: 10.1021/bi00507a030
[76] Miyazawa, S. and Jernigan, R.L. (1999), Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. Proteins, 34: 49-68. doi:10.1002/(SICI)1097-0134(19990101)34:1<49::AID-PROT5>3.0.CO;2-L
[77] Samuel Kakraba, Debra Knisley (2016). A graph-theoretic model of single point mutations in the cystic fibrosis transmembrane conductance regulator. Journal of Advances in biotechnology. ISSN 2348 - 6201 6:1
[78] C.A. Browne, H.P.J. Bennett, S. Solomon, The isolation of peptides by high-performance liquid chromatography using predicted elution positions, Analytical Biochemistry, Volume 124, Issue 1, 1982, Pages 201-208, ISSN 0003-2697, https://doi.org/10.1016/0003-2697(82)90238-X.
[79] Itoh, K., Foxman, B.M. and Fasman, G.D. (1976), The two β forms of poly (L-glutamic acid). Biopolymers, 15: 419-455. doi:10.1002/bip.1976.360150302
[80] Joël Janin, Shoshanna Wodak, Michael Levitt, Bernard Maigret, Conformation of amino acid side-chains in proteins, Journal of Molecular Biology, Volume 125, Issue 3, 1978, Pages 357-386, ISSN 0022-2836, https://doi.org/10.1016/0022-2836(78)90408-4.
[81] Finkelstein, A.V., Badretdinov, A.Y. and Ptitsyn, O.B. (1991), Physical reasons for secondary structure stability: α-Helices in short peptides. Proteins, 10: 287-299. doi:10.1002/prot.340100403
[82] Qian, N., & Sejnowski, T. J. (1988). Predicting the secondary structure of globular proteins using neural network models. Journal of Molecular Biology, 202(4), 865–884. doi:10.1016/0022-2836(88)90564-5
[83] B. Robson, E. Suzuki, Conformational properties of amino acid residues in globular proteins, Journal of Molecular Biology, 107, 3, 1976, 327-356, doi:10.1016/S0022-2836(76)80008
[84] Blaber M, Zhang XJ, Matthews BW. Structural basis of amino acid alpha helix propensity. Science. 1993 Jun 11;260(5114):1637-40. PubMed PMID: 8503008.

[85] Structure-based conformational preferences of amino acids Patrice Koehl, Michael Levitt Proceedings of the National Academy of Sciences Oct 1999, 96 (22) 12524-12529; DOI: 10.1073/pnas.96.22.12524Oobatake et al.,1985,
[86] Ptitsyn, O.B. and Finkelstein, A.V. (1983), Theory of protein secondary structure and algorithm of its prediction. Biopolymers, 22: 15-25. doi:10.1002/bip.360220105
[87] Nakashima H, Nishikawa K. The amino acid composition is different between the cytoplasmic and extracellular sides in membrane proteins. FEBS Lett. 1992 Jun 1;303(2-3):141-6. PubMed PMID: 1607012.Biou et al.,1988,
[88] Cyrus Chothia, The nature of the accessible and buried surfaces in proteins, Journal of Molecular Biology, Volume 105, Issue 1, 1976, Pages 1-12, ISSN 0022-2836, https://doi.org/10.1016/0022-2836(76)90191-1.
[89] Hydrophobicity, hydrophilicity, and the radial and orientational distributions of residues in native proteins S Rackovsky, H A Scheraga Proceedings of the National Academy of Sciences Dec 1977, 74 (12) 5248-5251; DOI: 10.1073/pnas.74.12.5248
[90] Amino acid preferences for specific locations at the ends of alpha helices JS Richardson, DC Richardson Science17 Jun 1988 : 1648-1652
[91] Computed conformational states of the 20 naturally occurring amino acid residues and of the prototype residue α-aminobutyric acid Max Vasquez, George Nemethy, and Harold A. Scheraga Macromolecules 1983 16 (7), 1043-1049 DOI: 10.1021/ma00241a004
[92] Punta, M. and Maritan, A. (2003), A knowledge-based scale for amino acid membrane propensity. Proteins, 50: 114-121. doi:10.1002/prot.10247
[93] Status of empirical methods for the prediction of protein backbone topography Frederick R. Maxfield and Harold A. Scheraga Biochemistry 1976 15 (23), 5138-5153 DOI: 10.1021/bi00668a030
[94] Hypothesis about the Mechanism of Protein Folding Seiji Tanaka and Harold A. Scheraga Macromolecules 1977 10 (2), 291-304 DOI: 10.1021/ma60056a015
[95] Monné M, Hermansson M, von Heijne G. A turn propensity scale for transmembrane helices. J Mol Biol. 1999 Apr 23;288(1):141-5. PubMed PMID: 10329132.
[96] Bastolla U, Porto M, Roman HE, Vendruscolo M. Principal eigenvector of contact matrices and hydrophobicity profiles in proteins. Proteins. 2005 Jan 1;58(1):22-30. PubMed PMID: 15523667.
[97] Isogai, Y., Némethy, G., Rackovsky, S., Leach, S.J. and Scheraga, H.A. (1980), Characterization of multiple bends in proteins. Biopolymers, 19: 1183-1210. doi:10.1002/bip.1980.360190607
[98] Mikita Suyama, Osamu Ohara, DomCut: prediction of inter-domain linker regions in amino acid sequences, Bioinformatics, Volume 19, Issue 5, 22 March 2003, Pages 673–674, https://doi.org/10.1093/bioinformatics/btg031
[99] Kyounghwa Bae, Bani K. Mallick, Christine G. Elsik, Prediction of protein interdomain linker regions by a hidden Markov model, Bioinformatics, Volume 21, Issue 10, 2005, Pages 2264–2270, https://doi.org/10.1093/bioinformatics/bti363
[100] Deep Directed Evolution of Solid Binding Peptides for Quantitative Big-data Generation. Deniz T. Yucesoy, Siddharth S. Rath, Jacob L. Rodriguez, Jonathan Francis-Landau, Oliver Nakano-Baker, Mehmet Sarikaya. bioRxiv 2021.01.26.428348; doi: https://doi.org/10.1101/2021.01.26.428348
[101] http://www.stat.yale.edu/Courses/1997-98/101/linmult.htm.
[102] A Generalized Similarity Metric for Predicting Peptide Binding Affinity. Jacob Rodriguez, Siddharth Rath, Jonathan Francis-Landau, Yekta Demirci, Burak Berk Ustundag, Mehmet Sarikaya. bioRxiv 654913; doi: https://doi.org/10.1101/654913

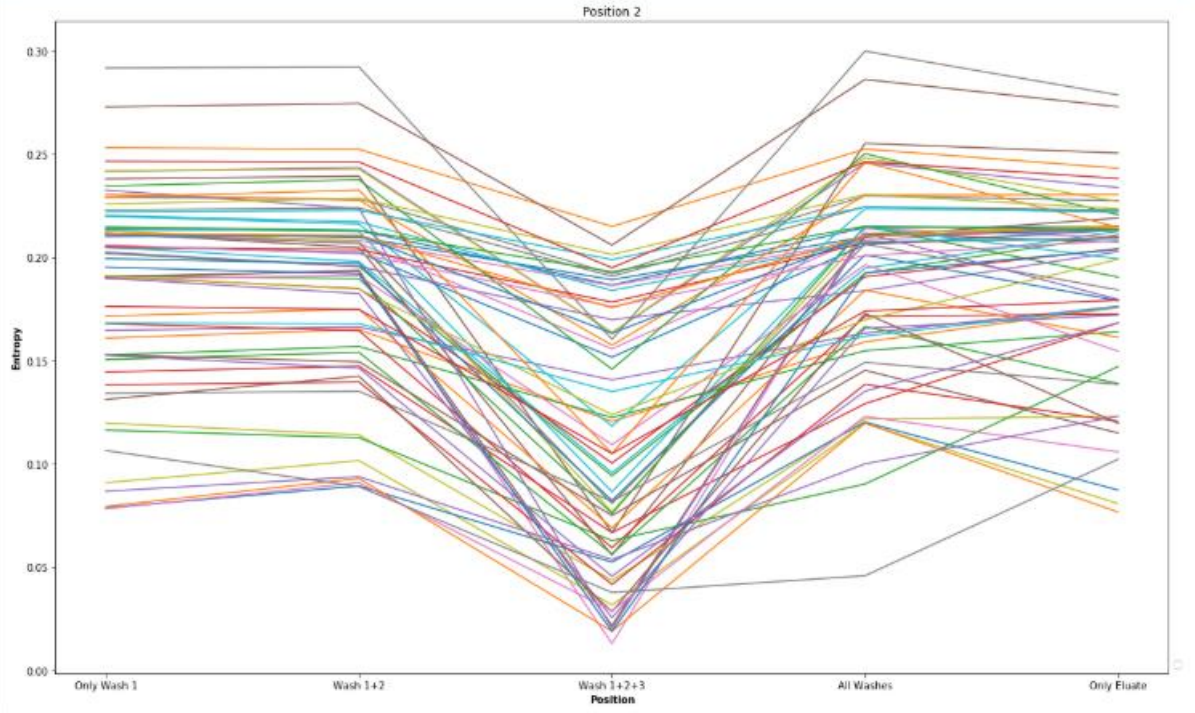
[103] de Oliveira, S., & Deane, C. (2017). Co-evolution techniques are reshaping the way we do structural bioinformatics. F1000Research, 6, 1224. https://doi.org/10.12688/f1000research.11543.1
[104] Emerging Device Applications for Semiconducting Two-Dimensional Transition Metal Dichalcogenides. Deep Jariwala, Vinod K. Sangwan, Lincoln J. Lauhon, Tobin J. Marks, and Mark C. Hersam. ACS Nano 2014 8 (2), 1102-1120 DOI: 10.1021/nn500064s
[105] Edwards-Gayle, C.J.C.; Hamley, I.W. Self-assembly of bioactive peptides, peptide conjugates, and peptide mimetic materials. Org. Biomol. Chem. 2017, 15, 5867–5876.
[106] Whitesides, G.M.; Boncheva, M. Beyond molecules: Self-assembly of mesoscopic and macroscopic components. Proc. Natl. Acad. Sci. USA 2002, 99, 4769–4774.
[107] Lee, S., Trinh, T., Yoo, M., Shin, J., Lee, H., Kim, J., Hwang, E., Lim, Y. B., & Ryou, C. (2019). Self-Assembling Peptides and Their Application in the Treatment of Diseases. International journal of molecular sciences, 20(23), 5850. https://doi.org/10.3390/ijms20235850
[108] Kowalewski, T. & Holtzman, D. M. In situ atomic force microscopy study of Alzheimer's β -amyloid peptide on different substrates: New insights into mechanism of β -sheet formation. Proceedings of the National Academy of Sciences 96, 3688–3693 (1999).
[109] Zhang, F. et al. Epitaxial growth of peptide nanofilaments on inorganic surfaces: Effects of interfacial hydrophobicity/hydrophilicity. Angewandte Chemie International Edition 45, 3611–3613 (2006).
[110] So, C. R., Tamerler, C. & Sarikaya, M. Adsorption, Diffusion, and Self-Assembly of an Engineered Gold-Binding Peptide on Au (111) Investigated by Atomic Force Microscopy. Angewandte Chemie International Edition 48, 5174–5177 (2009).
[111] Hayamizu, Y., So, C., Dag, S. et al. Bioelectronic interfaces by spontaneously organized peptides on 2D atomic single layer materials. Sci Rep 6, 33778 (2016). https://doi.org/10.1038/srep33778
[112] Yu GL, Jalil R, Belle B, Mayorov AS, Blake P, Schedin F, et al. Interaction phenomena in graphene seen through quantum capacitance. PNAS 2013;110:3282-6.
[113] Chhowalla M, Shin HS, Eda G, Li LJ, Loh KP, Zhang H. The chemistry of two-dimensional layered transition metal dichalcogenide nanosheets. Nat Chem 2014;5:263-75.
[114] Arun Kumar Singh, P. Kumar, D.J. Late, Ashok Kumar, S. Patel, Jai Singh. 2D layered transition metal dichalcogenides (MoS2): Synthesis, applications and theoretical aspects, App Mat Today,2018;13:242-270,

Appendix A: Graphs of properties filtered for correlation to binding at each peptide position.

```
[2, 9, 11, 29, 30, 36, 42, 54, 60, 67, 69, 70, 83, 103, 104, 114, 135, 145, 146, 252, 260, 263, 273, 281, 284, 344, 348, 351, 5
01]
```

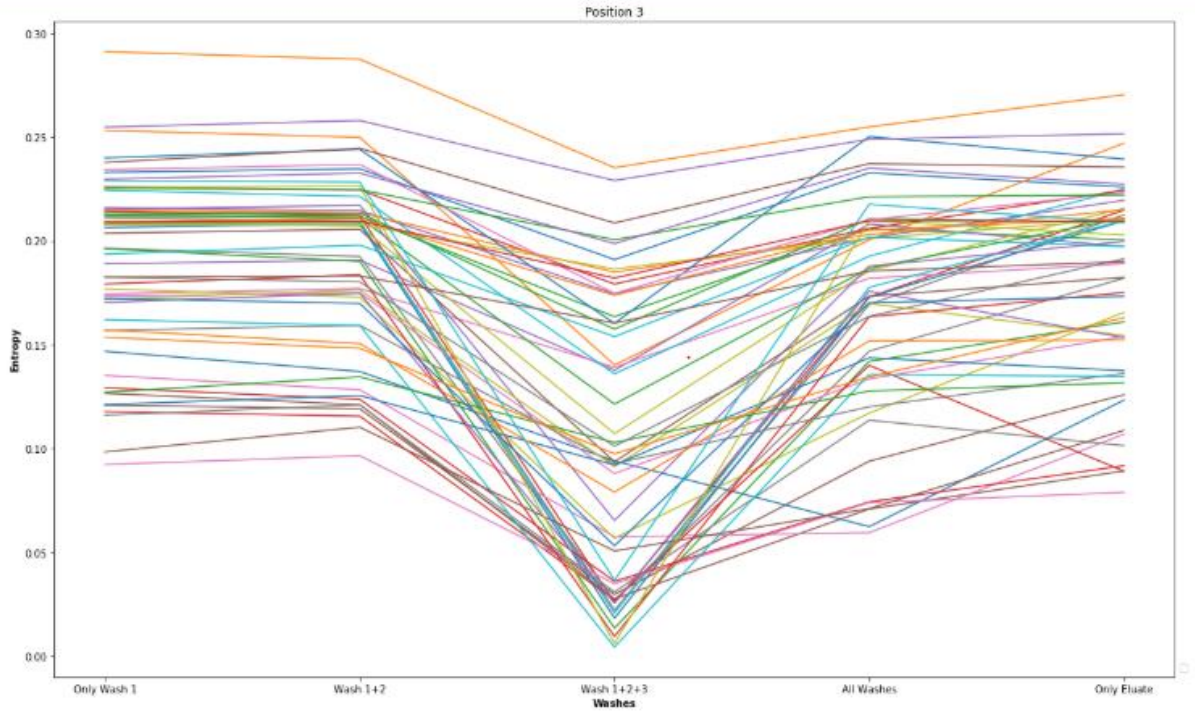



```
[6, 9, 10, 11, 28, 30, 34, 35, 37, 42, 57, 63, 69, 70, 76, 83, 89, 97, 103, 104, 107, 109, 120, 121, 132, 135, 137, 140, 143, 1
45, 146, 147, 148, 157, 196, 252, 262, 263, 264, 266, 267, 268, 272, 274, 275, 285, 288, 291, 313, 344, 345, 348, 349, 351, 35
2, 357, 380, 428, 439, 449, 454, 463, 489, 501, 515]
No handles with labels found to put in legend.
```

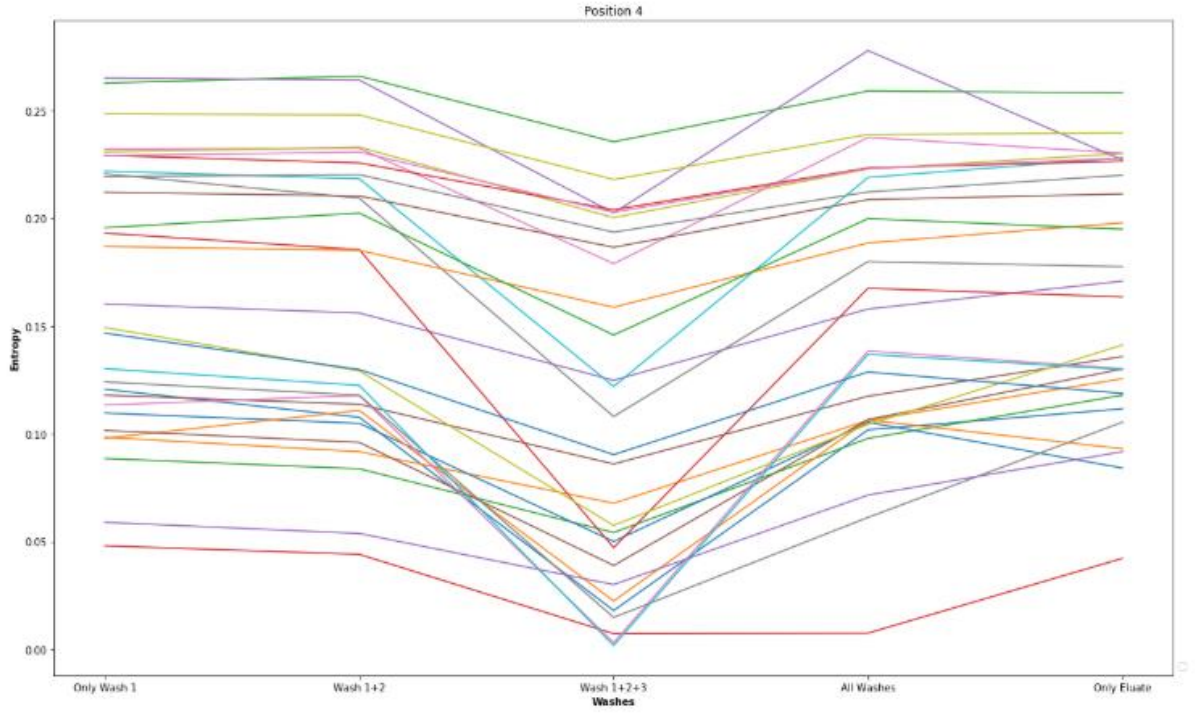


```
[28, 30, 37, 42, 44, 54, 66, 69, 70, 83, 89, 97, 103, 104, 107, 112, 114, 120, 121, 135, 136, 137, 140, 142, 145, 146, 147, 14
8, 157, 188, 198, 199, 238, 252, 255, 256, 257, 258, 259, 262, 263, 275, 280, 282, 283, 288, 343, 345, 348, 349, 351, 352, 374,
400, 439, 464]
```

```
No handles with labels found to put in legend.
```

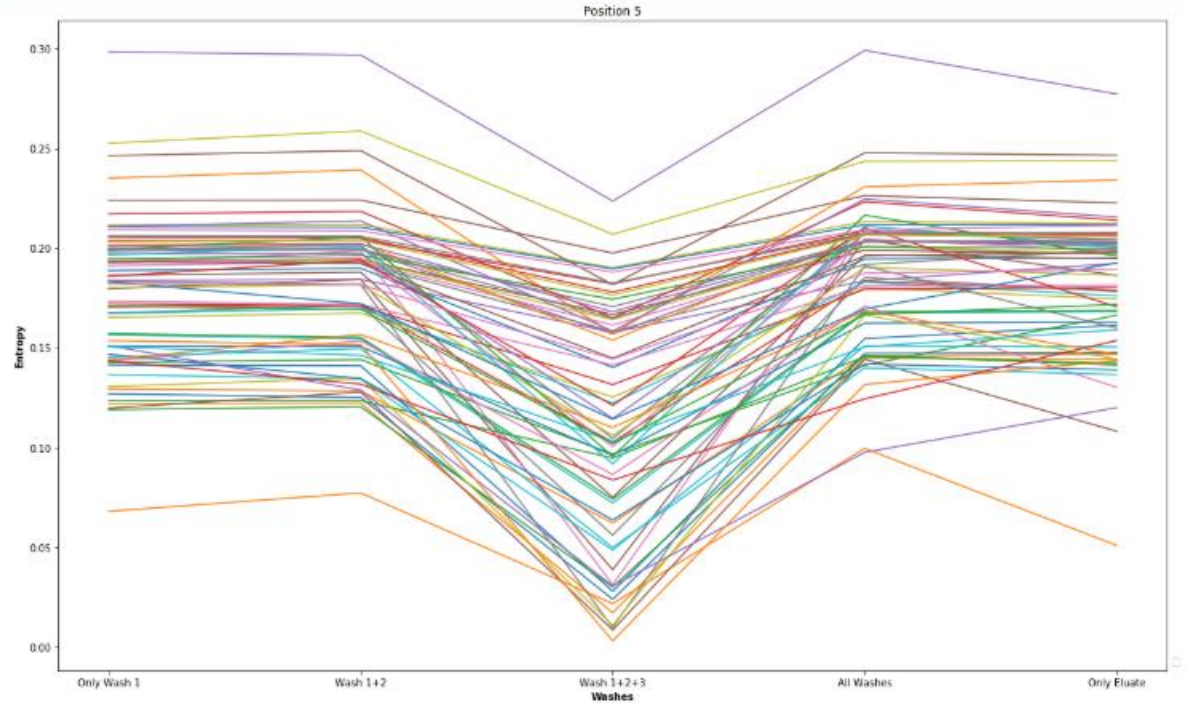



```
[3, 72, 74, 77, 128, 149, 178, 196, 198, 251, 252, 253, 255, 262, 265, 272, 273, 274, 275, 286, 289, 313, 315, 370, 379, 384, 4
28, 454, 501]
```

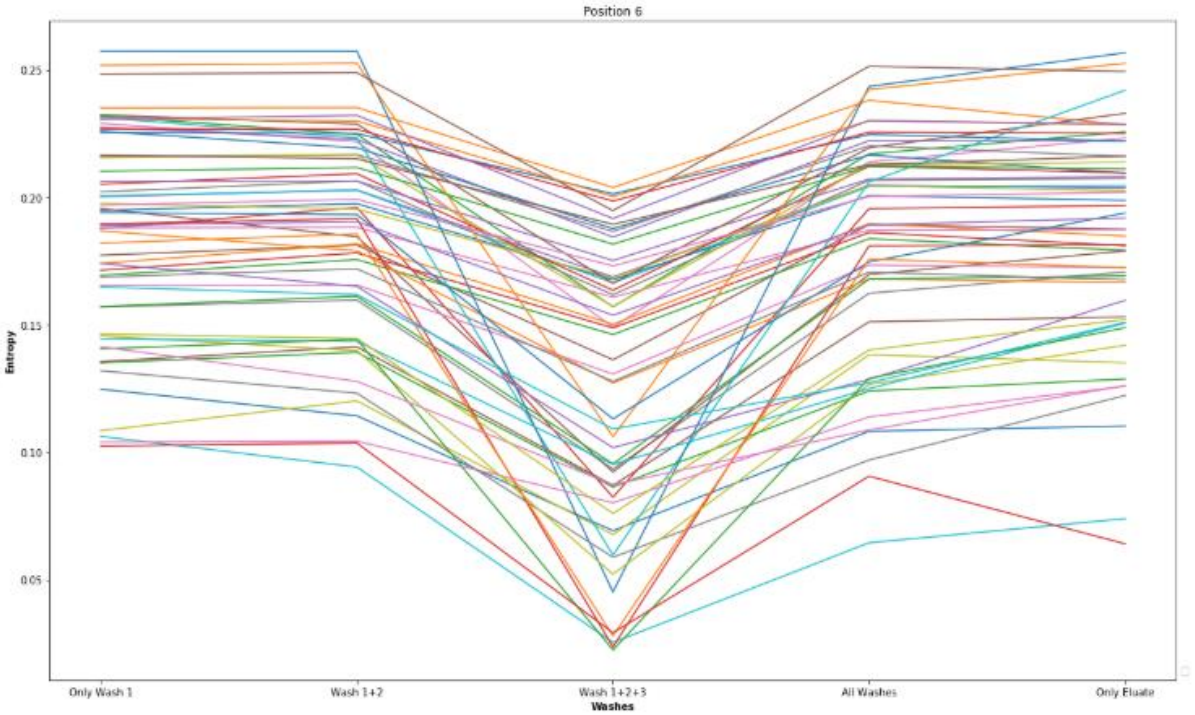


```
[9, 10, 11, 25, 29, 44, 48, 59, 65, 71, 72, 73, 74, 84, 85, 86, 91, 92, 93, 94, 105, 111, 124, 128, 129, 130, 131, 132, 134, 14
9, 155, 156, 168, 173, 174, 180, 182, 183, 186, 188, 189, 198, 222, 238, 250, 257, 258, 262, 263, 267, 275, 280, 282, 283, 289,
318, 334, 341, 343, 344, 345, 350, 351, 352, 407, 456, 457, 458, 470]
```

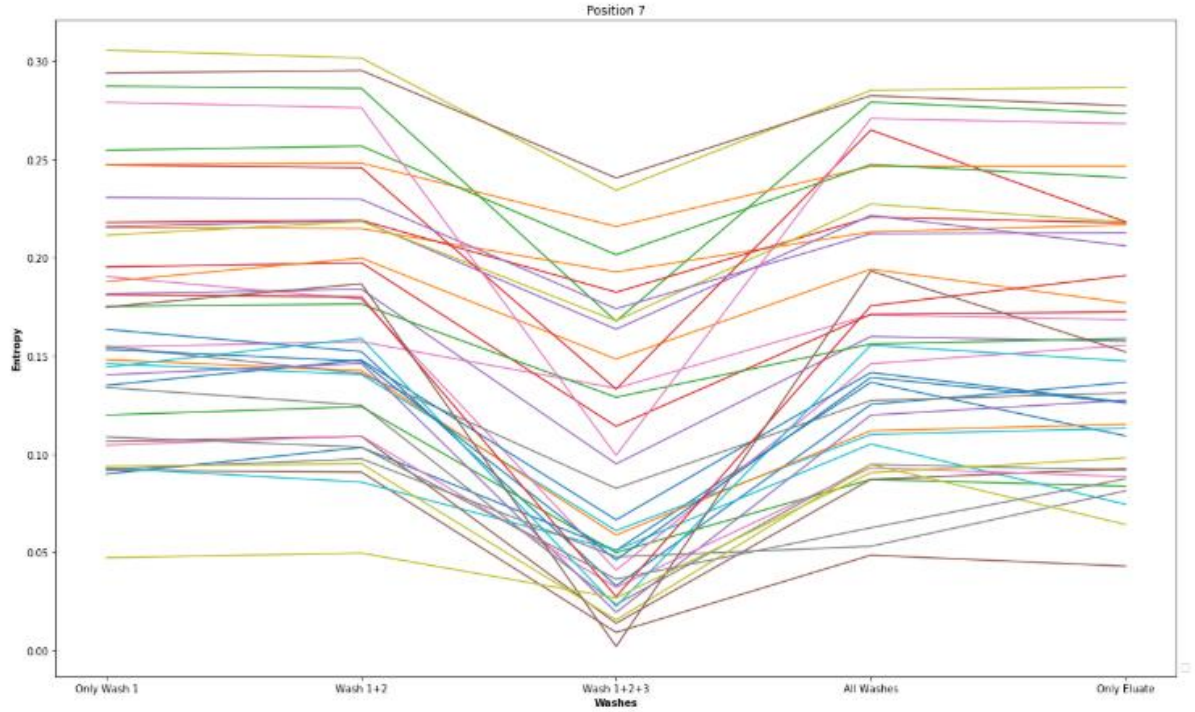
```
No handles with labels found to put in legend.
```



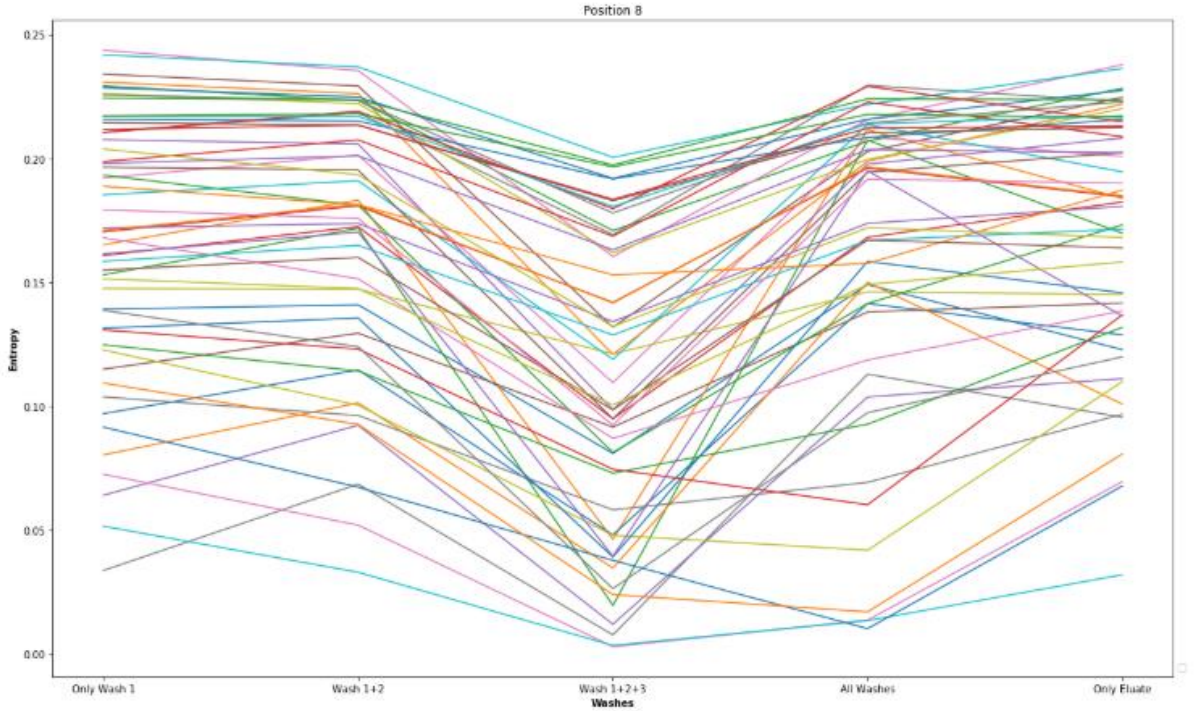

```
[3, 28, 42, 44, 59, 66, 69, 76, 78, 85, 86, 92, 93, 94, 103, 121, 126, 129, 132, 143, 145, 147, 155, 156, 173, 178, 197, 199, 2
38, 251, 255, 256, 258, 259, 262, 263, 264, 267, 275, 280, 282, 283, 288, 290, 315, 318, 350, 351, 352, 354, 374, 399, 400, 42
8, 439, 454, 456]
```



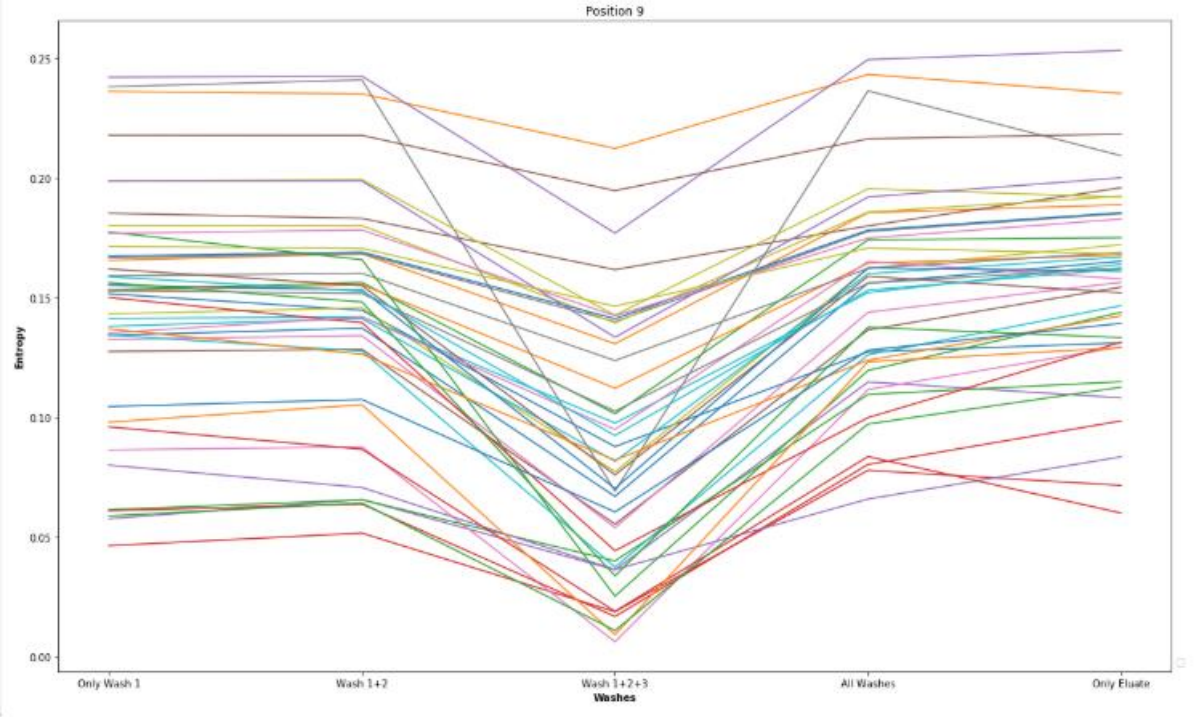
[2, 6, 30, 33, 37, 46, 76, 80, 97, 112, 114, 142, 143, 147, 148, 174, 196, 200, 250, 252, 258, 259, 262, 265, 267, 268, 269, 27
4, 284, 288, 289, 318, 343, 345, 348, 354, 357, 411, 525]



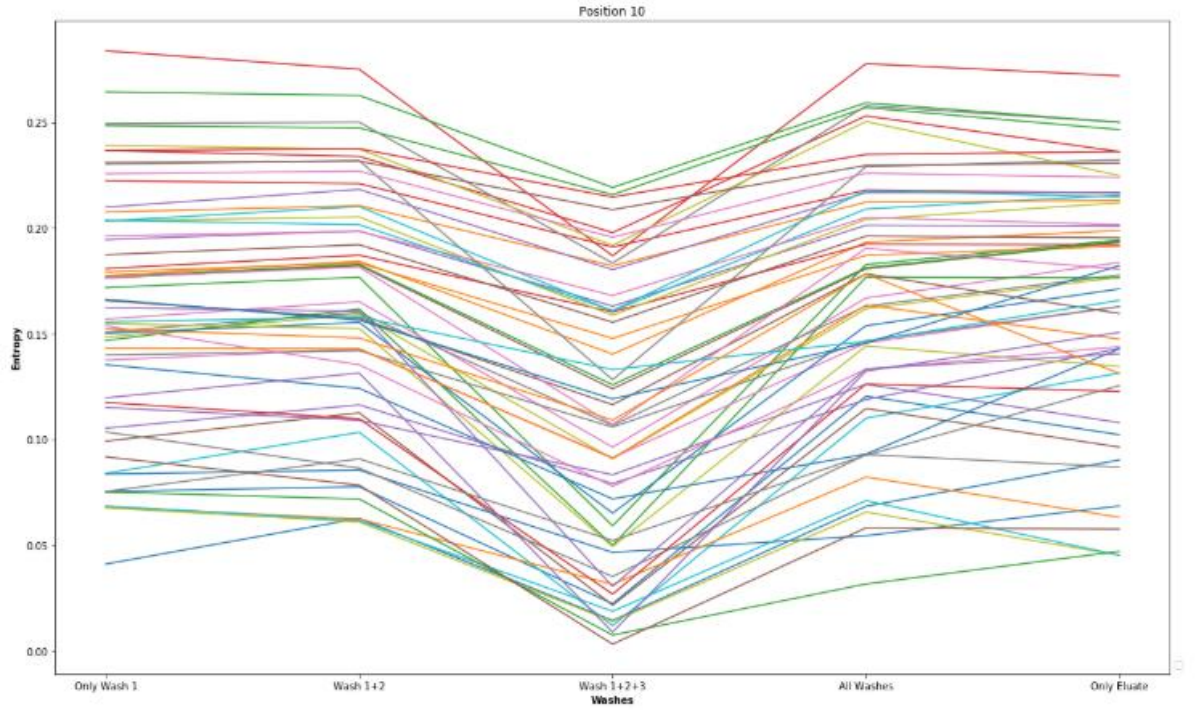
[3, 9, 10, 11, 36, 42, 44, 47, 54, 55, 58, 69, 70, 71, 77, 78, 83, 91, 103, 104, 107, 110, 121, 130, 131, 132, 136, 137, 140, 1
55, 157, 168, 180, 183, 188, 238, 254, 257, 267, 271, 272, 282, 286, 288, 289, 344, 345, 350, 352, 380, 411, 412, 446, 463]



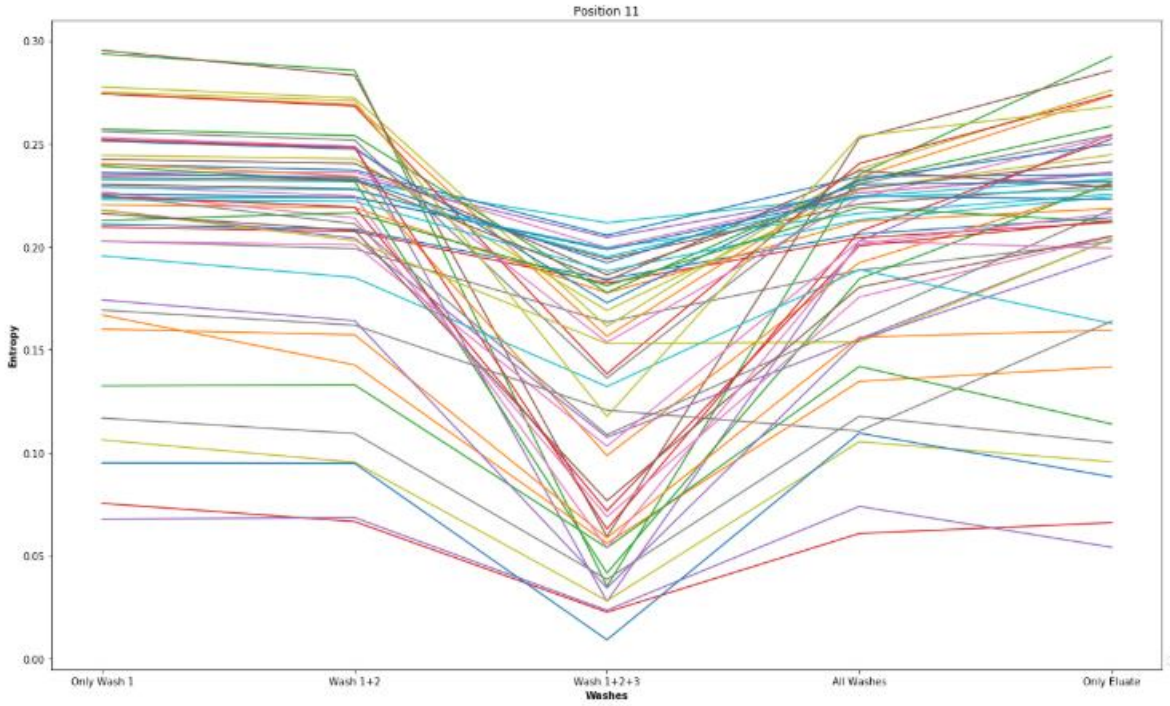
[3, 9, 10, 12, 41, 55, 58, 71, 72, 73, 74, 76, 77, 110, 111, 124, 125, 126, 128, 149, 168, 178, 196, 200, 251, 262, 263, 265, 266, 272, 273, 275, 289, 290, 313, 314, 315, 344, 351, 352, 379, 515, 525]



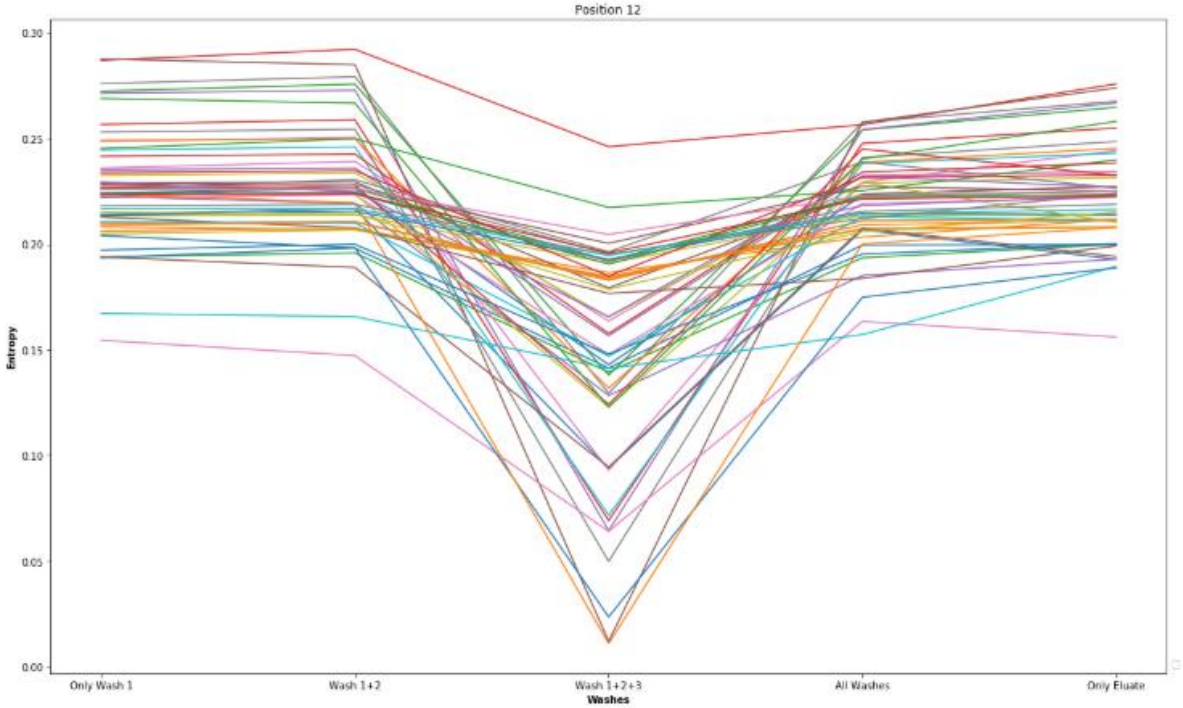
[2, 9, 11, 28, 58, 71, 72, 73, 74, 77, 79, 84, 90, 92, 105, 111, 124, 125, 129, 130, 149, 156, 168, 174, 182, 186, 188, 198, 200, 250, 252, 254, 255, 256, 258, 259, 262, 267, 268, 269, 275, 280, 287, 288, 289, 290, 315, 318, 343, 344, 345, 354, 357, 370, 379, 392, 439, 525]



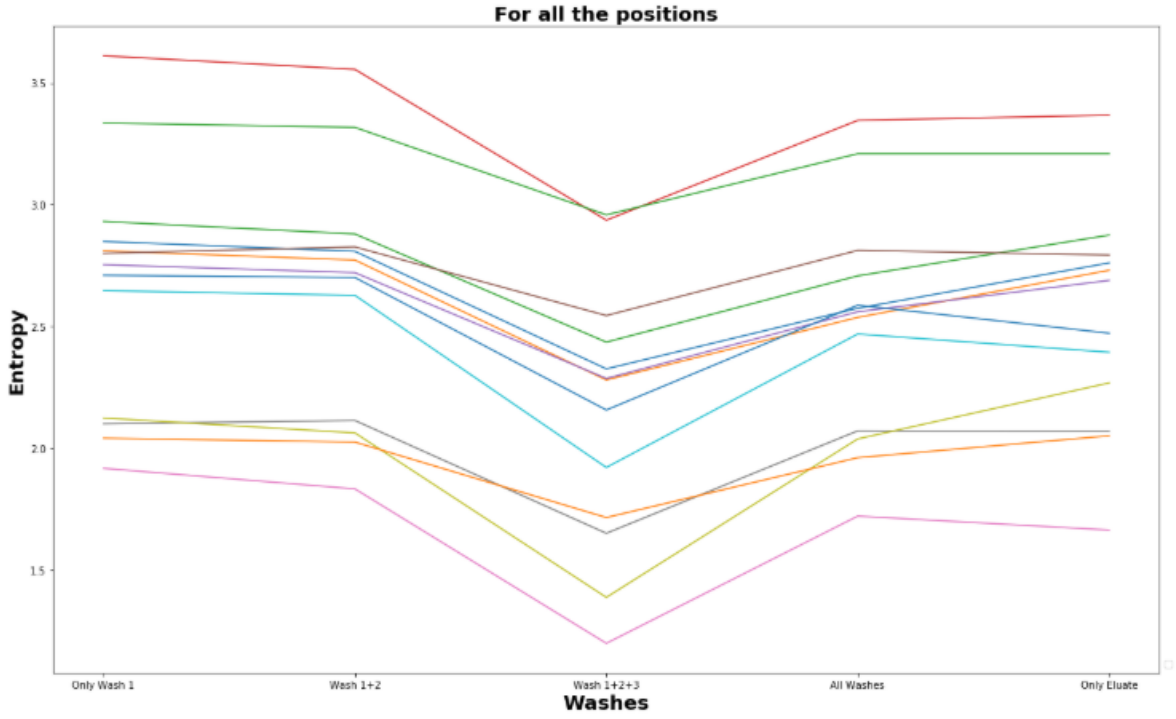
[6, 28, 30, 33, 34, 35, 36, 37, 42, 57, 63, 69, 70, 83, 89, 97, 103, 104, 107, 108, 109, 120, 121, 135, 136, 137, 140, 145, 14
6, 157, 178, 196, 198, 252, 253, 255, 256, 259, 262, 265, 280, 282, 283, 343, 345, 348, 349, 350, 357, 442, 501]



[3, 6, 29, 30, 34, 35, 36, 42, 57, 60, 66, 67, 69, 70, 83, 97, 103, 104, 106, 107, 108, 109, 135, 137, 140, 145, 157, 178, 181,
198, 199, 207, 218, 251, 255, 256, 282, 313, 343, 345, 348, 349, 357, 380, 397, 442, 447, 448, 454, 455, 464, 465, 472, 491]



```
[42, 69, 83, 97, 103, 198, 262, 282, 345, 348, 349, 351, 357]
No handles with labels found to put in legend.
```



Appendix B: Tables of properties filtered for correlation to binding at each peptide position.

Table B1 All the quantitative amino acid residue structural properties (Dataset available on Github)

id	Prop	authors
0	alpha-CH chemical shifts	Andersen et al., 1992
2	Signal sequence helical potential	Argos et al., 1982
3	Membrane-buried preference parameters	Argos et al., 1982
4	Conformational parameter of inner helix	Beghin-Dirkx, 1975
5	Conformational parameter of beta-structure	Beghin-Dirkx, 1975
6	Conformational parameter of beta-turn	Beghin-Dirkx, 1975
7	Average flexibility indices	Bhaskaran-Ponnuswamy, 1988
9	Information value for accessibility; average fraction 35%	Biou et al.,1988
10	Information value for accessibility; average fraction 23%	Biou et al.,1988
11	Retention coefficient in TFA	Browne et al., 1982
12	Retention coefficient in HFBA	Browne et al., 1982
13	Transfer free energy to surface	Bull-Breese, 1974
14	Apparent partial specific volume	Bull-Breese, 1974
15	alpha-NH chemical shifts	Bundi-Wuthrich, 1979
17	Spin-spin coupling constants 3JHalpha-NH	Bundi-Wuthrich, 1979
20	Steric parameter	Charton, 1981
21	Polarizability parameter	Charton-Charton, 1982
22	Free energy of solution in water, kcal/mole	Charton-Charton, 1982
23	The Chou-Fasman parameter of the coil conformation	Charton-Charton,1983
24	A parameter defined from the residuals obtained from the best correlation of the Chou-Fasman parameter of beta-sheet	Charton-Charton, 1983
25	The number of atoms in the side chain labelled 1+1	Charton-Charton,1983
26	The number of atoms in the side chain labelled 2+1	Charton-Charton,1983
27	The number of atoms in the side chain labelled 3+1	Charton-Charton,1983
28	The number of bonds in the longest chain	Charton-Charton, 1983
29	A parameter of charge transfer capability	Charton-Charton, 1983

30	A parameter of charge transfer donor capability	Charton-Charton, 1983
31	Average volume of buried residue	Chothia, 1975
32	Residue accessible surface area in tripeptide	Chothia, 1976
33	Residue accessible surface area in folded protein	Chothia, 1976
34	Proportion of residues 95% buried	Chothia, 1976
35	Proportion of residues 100% buried	Chothia, 1976
37	Normalized frequency of alpha-helix	Chou-Fasman, 1978b
38	Normalized frequency of beta-sheet	Chou-Fasman, 1978b
39	Normalized frequency of beta-turn	Chou-Fasman, 1978b
40	Normalized frequency of N-terminal helix	Chou-Fasman, 1978b
41	Normalized frequency of C-terminal helix	Chou-Fasman, 1978b
42	Normalized frequency of N-terminal non helical region	Chou-Fasman,1978b
43	Normalized frequency of C-terminal non helical region	Chou-Fasman,1978b
44	Normalized frequency of N-terminal beta-sheet	Chou-Fasman, 1978b
45	Normalized frequency of C-terminal beta-sheet	Chou-Fasman, 1978b
46	Normalized frequency of N-terminal non beta region	Chou-Fasman, 1978b
47	Normalized frequency of C-terminal non beta region	Chou-Fasman, 1978b
48	Frequency of the 1st residue in turn	Chou-Fasman, 1978b
49	Frequency of the 2nd residue in turn	Chou-Fasman, 1978b
50	Frequency of the 3rd residue in turn	Chou-Fasman, 1978b
51	Frequency of the 4th residue in turn	Chou-Fasman, 1978b
52	Normalized frequency of the 2nd and 3rd residues in turn	Chou-Fasman,1978b
53	Normalized hydrophobicity scales for alpha-proteins	Cid et al., 1992
54	Normalized hydrophobicity scales for beta-proteins	Cid et al., 1992
55	Normalized hydrophobicity scales for alpha+beta-proteins	Cid et al., 1992
56	Normalized hydrophobicity scales for alpha/beta-proteins	Cid et al., 1992
57	Normalized average hydrophobicity scales	Cid et al., 1992
58	Partial specific volume	Cohn-Edsall, 1943
59	Normalized frequency of middle helix	Crawford et al., 1973
61	Normalized frequency of turn	Crawford et al., 1973
62	Size	Dawson, 1972

63	Amino acid composition	Dayhoff et al., 1978a
65	Membrane preference for cytochrome b: MPH89	Degli Esposti et al., 1990
66	Average membrane preference: AMP07	Degli Esposti et al., 1990
67	Consensus normalized hydrophobicity scale	Eisenberg, 1984
68	Solvation free energy	Eisenberg-McLachlan, 1986
69	Atom-based hydrophobic moment	Eisenberg-McLachlan, 1986
70	Direction of hydrophobic moment	Eisenberg-McLachlan, 1986
71	Molecular weight	Fasman, 1976
72	Melting point	Fasman, 1976
73	Optical rotation	Fasman, 1976
74	pK-N	Fasman, 1976
75	pK-C	Fasman, 1976
76	Hydrophobic parameter pi	Fauchere-Pliska, 1983
77	Graph shape index	Fauchere et al., 1988
78	Smoothed upsilon steric parameter	Fauchere et al., 1988
79	Normalized van der Waals volume	Fauchere et al., 1988
80	STERIMOL length of the side chain	Fauchere et al., 1988
81	STERIMOL minimum width of the side chain	Fauchere et al., 1988
82	STERIMOL maximum width of the side chain	Fauchere et al., 1988
83	N.m.r. chemical shift of alpha-carbon	Fauchere et al., 1988
84	Localized electrical effect	Fauchere et al., 1988
85	Number of hydrogen bond donors	Fauchere et al., 1988
86	Number of full nonbonding orbitals	Fauchere et al., 1988
87	Positive charge	Fauchere et al., 1988
88	Negative charge	Fauchere et al., 1988
89	pK-a	RCOOH
91	Helix initiation parameter at posision i-1	Finkelstein et al., 1991
92	Helix initiation parameter at posision i,i+1,i+2	Finkelstein et al., 1991
93	Helix termination parameter at posision j-2,j-1,j	Finkelstein et al.,1991
94	Helix termination parameter at posision j+1	Finkelstein et al., 1991
96	Alpha-helix indices	Geisow-Roberts, 1980

97	Alpha-helix indices for alpha-proteins	Geisow-Roberts, 1980
98	Alpha-helix indices for beta-proteins	Geisow-Roberts, 1980
99	Alpha-helix indices for alpha/beta-proteins	Geisow-Roberts, 1980
100	Beta-strand indices	Geisow-Roberts, 1980
101	Beta-strand indices for beta-proteins	Geisow-Roberts, 1980
102	Beta-strand indices for alpha/beta-proteins	Geisow-Roberts, 1980
103	Aperiodic indices	Geisow-Roberts, 1980
104	Aperiodic indices for alpha-proteins	Geisow-Roberts, 1980
105	Aperiodic indices for beta-proteins	Geisow-Roberts, 1980
106	Aperiodic indices for alpha/beta-proteins	Geisow-Roberts, 1980
107	Hydrophobicity factor	Goldsack-Chalifoux, 1973
108	Residue volume	Goldsack-Chalifoux, 1973
109	Composition	Grantham, 1974
110	Polarity	Grantham, 1974
111	Volume	Grantham, 1974
112	Partition energy	Guy, 1985
113	Hydration number , Cited by Charton-Charton	Hopfinger, 1971
114	Hydrophilicity value	Hopp-Woods, 1981
115	Heat capacity	Hutchens, 1970
116	Absolute entropy	Hutchens, 1970
117	Entropy of formation	Hutchens, 1970
118	Normalized relative frequency of alpha-helix	Isogai et al., 1980
119	Normalized relative frequency of extended structure	Isogai et al.,1980
120	Normalized relative frequency of bend	Isogai et al., 1980
121	Normalized relative frequency of bend R	Isogai et al., 1980
122	Normalized relative frequency of bend S	Isogai et al., 1980
123	Normalized relative frequency of helix end	Isogai et al., 1980
124	Normalized relative frequency of double bend	Isogai et al., 1980
125	Normalized relative frequency of coil	Isogai et al., 1980
126	Average accessible surface area	Janin et al., 1978
127	Percentage of buried residues	Janin et al., 1978

128	Percentage of exposed residues	Janin et al., 1978
129	Ratio of buried and accessible molar fractions	Janin, 1979
130	Transfer free energy	Janin, 1979
131	Hydrophobicity	Jones, 1975
132	pK	#NAME?
133	Relative frequency of occurrence	Jones et al., 1992
134	Relative mutability	Jones et al., 1992
135	Amino acid distribution	Jukes et al., 1975
136	Sequence frequency	Jungck, 1978
137	Average relative probability of helix	Kanehisa-Tsong, 1980
138	Average relative probability of beta-sheet	Kanehisa-Tsong, 1980
139	Average relative probability of inner helix	Kanehisa-Tsong, 1980
140	Average relative probability of inner beta-sheet	Kanehisa-Tsong, 1980
141	Flexibility parameter for no rigid neighbors	Karplus-Schulz, 1985
142	Flexibility parameter for one rigid neighbor	Karplus-Schulz, 1985
143	Flexibility parameter for two rigid neighbors	Karplus-Schulz, 1985
144	The Kerr-constant increments	Khanarian-Moore, 1980
145	Net charge	Klein et al., 1984
147	Side chain interaction parameter	Krigbaum-Komoriya, 1979
148	Fraction of site occupied by water	Krigbaum-Komoriya, 1979
149	Side chain volume	Krigbaum-Komoriya, 1979
150	Hydropathy index	Kyte-Doolittle, 1982
151	Transfer free energy, CHP/water	Lawson et al., 1984
152	Hydrophobic parameter	Levitt, 1976
153	Distance between C-alpha and centroid of side chain	Levitt, 1976
154	Side chain angle theta	AAR
155	Side chain torsion angle phi	AAAR
156	Radius of gyration of side chain	Levitt, 1976
157	van der Waals parameter R0	Levitt, 1976
158	van der Waals parameter epsilon	Levitt, 1976
159	Normalized frequency of alpha-helix, with weights	Levitt, 1978

160	Normalized frequency of beta-sheet, with weights	Levitt, 1978
161	Normalized frequency of reverse turn, with weights	Levitt, 1978
162	Normalized frequency of alpha-helix, unweighted	Levitt, 1978
163	Normalized frequency of beta-sheet, unweighted	Levitt, 1978
164	Normalized frequency of reverse turn, unweighted	Levitt, 1978
165	Frequency of occurrence in beta-bends	Lewis et al., 1971
166	Conformational preference for all beta-strands	Lifson-Sander, 1979
167	Conformational preference for parallel beta-strands	Lifson-Sander,1979
168	Conformational preference for antiparallel beta-strands	Lifson-Sander,1979
169	Average surrounding hydrophobicity	Manavalan-Ponnuswamy, 1978
172	Normalized frequency of zeta R	Maxfield-Scheraga, 1976
173	Normalized frequency of left-handed alpha-helix	Maxfield-Scheraga,1976
174	Normalized frequency of zeta L	Maxfield-Scheraga, 1976
175	Normalized frequency of alpha region	Maxfield-Scheraga, 1976
176	Refractivity , Cited by Jones	McMeekin et al., 1964
177	Retention coefficient in HPLC, pH7.4	Meek, 1980
178	Retention coefficient in HPLC, pH2.1	Meek, 1980
179	Retention coefficient in NaClO4	Meek-Rossetti, 1981
180	Retention coefficient in NaH2PO4	Meek-Rossetti, 1981
181	Average reduced distance for C-alpha	Meirovitch et al., 1980
182	Average reduced distance for side chain	Meirovitch et al., 1980
183	Average side chain orientation angle	Meirovitch et al., 1980
184	Effective partition energy	Miyazawa-Jernigan, 1985
186	Normalized frequency of bata-structure	Nagano, 1973
188	AA composition of total proteins	Nakashima et al., 1990
189	SD of AA composition of total proteins	Nakashima et al., 1990
190	AA composition of mt-proteins	Nakashima et al., 1990
191	Normalized composition of mt-proteins	Nakashima et al., 1990
192	AA composition of mt-proteins from animal	Nakashima et al., 1990
193	Normalized composition from animal	Nakashima et al., 1990
194	AA composition of mt-proteins from fungi and plant	Nakashima et al., 1990

195	Normalized composition from fungi and plant	Nakashima et al., 1990
196	AA composition of membrane proteins	Nakashima et al., 1990
197	Normalized composition of membrane proteins	Nakashima et al., 1990
198	Transmembrane regions of non-mt-proteins	Nakashima et al., 1990
199	Transmembrane regions of mt-proteins	Nakashima et al., 1990
200	Ratio of average and computed composition	Nakashima et al., 1990
201	AA composition of CYT of single-spanning proteins	Nakashima-Nishikawa,1992
202	AA composition of CYT2 of single-spanning proteins	Nakashima-Nishikawa, 1992
203	AA composition of EXT of single-spanning proteins	Nakashima-Nishikawa,1992
204	AA composition of EXT2 of single-spanning proteins	Nakashima-Nishikawa, 1992
205	AA composition of MEM of single-spanning proteins	Nakashima-Nishikawa,1992
206	AA composition of CYT of multi-spanning proteins	Nakashima-Nishikawa,1992
207	AA composition of EXT of multi-spanning proteins	Nakashima-Nishikawa,1992
208	AA composition of MEM of multi-spanning proteins	Nakashima-Nishikawa,1992
209	8 A contact number	Nishikawa-Ooi, 1980
210	14 A contact number	Nishikawa-Ooi, 1986
211	Transfer energy, organic solvent/water	Nozaki-Tanford, 1971
212	Average non-bonded energy per atom	Oobatake-Ooi, 1977
213	Short and medium range non-bonded energy per atom	Oobatake-Ooi, 1977
214	Long range non-bonded energy per atom	Oobatake-Ooi, 1977
215	Average non-bonded energy per residue	Oobatake-Ooi, 1977
216	Short and medium range non-bonded energy per residue	Oobatake-Ooi,1977
217	Optimized beta-structure-coil equilibrium constant	Oobatake et al.,1985
218	Optimized propensity to form reverse turn	Oobatake et al., 1985
219	Optimized transfer energy parameter	Oobatake et al., 1985
220	Optimized average non-bonded energy per atom	Oobatake et al., 1985
221	Optimized side chain interaction parameter	Oobatake et al., 1985
222	Normalized frequency of alpha-helix from LG	Palau et al., 1981
223	Normalized frequency of alpha-helix from CF	Palau et al., 1981
224	Normalized frequency of beta-sheet from LG	Palau et al., 1981

225	Normalized frequency of beta-sheet from CF	Palau et al., 1981
226	Normalized frequency of turn from LG	Palau et al., 1981
227	Normalized frequency of turn from CF	Palau et al., 1981
228	Normalized frequency of alpha-helix in all-alpha class	Palau et al., 1981
229	Normalized frequency of alpha-helix in alpha+beta class	Palau et al.,1981
230	Normalized frequency of alpha-helix in alpha/beta class	Palau et al.,1981
231	Normalized frequency of beta-sheet in all-beta class	Palau et al.,1981
232	Normalized frequency of beta-sheet in alpha+beta class	Palau et al., 1981
233	Normalized frequency of beta-sheet in alpha/beta class	Palau et al., 1981
234	Normalized frequency of turn in all-alpha class	Palau et al., 1981
235	Normalized frequency of turn in all-beta class	Palau et al., 1981
236	Normalized frequency of turn in alpha+beta class	Palau et al., 1981
237	Normalized frequency of turn in alpha/beta class	Palau et al., 1981
238	HPLC parameter	Parker et al., 1986
239	Partition coefficient	Pliska et al., 1981
240	Surrounding hydrophobicity in folded form	Ponnuswamy et al., 1980
241	Average gain in surrounding hydrophobicity	Ponnuswamy et al., 1980
242	Average gain ratio in surrounding hydrophobicity	Ponnuswamy et al.,1980
243	Surrounding hydrophobicity in alpha-helix	Ponnuswamy et al., 1980
244	Surrounding hydrophobicity in beta-sheet	Ponnuswamy et al., 1980
245	Surrounding hydrophobicity in turn	Ponnuswamy et al., 1980
246	Accessibility reduction ratio	Ponnuswamy et al., 1980
247	Average number of surrounding residues	Ponnuswamy et al., 1980
248	Intercept in regression analysis	Prabhakaran-Ponnuswamy, 1982
249	Slope in regression analysis x 1.0E1	Prabhakaran-Ponnuswamy, 1982
250	Correlation coefficient in regression analysis	Prabhakaran-Ponnuswamy,1982
251	Hydrophobicity	Prabhakaran, 1990
252	Relative frequency in alpha-helix	Prabhakaran, 1990
253	Relative frequency in beta-sheet	Prabhakaran, 1990
254	Relative frequency in reverse-turn	Prabhakaran, 1990
255	Helix-coil equilibrium constant	Ptitsyn-Finkelstein, 1983

256	Beta-coil equilibrium constant	Ptitsyn-Finkelstein, 1983
257	Weights for alpha-helix at the window position of -6	Qian-Sejnowski, 1988
258	Weights for alpha-helix at the window position of -5	Qian-Sejnowski, 1988
259	Weights for alpha-helix at the window position of -4	Qian-Sejnowski, 1988
260	Weights for alpha-helix at the window position of -3	Qian-Sejnowski, 1988
261	Weights for alpha-helix at the window position of -2	Qian-Sejnowski, 1988
262	Weights for alpha-helix at the window position of -1	Qian-Sejnowski, 1988
263	Weights for alpha-helix at the window position of 0	Qian-Sejnowski,1988
264	Weights for alpha-helix at the window position of 1	Qian-Sejnowski,1988
265	Weights for alpha-helix at the window position of 2	Qian-Sejnowski,1988
266	Weights for alpha-helix at the window position of 3	Qian-Sejnowski,1988
267	Weights for alpha-helix at the window position of 4	Qian-Sejnowski,1988
268	Weights for alpha-helix at the window position of 5	Qian-Sejnowski,1988
269	Weights for alpha-helix at the window position of 6	Qian-Sejnowski,1988
270	Weights for beta-sheet at the window position of -6	Qian-Sejnowski,1988
271	Weights for beta-sheet at the window position of -5	Qian-Sejnowski,1988
272	Weights for beta-sheet at the window position of -4	Qian-Sejnowski,1988
273	Weights for beta-sheet at the window position of -3	Qian-Sejnowski,1988
274	Weights for beta-sheet at the window position of -2	Qian-Sejnowski,1988
275	Weights for beta-sheet at the window position of -1	Qian-Sejnowski,1988
276	Weights for beta-sheet at the window position of 0	Qian-Sejnowski,1988
277	Weights for beta-sheet at the window position of 1	Qian-Sejnowski,1988
278	Weights for beta-sheet at the window position of 2	Qian-Sejnowski,1988
279	Weights for beta-sheet at the window position of 3	Qian-Sejnowski,1988
280	Weights for beta-sheet at the window position of 4	Qian-Sejnowski,1988
281	Weights for beta-sheet at the window position of 5	Qian-Sejnowski,1988
282	Weights for beta-sheet at the window position of 6	Qian-Sejnowski,1988
283	Weights for coil at the window position of -6	Qian-Sejnowski, 1988
284	Weights for coil at the window position of -5	Qian-Sejnowski, 1988
285	Weights for coil at the window position of -4	Qian-Sejnowski, 1988
286	Weights for coil at the window position of -3	Qian-Sejnowski, 1988

287	Weights for coil at the window position of -2	Qian-Sejnowski, 1988
288	Weights for coil at the window position of -1	Qian-Sejnowski, 1988
289	Weights for coil at the window position of 0	Qian-Sejnowski, 1988
290	Weights for coil at the window position of 1	Qian-Sejnowski, 1988
291	Weights for coil at the window position of 2	Qian-Sejnowski, 1988
292	Weights for coil at the window position of 3	Qian-Sejnowski, 1988
293	Weights for coil at the window position of 4	Qian-Sejnowski, 1988
294	Weights for coil at the window position of 5	Qian-Sejnowski, 1988
295	Weights for coil at the window position of 6	Qian-Sejnowski, 1988
298	Side chain orientational preference	Rackovsky-Scheraga, 1977
299	Average relative fractional occurrence in A0	i
300	Average relative fractional occurrence in AR	i
301	Average relative fractional occurrence in AL	i
302	Average relative fractional occurrence in EL	i
303	Average relative fractional occurrence in E0	i
304	Average relative fractional occurrence in ER	i
305	Average relative fractional occurrence in A0	i-1
306	Average relative fractional occurrence in AR	i-1
307	Average relative fractional occurrence in AL	i-1
308	Average relative fractional occurrence in EL	i-1
309	Average relative fractional occurrence in E0	i-1
310	Average relative fractional occurrence in ER	i-1
311	Value of theta	i
312	Value of theta	i-1
313	Transfer free energy from chx to wat	Radzicka-Wolfenden, 1988
314	Transfer free energy from oct to wat	Radzicka-Wolfenden, 1988
315	Transfer free energy from vap to chx	Radzicka-Wolfenden, 1988
316	Transfer free energy from chx to oct	Radzicka-Wolfenden, 1988
317	Transfer free energy from vap to oct	Radzicka-Wolfenden, 1988
318	Accessible surface area	Radzicka-Wolfenden, 1988

319	Energy transfer from out to in	95%buried
320	Mean polarity	Radzicka-Wolfenden, 1988
321	Relative preference value at N"	Richardson-Richardson, 1988
322	Relative preference value at N'	Richardson-Richardson, 1988
323	Relative preference value at N-cap	Richardson-Richardson, 1988
324	Relative preference value at N1	Richardson-Richardson, 1988
325	Relative preference value at N2	Richardson-Richardson, 1988
326	Relative preference value at N3	Richardson-Richardson, 1988
327	Relative preference value at N4	Richardson-Richardson, 1988
328	Relative preference value at N5	Richardson-Richardson, 1988
329	Relative preference value at Mid	Richardson-Richardson, 1988
330	Relative preference value at C5	Richardson-Richardson, 1988
331	Relative preference value at C4	Richardson-Richardson, 1988
332	Relative preference value at C3	Richardson-Richardson, 1988
333	Relative preference value at C2	Richardson-Richardson, 1988
334	Relative preference value at C1	Richardson-Richardson, 1988
335	Relative preference value at C-cap	Richardson-Richardson, 1988
336	Relative preference value at C'	Richardson-Richardson, 1988
337	Relative preference value at C"	Richardson-Richardson, 1988
338	Information measure for alpha-helix	Robson-Suzuki, 1976
339	Information measure for N-terminal helix	Robson-Suzuki, 1976
340	Information measure for middle helix	Robson-Suzuki, 1976
341	Information measure for C-terminal helix	Robson-Suzuki, 1976
342	Information measure for extended	Robson-Suzuki, 1976
343	Information measure for pleated-sheet	Robson-Suzuki, 1976
344	Information measure for extended without H-bond	Robson-Suzuki, 1976
345	Information measure for turn	Robson-Suzuki, 1976
346	Information measure for N-terminal turn	Robson-Suzuki, 1976

347	Information measure for middle turn	Robson-Suzuki, 1976
348	Information measure for C-terminal turn	Robson-Suzuki, 1976
349	Information measure for coil	Robson-Suzuki, 1976
350	Information measure for loop	Robson-Suzuki, 1976
351	Hydration free energy	Robson-Osguthorpe, 1979
352	Mean area buried on transfer	Rose et al., 1985
353	Mean fractional area loss	Rose et al., 1985
354	Side chain hydropathy, uncorrected for solvation	Roseman, 1988
355	Side chain hydropathy, corrected for solvation	Roseman, 1988
356	Loss of Side chain hydropathy by helix formation	Roseman, 1988
357	Transfer free energy , Cited by Charton-Charton	Simon, 1976
358	Principal component I	Sneath, 1966
359	Principal component II	Sneath, 1966
360	Principal component III	Sneath, 1966
361	Principal component IV	Sneath, 1966
362	Zimm-Bragg parameter s at 20 C	Sueki et al., 1984
363	Zimm-Bragg parameter sigma x 1.0E4	Sueki et al., 1984
364	Optimal matching hydrophobicity	Sweet-Eisenberg, 1983
366	Normalized frequency of isolated helix	Tanaka-Scheraga, 1977
367	Normalized frequency of extended structure	Tanaka-Scheraga, 1977
368	Normalized frequency of chain reversal R	Tanaka-Scheraga, 1977
369	Normalized frequency of chain reversal S	Tanaka-Scheraga, 1977
370	Normalized frequency of chain reversal D	Tanaka-Scheraga, 1977
371	Normalized frequency of left-handed helix	Tanaka-Scheraga, 1977
373	Normalized frequency of coil	Tanaka-Scheraga, 1977
374	Normalized frequency of chain reversal	Tanaka-Scheraga, 1977
375	Relative population of conformational state A	Vasquez et al., 1983
376	Relative population of conformational state C	Vasquez et al., 1983
377	Relative population of conformational state E	Vasquez et al., 1983
378	Electron-ion interaction potential	Veljkovic et al., 1985
379	Bitterness	Venanzi, 1984

380	Transfer free energy to lipophilic phase	von Heijne-Blomberg, 1979
381	Average interactions per side chain atom	Warme-Morgan, 1978
382	RF value in high salt chromatography	Weber-Lacey, 1978
383	Propensity to be buried inside	Wertz-Scheraga, 1978
384	Free energy change of epsilon to epsilon	i
385	Free energy change of alpha to alpha	Ri
386	Free energy change of epsilon to alpha	i
387	Polar requirement	Woese, 1973
388	Hydration potential	Wolfenden et al., 1981
389	Principal property value z1	Wold et al., 1987
390	Principal property value z2	Wold et al., 1987
391	Principal property value z3	Wold et al., 1987
392	Unfolding Gibbs energy in water, pH7.0	Yutani et al., 1987
393	Unfolding Gibbs energy in water, pH9.0	Yutani et al., 1987
394	Activation Gibbs energy of unfolding, pH7.0	Yutani et al., 1987
395	Activation Gibbs energy of unfolding, pH9.0	Yutani et al., 1987
396	Dependence of partition coefficient on ionic strength	Zaslavsky etal., 1982
398	Bulkiness	Zimmerman et al., 1968
400	Isoelectric point	Zimmerman et al., 1968
401	RF rank	Zimmerman et al., 1968
402	Normalized positional residue frequency at helix termini N4'	Aurora-Rose, 1998
404	Normalized positional residue frequency at helix termini N''	Aurora-Rose, 1998
405	Normalized positional residue frequency at helix termini N'	Aurora-Rose, 1998
406	Normalized positional residue frequency at helix termini Nc	Aurora-Rose, 1998
407	Normalized positional residue frequency at helix termini N1	Aurora-Rose, 1998
408	Normalized positional residue frequency at helix termini N2	Aurora-Rose, 1998
409	Normalized positional residue frequency at helix termini N3	Aurora-Rose, 1998
410	Normalized positional residue frequency at helix termini N4	Aurora-Rose, 1998
411	Normalized positional residue frequency at helix termini N5	Aurora-Rose, 1998

412	Normalized positional residue frequency at helix termini C5	Aurora-Rose, 1998
413	Normalized positional residue frequency at helix termini C4	Aurora-Rose, 1998
414	Normalized positional residue frequency at helix termini C3	Aurora-Rose, 1998
415	Normalized positional residue frequency at helix termini C2	Aurora-Rose, 1998
416	Normalized positional residue frequency at helix termini C1	Aurora-Rose, 1998
417	Normalized positional residue frequency at helix termini Cc	Aurora-Rose, 1998
418	Normalized positional residue frequency at helix termini C'	Aurora-Rose, 1998
419	Normalized positional residue frequency at helix termini C''	Aurora-Rose, 1998
422	Delta G values for the peptides extrapolated to 0 M urea	O'Neil-DeGrado, 1990
423	Helix formation parameters	delta delta G
424	Normalized flexibility parameters , average	B-values
425	Normalized flexibility parameters for each residue surrounded by none rigid neighbours	B-values
426	Normalized flexibility parameters for each residue surrounded by one rigid neighbours	B-values
427	Normalized flexibility parameters for each residue surrounded by two rigid neighbours	B-values
428	Free energy in alpha-helical conformation	Munoz-Serrano, 1994
429	Free energy in alpha-helical region	Munoz-Serrano, 1994
430	Free energy in beta-strand conformation	Munoz-Serrano, 1994
431	Free energy in beta-strand region	Munoz-Serrano, 1994
433	Free energies of transfer of AcWl-X-LL peptides from bilayer interfaceto water	Wimley-White, 1996
434	Thermodynamic beta sheet propensity	Kim-Berg, 1993
435	Turn propensity scale for transmembrane helices	Monne et al., 1999
436	Alpha helix propensity of position 44 in T4 lysozyme	Blaber et al.,1993
437	p-Values of mesophilic proteins based on the distributions of B values	Parthasarathy-Murthy, 2000
438	p-Values of thermophilic proteins based on the distributions of B values	Parthasarathy-Murthy, 2000
439	Distribution of amino acid residues in the 18 non-redundant familiesofthermophilic proteins	Kumar et al., 2000
440	Distribution of amino acid residues in the 18 non-redundant familiesofmesophilic proteins	Kumar et al., 2000
441	Distribution of amino acid residues in the alpha-helices in thermophilic proteins	Kumar et al., 2000
442	Distribution of amino acid residues in the alpha-helices in mesophilicproteins	Kumar et al., 2000

443	Side-chain contribution to protein stability	kJ/mol
444	Propensity of amino acids within pi-helices	Fodje-Al-Karadaghi, 2002
445	Hydropathy scale based on self-information values in the two-state model	5% accessibility
446	Hydropathy scale based on self-information values in the two-state model	9% accessibility
447	Hydropathy scale based on self-information values in the two-state model	16% accessibility
448	Hydropathy scale based on self-information values in the two-state model	20% accessibility
449	Hydropathy scale based on self-information values in the two-state model	25% accessibility
450	Hydropathy scale based on self-information values in the two-state model	36% accessibility
451	Hydropathy scale based on self-information values in the two-state model	50% accessibility
452	Averaged turn propensities in a transmembrane helix	Monne et al., 1999
453	Alpha-helix propensity derived from designed sequences	Koehl-Levitt, 1999
454	Beta-sheet propensity derived from designed sequences	Koehl-Levitt,1999
455	Composition of amino acids in extracellular proteins	percent
456	Composition of amino acids in anchored proteins	percent
457	Composition of amino acids in membrane proteins	percent
458	Composition of amino acids in intracellular proteins	percent
459	Composition of amino acids in nuclear proteins	percent
460	Surface composition of amino acids in intracellular proteins of thermophiles	percent
461	Surface composition of amino acids in intracellular proteins of mesophiles	percent
462	Surface composition of amino acids in extracellular proteins of mesophiles	percent
463	Surface composition of amino acids in nuclear proteins	percent
464	Interior composition of amino acids in intracellular proteins of thermophiles	percent
465	Interior composition of amino acids in intracellular proteins of mesophiles	percent
466	Interior composition of amino acids in extracellular proteins of mesophiles	percent
467	Interior composition of amino acids in nuclear proteins	percent
468	Entire chain composition of amino acids in intracellular proteins ofthermophiles	percent
469	Entire chain composition of amino acids in intracellular proteins ofmesophiles	percent
470	Entire chain composition of amino acids in extracellular proteins ofmesophiles	percent

471	Entire chain compositino of amino acids in nuclear proteins	percent
482	Amphiphilicity index	Mitaku et al., 2002
483	Volumes including the crystallographic waters using the ProtOr	Tsaietal., 1999
484	Volumes not including the crystallographic waters using the ProtOr	Tsai et al., 1999
485	Electron-ion interaction potential values	Cosic, 1994
486	Hydrophobicity scales	Ponnuswamy, 1993
487	Hydrophobicity coefficient in RP-HPLC, C18 with 0.1%TFA/MeCN/H2O	Wilceet al. 1995
488	Hydrophobicity coefficient in RP-HPLC, C8 with 0.1%TFA/MeCN/H2O	Wilceet al. 1995
489	Hydrophobicity coefficient in RP-HPLC, C4 with 0.1%TFA/MeCN/H2O	Wilceet al. 1995
490	Hydrophobicity coefficient in RP-HPLC, C18 with 0.1%TFA/2-PrOH/MeCN/H2O	Wilce et al. 1995
491	Hydrophilicity scale	Kuhn et al., 1995
492	Retention coefficient at pH 2	Guo et al., 1986
493	Modified Kyte-Doolittle hydrophobicity scale	Juretic et al., 1998
494	Interactivity scale obtained from the contact matrix	Bastolla et al.,2005
495	Interactivity scale obtained by maximizing the mean of correlationcoefficient over single-domain globular proteins	Bastolla et al., 2005
496	Interactivity scale obtained by maximizing the mean of correlationcoefficient over pairs of sequences sharing the TIM barrel fold	Bastolla et al., 2005
497	Linker propensity index	Suyama-Ohara, 2003
498	Knowledge-based membrane-propensity scale from 1D_Helix in MPtopo databases	Punta-Maritan, 2003
499	Knowledge-based membrane-propensity scale from 3D_Helix in MPtopo databases	Punta-Maritan, 2003
500	Linker propensity from all dataset	George-Heringa, 2003
501	Linker propensity from 1-linker dataset	George-Heringa, 2003
502	Linker propensity from 2-linker dataset	George-Heringa, 2003
503	Linker propensity from 3-linker dataset	George-Heringa, 2003
504	Linker propensity from small dataset	linker length is less than sixresidues
505	Linker propensity from medium dataset	linker length is between sixand14 residues
506	Linker propensity from long dataset	linker length is greater than 14residues
507	Linker propensity from helical dataset	annotated by DSSP
508	Linker propensity from non-helical dataset	annotated by DSSP

509	The stability scale from the knowledge-based atom-atom potential	Zhou-Zhou, 2004
510	The relative stability scale extracted from mutation experiments	Zhou-Zhou, 2004
511	Buriability	Zhou-Zhou, 2004
512	Linker index	Bae et al., 2005
513	Mean volumes of residues buried in protein interiors	Harpaz et al.,1994
514	Average volumes of residues	Pontius et al., 1996
515	Hydrostatic pressure asymmetry index, PAI	Di Giulio, 2005
517	Average internal preferences	Olsen, 1980
518	Hydrophobicity-related index	Kidera et al., 1985
519	Apparent partition energies calculated from Wertz-Scheraga index	Guy,1985
521	Apparent partition energies calculated from Janin index	Guy, 1985
522	Apparent partition energies calculated from Chothia index	Guy, 1985
525	Weights from the IFH scale	Jacobs-White, 1989
526	Hydrophobicity index, 3.0 pH	Cowan-Whittaker, 1990
527	Scaled side chain hydrophobicity values	Black-Mould, 1991
528	Hydrophobicity scale from native protein structures	Casari-Sippl, 1992
529	NNEIG index	Cornette et al., 1987
530	SWEIG index	Cornette et al., 1987
531	PRIFT index	Cornette et al., 1987
532	PRILS index	Cornette et al., 1987
533	ALTFT index	Cornette et al., 1987
534	ALTLS index	Cornette et al., 1987
535	TOTFT index	Cornette et al., 1987
536	TOTLS index	Cornette et al., 1987
537	Relative partition energies derived by the Bethe approximation	Miyazawa-Jernigan, 1999
538	Optimized relative partition energies - method A	Miyazawa-Jernigan,1999
539	Optimized relative partition energies - method B	Miyazawa-Jernigan,1999
540	Optimized relative partition energies - method C	Miyazawa-Jernigan,1999
541	Optimized relative partition energies - method D	Miyazawa-Jernigan,1999
543	Hydrophobicity index	Fasman, 1989

544	Number of vertices	order of the graph
545	Number of edges	size of the graph
546	Total weighted degree of the graph	obtained by adding all the weightsof all the vertices
547	Weighted domination number	Karkbara-Knisley, 2016
548	Average eccentricity	Karkbara-Knisley, 2016
549	Radius	minimum eccentricity
550	Diameter	maximum eccentricity
551	Average weighted degree	total degree, divided by the number of vertices
552	Maximum eigenvalue of the weighted Laplacian matrix of the graph	Karkbara-Knisley, 2016
553	Minimum eigenvalue of the weighted Laplacian matrix of the graph	Karkbara-Knisley, 2016
554	Average eigenvalue of the Laplacian matrix of the the graph	Karkbara-Knisley, 2016
555	Second smallest eigenvalue of the Laplacian matrix of the graph	Karkbara-Knisley, 2016
556	Weighted domination number using the atomic number	Karkbara-Knisley, 2016
557	Average weighted eccentricity based on the the atomic number	Karkbara-Knisley, 2016
558	Weighted radius based on the atomic number	minimum eccentricity
559	Weighted diameter based on the atomic number	maximum eccentricity
560	Total weighted atomic number of the graph	obtained by summing all theatomic number of each of the vertices in the graph
561	Average weighted atomic number or degree based on atomic number in thegraph	Karkbara-Knisley, 2016
562	Weighted maximum eigenvalue based on the atomic numbers	Karkbara-Knisley, 2016
563	Weighted minimum eigenvalue based on the atomic numbers	Karkbara-Knisley, 2016
564	Weighted average eigenvalue based on the atomic numbers	Karkbara-Knisley, 2016
565	Weighted second smallest eigenvalue of the weighted Laplacian matrix	Karkbara-Knisley, 2016

Position	Property Indices
1	2, 9, 11, 29, 30, 36, 42, 54, 60, 67, 69, 70, 83, 103, 104, 114, 135, 145, 146, 252, 260, 263, 273, 281, 284, 344, 348, 351, 501
2	6, 9, 10, 11, 28, 30, 34, 35, 37, 42, 57, 63, 69, 70, 76, 83, 89, 97, 103, 104, 107, 109, 120, 121, 132, 135, 137, 140, 143, 145, 146, 147, 148, 157, 196, 252, 262, 263, 264, 266, 267, 268, 272, 274, 275, 285, 288, 291, 313, 344, 345, 348, 349, 351, 352, 357, 380, 428, 439, 449, 454, 463, 489, 501, 515
3	28, 30, 37, 42, 44, 54, 66, 69, 70, 83, 89, 97, 103, 104, 107, 112, 114, 120, 121, 135, 136, 137, 140, 142, 145, 146, 147, 148, 157, 188, 198, 199, 238, 252, 255, 256, 257, 258, 259, 262, 263, 275, 280, 282, 283, 288, 343, 345, 348, 349, 351, 352, 374, 400, 439, 464
4	3, 72, 74, 77, 128, 149, 178, 196, 198, 251, 252, 253, 255, 262, 265, 272, 273, 274, 275, 286, 289, 313, 315, 370, 379, 384, 428, 454, 501
5	9, 10, 11, 25, 29, 44, 48, 59, 65, 71, 72, 73, 74, 84, 85, 86, 91, 92, 93, 94, 105, 111, 124, 128, 129, 130, 131, 132, 134, 149, 155, 156, 168, 173, 174, 180, 182, 183, 186, 188, 189, 198, 222, 238, 250, 257, 258, 262, 263, 267, 275, 280, 282, 283, 289, 318, 334, 341, 343, 344, 345, 350, 351, 352, 407, 456, 457, 458, 470
6	3, 28, 42, 44, 59, 66, 69, 76, 78, 85, 86, 92, 93, 94, 103, 121, 126, 129, 132, 143, 145, 147, 155, 156, 173, 178, 197, 199, 238, 251, 255, 256, 258, 259, 262, 263, 264, 267, 275, 280, 282, 283, 288, 290, 315, 318, 350, 351, 352, 354, 374, 399, 400, 428, 439, 454, 456
7	2, 6, 30, 33, 37, 46, 76, 80, 97, 112, 114, 142, 143, 147, 148, 174, 196, 200, 250, 252, 258, 259, 262, 265, 267, 268, 269, 274, 284, 288, 289, 318, 343, 345, 348, 354, 357, 411, 525
8	3, 9, 10, 11, 36, 42, 44, 47, 54, 55, 58, 69, 70, 71, 77, 78, 83, 91, 103, 104, 107, 110, 121, 130, 131, 132, 136, 137, 140, 155, 157, 168, 180, 183, 188, 238, 254, 257, 267, 271, 272, 282, 286, 288, 289, 344, 345, 350, 352, 380, 411, 412, 446, 463
9	3, 9, 10, 12, 41, 55, 58, 71, 72, 73, 74, 76, 77, 110, 111, 124, 125, 126, 128, 149, 168, 178, 196, 200, 251, 262, 263, 265, 266, 272, 273, 275, 289, 290, 313, 314, 315, 344, 351, 352, 379, 515, 525
10	2, 9, 11, 28, 58, 71, 72, 73, 74, 77, 79, 84, 90, 92, 105, 111, 124, 125, 129, 130, 149, 156, 168, 174, 182, 186, 188, 198, 200, 250, 252, 254, 255, 256, 258, 259, 262, 267, 268, 269, 275, 280, 287, 288, 289, 290, 315, 318, 343, 344, 345, 354, 357, 370, 379, 392, 439, 525
11	6, 28, 30, 33, 34, 35, 36, 37, 42, 57, 63, 69, 70, 83, 89, 97, 103, 104, 107, 108, 109, 120, 121, 135, 136, 137, 140, 145, 146, 157, 178, 196, 198, 252, 253, 255, 256, 259, 262, 265, 280, 282, 283, 343, 345, 348, 349, 350, 357, 442, 501
12	3, 6, 29, 30, 34, 35, 36, 42, 57, 60, 66, 67, 69, 70, 83, 97, 103, 104, 106, 107, 108, 109, 135, 137, 140, 145, 157, 178, 181, 198, 199, 207, 218, 251, 255, 256, 282, 313, 343, 345, 348, 349, 357, 380, 397, 442, 447, 448, 454, 455, 464, 465, 472, 491
All	42, 69, 83, 97, 103, 198, 262, 282, 345, 348, 349, 351, 357

Appendix C: Peptide classifications

Table 1 DAG with 527 properties

Total (16569)	**Super (3)**	**Strong (451)**	**Weak (412)**	**Medium (15703)**
Predicted Super (296)	1 True positive	8 False positive	13 False positive	274 False positive
Predicted Strong (931)	0 False positive	143 True positive	6 False positive	782 False positive
Predicted Weak (346)	0 False positive	2 False positive	26 True positive	318 False positive
Predicted Medium (14996)	2 False positive	298 False positive	367 False positive	14329 True positive

Table 2 DAG with 292 unique properties

Total (16569)	**Super (3)**	**Strong (451)**	**Weak (412)**	**Medium (15703)**
Predicted Super (316)	1 True positive	9 False positive	13 False positive	293 False positive
Predicted Strong (864)	0 False positive	136 True positive	5 False positive	723 False positive
Predicted Weak (1166)	0 False positive	21 False positive	85 True positive	1080 False positive
Predicted Medium (14223)	2 False positive	285 False positive	309 False positive	13627 True positive

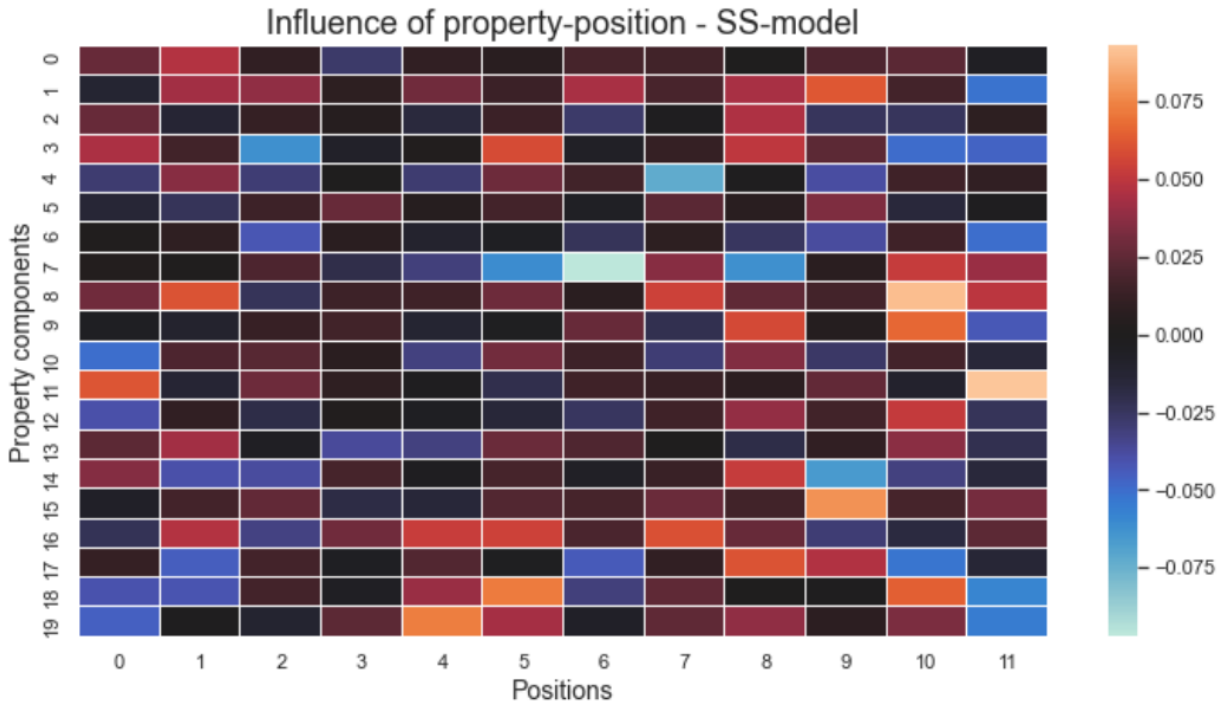
Table 3 DAG with 196 Shannon Entropy validated properties

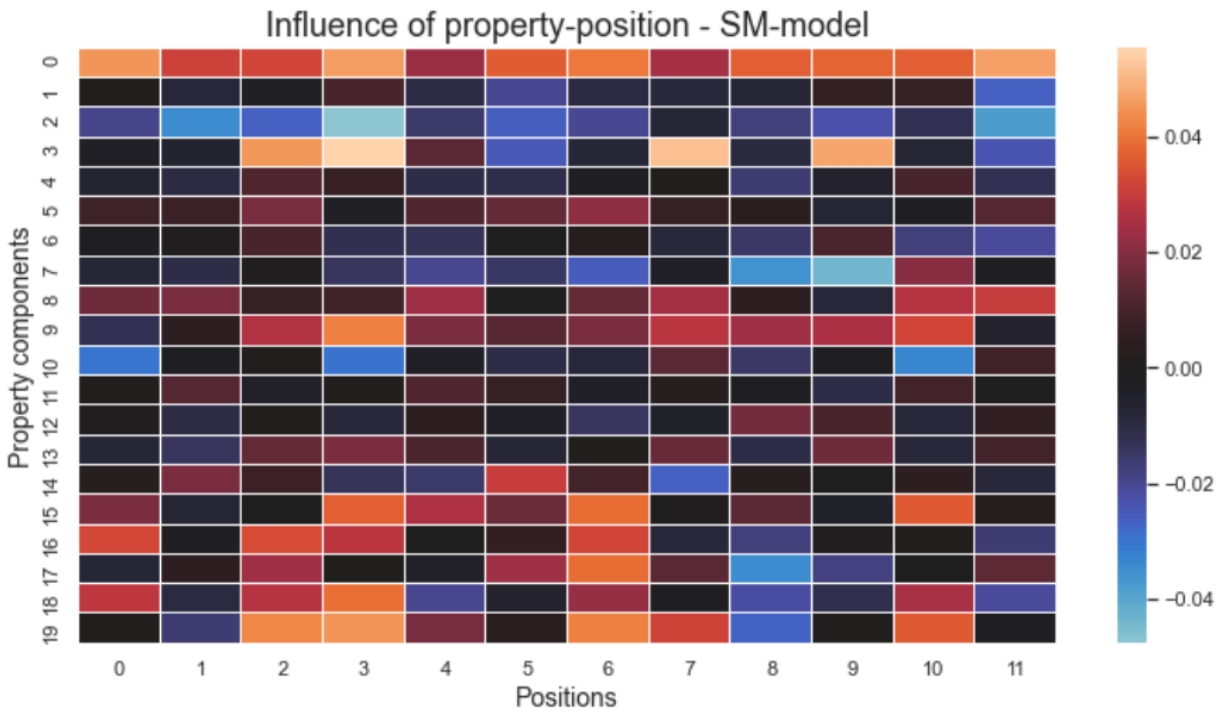
Total (16569)	**Super (3)**	**Strong (451)**	**Weak (412)**	**Medium (15703)**
Predicted Super (265)	1 True positive	9 False positive	16 False positive	239 False positive
Predicted Strong (880)	0 False positive	135 True positive	5 False positive	740 False positive
Predicted Weak (461)	0 False positive	24 False positive	36 True positive	401 False positive
Predicted Medium (14963)	2 False positive	283 False positive	355 False positive	14324 True positive

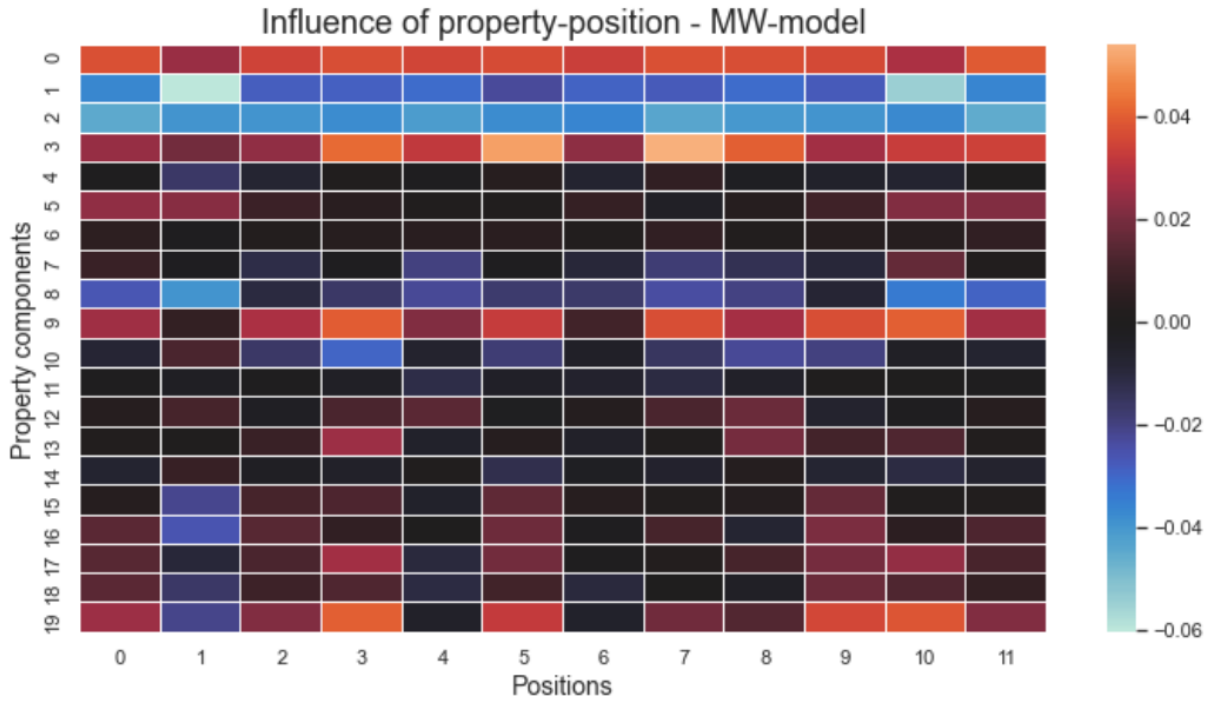
Table 4 DAG with 20 Principle components of the 527 properties

Total (16569)	**Super (3)**	**Strong (451)**	**Weak (412)**	**Medium (15703)**
Predicted Super (609)	1 True positive	14 False positive	23 False positive	571 False positive
Predicted Strong (948)	0 False positive	111 True positive	9 False positive	828 False positive
Predicted Weak (1112)	0 False positive	18 False positive	83 True positive	1011 False positive
Predicted Medium (13900)	2 False positive	308 False positive	297 False positive	13293 True positive

Appendix D: Influence of position and property component as per the 20 property component models






Influence of property-position - MW-model
Property components
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
Positions
0
1
2
3
4
5
6
7
8
9
10
11
0.04
0.02
0.00
−0.02
−0.04
−0.06

Appendix E: Some of the super binder candidates generated (out of 12000 +)

Peptide predictions	**COAM**
WPFDLRIKWEGF	1.236025
WPFDLRIKWEGL	1.22878
WPFDLRIKWEIF	1.201815
WPFDLRIKWEIL	1.19457
WPFDLRIKWDGF	1.228828
WPFDLRIKWDGL	1.221582
WPFDLRIKWDIF	1.194618
WPFDLRIKWDIL	1.187372
WPFDLRIKLEGF	1.199591
WPFDLRIKLEGL	1.192346
WPFDLRIKLEIF	1.165382
WPFDLRIKLEIL	1.158136