

STRUCTURAL AND TOPOLOGICAL VARIATIONS IN AMINO ACIDS ENCODED BY
SYNONYMOUS CODONS

by

Shengyuan Wang
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Bioinformatics and Computational Biology

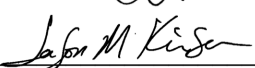
Committee:



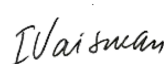
Dr. Iosif Vaisman, Committee Chair



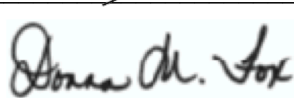
Dr. Dmitri Klimov, Committee Member



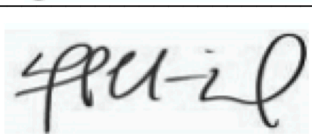
Dr. Jason Kinser, Committee Member



Dr. Iosif Vaisman, Director, School of
Systems Biology



Dr. Donna M. Fox, Associate Dean,
Office of Student Affairs & Special
Programs, College of Science



Dr. Fernando R. Miralles-Wilhelm
Dean, College of Science

Date: 2/14/2022

Spring Semester 2022
George Mason University
Fairfax, VA

Structural and Topological Variations in Amino Acids Encoded by Synonymous Codons

A Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy at George Mason University

by

Shengyuan Wang
Master of Science
North Carolina Central University, 2016
Bachelor of Science
China Agricultural University, 2013

Director: Iosif Vaisman, Professor
Department of Bioinformatics and Computational Biology

Spring Semester 2022
George Mason University
Fairfax, VA

Copyright 2022 Shengyuan Wang
All Rights Reserved

ACKNOWLEDGEMENTS

I would like to thank my mentor, Dr. Iosif Vaisman, for helping me and guiding me in my dissertation research. I would like to thank my committee, Dr. Dmitri Klimov and Dr. Jason Kinser, for offering me valuable critiques and suggestions to help me improve my work. I would like to thank the School of Systems Biology for providing me resources to conduct my research and complete my PhD study. I would like to thank my parents for their unconditional support. Finally, I want to thank my girlfriend, Dr. Suwei Wang, who always supports me in work and life.

TABLE OF CONTENTS

	Page
List of Tables	vii
List of Figures	ix
List of Equations	xiv
List of Abbreviations	xv
Abstract	xv
Chapter 1 : OVERVIEW	1
1.1 Introduction to silent mutations in proteins.....	1
1.2 Introduction to Delaunay tessellation method in protein structure analysis	3
1.3 Introduction to computational mutagenesis methods.....	4
1.4 Overview of the work presented here	4
Chapter 2 : FOUR-BODY POTENTIAL ANALYSIS USING DELAUNAY TESSELLATION AND THEIR APPLICATION IN CANCER-CAUSING SILENT MUTATION TOPLOGICAL ANALYSIS	6
2.1 Overview	6
2.2 Materials and methods	6
2.2.1 Protein sequences and structures database construction.....	6
2.2.2 Secondary structure preference of synonymous codons.....	8
2.2.3 Knowledge-based potential estimation.....	9
2.2.4 Comparison of amino acid level knowledge-based potential results with a previous study.....	11
2.2.5 Examining potentials distribution difference between synonymous codons....	12
2.2.6 Investigating protein structure and function change by silent mutations through knowledge-based potentials.....	13
2.2.6.1 Dataset creation.....	13
2.2.6.2 Four-body statistical potentials and computational mutagenesis.....	15
2.2.6.3 Protein variant feature vectors	15
2.3 Results and Discussions	16

2.3.1 Knowledge-based potential results	16
2.3.2 Secondary structure preference of synonymous codons results	20
2.3.3 Amino acid level potential estimations comparing with a previous study	21
2.3.4 Learning curves of potential estimations in the <i>10220culled</i> dataset.....	28
2.3.5 Potentials distribution difference between synonymous codons.....	30
2.3.6 Potentials comparison among protein subgroups	35
2.3.7 Linking protein structure and functions by silent mutations	37
2.3.8 Conclusions	40
Chapter 3 : EXPLORING FITNESS AND ACTIVITY OF PROTEIN MUTANTS WITH COMPUTATIONAL MUTAGENESIS AND MACHINE LEARNING TECHNIQUES	42
3.1 Overview	42
3.2 Materials and Methods	43
3.2.1 Four-body statistical potential estimation using Delaunay tessellation in β - lactamase	43
3.2.2 Computational mutagenesis on β -lactamase	44
3.2.3 β -lactamases variant feature vectors, machine learning, and model evaluation	45
3.2.4 Computational mutagenesis methods on protein activities	47
3.2.5 Computational mutagenesis methods on double mutations	49
3.3 Results and discussions	50
3.3.1 Fitness scores summary of β -lactamases	50
3.3.2 Residual scores distinguish between categories of β -lactamases amino acids.	53
3.3.3 Machine learning models for predicting β -lactamases variant fitness	53
3.3.4 Learning curves exploration in β -lactamase models	62
3.3.5 Inclusion of deep learning exploration	64
3.3.6 Mutagenesis on protein activities	66
3.3.7 Mutagenesis on double residue mutations.....	69
3.4 Conclusions	73
Chapter 4 : VALGUSHEL, A NOVEL METHOD TO IDENTIFY AND CHARACTERIZE KINKED ALPHA-HELICES.....	75
4.1 Overview	75
4.2 Introduction	77
4.3 Materials and Methods	80

4.3.1 ValgusHel Method Definition	80
4.3.1.1 Part 1: ValgusHel-geometry	80
4.3.1.2 Part 2: ValgusHel-topology	82
4.3.1.2.1 Part 2.1: t-numbers.....	82
4.3.1.2.2 Part 2.2: N-gram.....	83
4.3.2 Data set preparation	85
4.3.3 Kinked α -helices identification using ValgusHel geometry method.....	86
4.3.4 Association between sequence and structure in kinked α -helices	87
4.3.5 Prediction of normal, curve, and kinked residues based on variables (DSSP or Delaunay tessellation) in Random Forest Classification.....	89
4.3.6 Kinked α -helices identification using ValgusHel-topology method	89
4.4 Results and discussion.....	90
4.4.1 Characteristics of calculated helix angles.....	90
4.4.2 Validation of ValgusHel geometry method against HELANAL-plus, MC-HELAN, and AHAH	93
4.4.3 Amino acid residues frequency in different α -helices groups	95
4.4.4 Sequence similarity in different cluster and α -helices groups.....	96
4.4.5 Kinked and normal α -helices prediction using DSSP and Local profile.....	108
4.4.6 Random forest classification using Delaunay simplex type descriptor (t-numbers)	110
4.4.7 Agreement between ValgusHel-geometry and ValgusHel-topology methods using N-gram	112
4.5 CONCLUSIONS	115
Chapter 5 : CONCLUSIONS.....	116
APPENDIX.....	118
REFERENCES	119
BIOGRAPHY	125

LIST OF TABLES

Table	Page
Table 2-1 Proteins studied for silent mutations causing cancers	14
Table 2-2 Knowledge-based potential score examples in 10220culled dataset in this dissertation and in the dataset in a previous study	24
Table 2-3 BAR results of Naïve Bayes and Random Forest using 10-fold cross-validation (WEKA).....	39
Table 3-1 Protein mutation activity datasets.....	48
Table 3-2 Beta-lactamases protein structure-function relationship. All refers to the collection of all 4997 β -lactamase variants with experimental fitness data. C/NC is a subset of these variants and represents conservative/non-conservative amino acid substitutions of natural residues. The data in the right table is the average of the residual scores for the relevant subset of mutants. All numbers in parentheses on the graph or table row/column headers are counts of the total number of mutants in the subset.	52
Table 3-3 LOOCV performance on β -lactamases variant data sets at amino acid level. .	55
Table 3-4 LOOCV performance on β -lactamases variant data sets at codon level.	59
Table 3-5 Mean random forest leave-one-out cross-validation (LOOCV) prediction performance (β -lactamases variant Local Profiles) based on side chain polarities of the native and new amino acids at the mutated position.....	61
Table 3-6 Mean random forest leave-one-out cross-validation (LOOCV) prediction performance (β -lactamases variant Local Profiles) based on depth and secondary structure.....	62
Table 3-7 Comparison of 10-fold cross-validation prediction performance (β -lactamases variant Local Profiles) at the amino acid or codon level using Random Forest Classification (RFC) or Artificial Neural Networks (ANN).....	65
Table 3-8 Comparison of 10-fold cross-validation prediction performance (β -lactamases variant Local Profiles) at the amino acid or codon level using Random Forest Regression (RFR) or Artificial Neural Networks (ANN).....	66
Table 3-9 10-fold cross-validation, random forest regression results.....	67
Table 3-10 10-fold cross-validation, random forest classification, median, rescaled	68
Table 3-11 Model performance by using different training and test datasets in two protein activity categories	69
Table 3-12 10-fold cross validation, median, classification, instance =270,990	70
Table 3-13 10-fold cross validation, median, classification, instance number varies	71
Table 3-14 Residual scores, mutation fitness scores, and double mutation fitness score Pearson's correlation coefficients	72
Table 4-1 Simplex type stacking index in an example sequence.....	84

Table 4-2 Sequence identity and helix angles by clustering sequences.....	98
Table 4-3 Mean sequence similarity score with SD of different sequence groups by clustering helix angles. N=752 in each group.....	99
Table 4-4 Mean sequence identity score with SD of different sequence groups by clustering helix angles. N=752 in each group.....	100
Table 4-5 Clustering results of α -helix fragments (N=775)	101
Table 4-6 DSSP random forest 10-fold, evenly data set (n=15,487 for each class).	108
Table 4-7 Local profile random forest 10-fold, evenly data set (n=11,747 for each class).	109
Table 4-8 10-fold performance results for each class using random forest classification	112

LIST OF FIGURES

Figure	Page
Figure 2-1 Frequency of amino acids in 10220culled dataset.	17
Figure 2-2 The distribution of volume and tetrahedrality of simplices in different simplex types.	18
Figure 2-3 Log-likelihood ratio of Delaunay simplices at (A) amino acid level and (B) codon level. Simplex compositions were ranked from high potentials to low potentials on X-axis.	19
Figure 2-4 Chi-square test results of the observed and expected number of synonymous codons in 18 amino acids in different secondary structures. P-values ≥ 0.05 are highlighted in yellow.	21
Figure 2-5 Knowledge-based potential results comparison from 10220culled dataset and Taylor's (2006) dataset. (A) Scatter plot showing the correlation of estimated potentials for the same simplex composition between the two studies. A regression line was shown with estimated a correlation slope with an R-square. (B) Density plots of each simplex composition. Number of each simplex is the number of simplices in each simplex composition represented by amino acid residues.	23
Figure 2-6 Distribution of number of simplices in 10220culled and Taylor's dataset. R-square in Y-axis (left) is the correlation R-square between potentials estimated in 10220culled and potentials estimated in Taylor's study for the same simplex composition. Number of scatter potentials in Y-axis (right) is the number of unique simplex compositions. Least number of simplices in X-axis is the cutoff to restrict simplex compositions with a number of simplices no less than the cutoffs.	25
Figure 2-7 Impact of fLPS p-value and least number of simplices on the correlation R-square between potentials for the same simplex composition in 10220culled and Taylor's study. Y-axis is fLPS p-value in exponent of power of 10. Different color shows the effect of removing simplex compositions containing simplices less than N (0, 20, 40, 60, 80, or 100).	27
Figure 2-8 Impact of minimum number of simplices cutoff and Delaunay simplex edge length cutoffs on the correlation R-square between potentials for the same simplex composition in 10220culled and Taylor's study. Cutoff unit is angstrom (Å). Minimum number of simplices in X-axis is the cutoff to restrict simplex compositions with a number of simplices no less than the cutoffs.	28
Figure 2-9 Learning curve of number of training set. R-square is the correlation R-square between mean knowledge-based potentials in 10220culled dataset and the ones in the training set.	30

Figure 2-10 Z test results of the mean potentials of synonymous codon pairs of <i>xiziziti</i> and <i>xiziziti</i> , where x_i is the i -th synonymous codon of amino acid X and y_i, z_i, t_i are i th synonymous codons of amino acid Y, Z, T, and i (2, 3, 4, 6) is the number of synonymous codons depending on its specific amino acid. X is different from Y, Z, T while Y, Z, T may be the same amino acid. P-values of 0.05 is highlighted in the red dashed line.	31
Figure 2-11 Z test results of the mean potentials of synonymous codon pairs of <i>xiziziti</i> and <i>miyiziti</i> , where x_i and m_i is the i th synonymous codon of amino acid X and M respectively, and y_i, z_i, t_i are i th synonymous codons of amino acid Y, Z, T, and i (2, 3, 4, 6) is the number of synonymous codons depending on its specific amino acid. X/M is different from Y, Z, T while Y, Z, T may be the same amino acid. X and M can be different or same amino acid. P-values of 0.05 is highlighted in the red dashed line.	32
Figure 2-12 Z test results of the mean potentials of synonymous codon pairs of <i>xixiyizi</i> and <i>xixiyizi</i> , where x_i is the i -th synonymous codon of amino acid X and y_i, z_i are i -th synonymous codons of amino acid Y, Z, and i (2, 3, 4, 6) is the number of synonymous codons depending on its specific amino acid. X is different from Y, Z while Y, Z may be the same amino acid. X and M can are different amino acids. P-values of 0.05 is highlighted in the red dashed line.	33
Figure 2-13 Z test results of the mean potentials of synonymous codon pairs of <i>xixiyizi</i> and <i>mimiyizi</i> , where x_i and m_i is the i -th synonymous codon of amino acid X and M, respectively, and y_i, z_i are i -th synonymous codons of amino acid Y, Z, and i (2, 3, 4, 6) is the number of synonymous codons depending on its specific amino acid. X is different from Y, Z while Y, Z may be the same amino acid. P-values of 0.05 is highlighted in the red dashed line.	34
Figure 2-14 Potential comparison between enzymes vs. non-enzymes at (A) amino acid level, (C) codon level, (D) codon level (heatmap). Enzymes are random split into half and their potential scatter plot shows in (B).	36
Figure 2-15 Potential comparison between hydrolases vs. non-hydrolases at (A) amino acid level, (B) codon level and (C) codon level (heatmap)	37
Figure 2-16 Distribution of residue scores between cancer causing silent mutations vs. non-cancer-causing silent mutations in (A) histogram with all gene combined and (B) boxplot for each individual gene.....	38
Figure 2-17 Correlation between residual scores and SynMICdb scores.	40
Figure 3-1 CMP scores for the β -lactamases amino acids vs. their residue environment scores. At each amino acid position, the 19 β -lactamase variants were categorized into hydrophobic, polar, and uncharged groups.....	53
Figure 3-2 Random forest classification leave-one-out cross-validation (LOOCV) prediction array for all 4,997 β -lactamases variants (Residual Profiles data set). Collectively, these predictions yield the performance summary data in the top row. Columns correspond to the β -lactamases amino acid positions, and rows represent the 19 different types of residue replacements with wild type. A β -lactamases variant is labeled correct (green) if its experimental and predicted fitness categories are identical; otherwise, the variant is labeled incorrect (brown).....	56

Figure 3-3 Evaluating the significance of β -lactamases prediction performance. The scatter plot compares the tree regression leave-one-out cross-validation (LOOCV) predicted values obtained for β -lactamase variability values (using a local feature data set) with their experimental measurements.....	57
Figure 3-4 Random forest classification leave-one-out cross-validation (LOOCV) prediction array for all 16,043 β -lactamases variants (Residual Profiles data set) at codon level. Collectively, these predictions yield the performance summary data in the top row. Columns correspond to the β -lactamases amino acid positions, and rows represent the 61 different types of residue replacements with codon notation. A β -lactamases variant is labeled correct (green) if its experimental and predicted fitness categories are identical; otherwise, the variant is labeled incorrect (red).....	60
Figure 3-5 Learning curves. The plots reveal the degree to which performance is improved as the number of TEM variants in the training set is increased. Each point represents the average over ten runs of 10-fold CV, and the error bars indicate the standard deviation. Plots were generated by using both types of TEM variant data sets (Residual Profiles and Local Profiles feature vectors) with both random forest classification and tree regression. PPV=positive predictive value; BAR=balanced accuracy rate; MCC=Matthew's correlation coefficient; AUC= area under the curve; RMSE=root mean squared error.	63
Figure 3-6 Model performance by including different number of instances in the 10-fold cross-validation. (A) Using residual profile as inputs. (B) Using local profile as inputs. 71	
Figure 3-7 Pearson's correlation coefficients between variables and fitness class at (A) residue profile and (B) local profile.....	73
Figure 4-1 Illustration of defining a helix angle using ValgusHel-geometry. Axis 1 is the axis of the cylinder formed by residue i, i+1, ..., i+5 and axis 2 is the axis of the cylinder formed by residue i+3, i+4, ..., i+8. Residue i+4 is the center amino acid residue in residues i to i+8, and the calculated helix angle between axis 1 and axis 2 is annotated on it. The fragment containing residue i to i+8 is classified as normal α -helix (helix angle $\leq 19^\circ$), kinked α -helix (helix angle $> 30^\circ$), or curved α -helix (helix angle $19-30^\circ$).	82
Figure 4-2 Helix angles distribution by using ValgusHel method. A total of 278,010 helix angles were estimated in Our data set (Figure 2A). Figure 2B shows the zoom in results when helix angles $\geq 20^\circ$	91
Figure 4-3 The relationship between Root Mean Square Deviation (RMSD) and calculated helix angles ($^\circ$) by using ValgusHel-geometry. Error bars are standard deviations.	92
Figure 4-4 Number of α -helices categorized as kinked α -helices by using ValgusHel-geometry, HELANAL-Plus and MC-HELAN methods in a data set provided on the MC-HELAN server. The total number of α -helices was 887. The edges of the triangle represent the agreement of α -helices classified as kinked α -helices by two methods on the vertices. The circle in the middle represent the agreement of α -helices classified as kinked α -helices by the three methods.....	94
Figure 4-5 Percent of α -helices categorized as kinked α -helices by using ValgusHel-geometry, HELANAL-Plus and AHAH methods in the data set provided on AHAH	

website. The total number of α -helices was 177. The edges of the triangle represent the agreement of α -helices classified as kinked α -helices by two methods on the vertices. The circle in the middle represent the agreement of α -helices classified as kinked α -helices by the three methods.	94
Figure 4-6. Difference of amino acid frequencies (%) in (A) all residues and (B) all helices. Difference of frequency (%)=frequency (%) _i in kinked residue/helix group-frequency (%) _i in normal residue/helix group. i (1 to 20) represents 20 natural amino acids.	96
Figure 4-7 Density of sequence helix angles in each of the six sequence clusters. The calculated helix angles (°) were obtained by using ValgusHel-geometry. Cluster 1-6 were top sequence clusters obtained from Jalview based on sequence similarity. Since Jalview limits number of input entries, we randomly selected and input 15,000 out of 95,210 sequences.	97
Figure 4-8 Kinked helix cluster motif logo. (A)-(E) are from clusters 1 to 5 in Table 4-5. Numbers in x-axis are residue number within the 9-residue fragment. Bit score in y-axis is measurement of certainty. Higher bit score indicates higher certainty to observe the amino acid at the residue position in x-axis.	102
Figure 4-9 Normal helix cluster motif logo. (A)-(H) are from clusters 1 to 8 in Table 4-5. Numbers in x-axis are residue number within the 9-residue fragment. Bit score in y-axis is measurement of certainty. Higher bit score indicates higher certainty to observe the amino acid at the residue position in x-axis.	103
Figure 4-10 Predictions made by Jpred 4 among (A) normal helix clusters and (B) kinked helix clusters. Numbers in x-axis are residue number within the 9-residue fragment. Jpred 4 confidence uses the y-axis scale on the right (scale 0 to 10).	104
Figure 4-11 Boxplot of helix clusters among kink and normal helix clusters. Median, 25 th percentile (Q1), 75 th percentile (Q3), minimum (Q1 – 1.5IQR) and maximum (Q3 +1.5IQR) are shown. Outliers are shown as dots. IQR = interquartile range.	105
Figure 4-12 The relationship between a 9-residue sequence RMSD and the sum of two 4-residue sequence RMSD. In the 9-residue sequence (N=775), the center residue is removed, producing two sequence fragments, each containing 4 residues. The 9-residue sequence is aligned and compared with a model ideal helix. (A) average of helix angles at different residue locations among ideal, normal, kinked, or curve α -helices. Residue 0 is the center residue of the 9-residue sequence fragment. (B) center residue 0 was removed from the 9-residue sequence fragment, resulting in two 4-residue fragments. The sum of the RMSD from the two 4-residue fragment is the y-axis. Color dots and lines represented the classification of the 9-residue sequence. (C) center residue 0 was removed from the 9-residue sequence fragment, resulting in two 4-residue fragments. The sum of the RMSD from the two 4-residue fragment is the y-axis. Color dots and lines represented the classification of the two 4-residue fragments.	107
Figure 4-13 Percentage of simplex types in normal, curved, and kinked helices. Simplex types (0, 1, 2, 3, and 4) are adopted from Taylor et al. (2015). Figure 8(B) is a zoom in for Figure 8(A) to show details of percentage lower than 1%.	111
Figure 4-14 Agreement between topological and geometrical methods at different fragment length and N-grams. True Positive (TP) is the number of α -helices classified as	

kinked α -helices by both the ValgusHel-geometry and ValgusHel-topology. Sum is the number of total α -helices included. TP/SUM reflects the agreement between topological and geometry methods. 114

LIST OF EQUATIONS

Equation	Page
Equation 2-1	10
Equation 2-2	10
Equation 2-3	11
Equation 2-4	11
Equation 2-5	15
Equation 4-1	83
Equation 4-2	85

LIST OF ABBREVIATIONS

Amino Acid.....	AA
Area Under Curve	AUC
Artificial Neural Network	ANN
ATP-binding cassette sub-family B member 1	ABCB1
Balanced Accuracy rate	BAR
Comprehensive Mutation Profile.....	CMP
Conservative	C
Cross Validation.....	CV
DNA DataBank of Japan	DDBJ
European Molecular Biology Laboratory	EMBL
False Negative.....	FN
False Positive	FP
Feed-Forward Neural Network	FFNN
Fibrosis Transmembrane Conductance Regulator	FTCR
IgG-binding domain of protein G	GB1
Leave-one-out cross-validation.....	LOOCV
Matthew's correlation coefficient.....	MCC
National Center for Biotechnology Information.....	NCBI
Non-Conservative	NC
Positive Predictive Value.....	PPV
Protein Data Bank	PDB
Random Forest Classification	RFC
Random Forest Regression	RFR
Residue Environment Score.....	RES
Residue Score.....	RS
Sensitivity	Se
Single Nucleotide Polymorphism	SNP
Specificity	Sp
The Root Mean Square Error.....	RMSE
True Negative.....	TN
True Positive	TP

ABSTRACT

STRUCTURAL AND TOPOLOGICAL VARIATIONS IN AMINO ACIDS ENCODED BY SYNONYMOUS CODONS

Shengyuan Wang, Ph.D.

George Mason University, 2022

Dissertation Director: Dr. Iosif Vaisman

This dissertation explores the relationship between protein structural and topological properties using computational geometry approach. A representative nonredundant dataset containing 10,220 individual protein chains with known structures was created, and each amino acid residue in the set was matched to the corresponding codon. The Delaunay tessellation of all proteins in the dataset resulted in the four-body statistical potentials with both 20 letter amino acid alphabet and 61 letter codon alphabet. Compositional, geometric, and topological patterns in the codon based representation were identified and influence of the synonymous codons on protein structure was assessed. Both amino acid and codon based potentials were extensively tested for reliability and consistency and their performance in a number of applications was evaluated. Computational mutagenesis approach, where the new potentials were used in the machine learning models for predicting protein fitness and activity changes caused by mutations, demonstrated high accuracy of the predictions. In addition, a new method for

accurate identification of kinked α -helices by using both geometric and topological parameters was developed.

CHAPTER 1 : OVERVIEW

1.1 Introduction to silent mutations in proteins

Synonymous codons are translated into the same amino acids, and they were originally thought to have no effect on protein function or biocompatibility or evolutionary processes, hence the so-called "silent mutations" [1-3]. Recent evidence, however, shows that more than 50 human diseases afflicting most organ systems have been identified to be associated with synonymous mutations [1, 4], such as congenital disorder of glycosylation type 1d (mutations G55G in ALG3 gene) [5], familial adenomatous polyposis (mutations R623R H652H; R653R in APC gene) [6, 7], androgen-insensitivity syndrome (mutations S888S in AR gene), [6] etc. Chen and co-workers conducted a survey of 21,429 diseases-Single Nucleotide Polymorphism (SNPs) associations from a pool of 2,113 studies exploring human diseases and their association with non-synonymous SNPs and synonymous SNPs. Their results showed that a similar likelihood and effect size for human diseases association was observed from non-synonymous SNPs and synonymous SNPs [8]. Their findings are consistent with the estimate that 5%~10% of human genes contain at least one region where synonymous mutations could be harmful by Chamary and Hurst [4]. The understanding of synonymous codons has important clinical and technological implications. In addition to their impact on disease risks described above, it is becoming more and more evident the

SNPs have a major role in individual differences observed in how patients respond to medical treatments in terms of efficacy and effectiveness, adverse effects and disease progression. For example, synonymous mutations in transporter ATP-binding cassette sub-family B member 1 (ABCB1) have shown implications in drug resistance to chemotherapeutic agents [9]. As personalized medicine is under intense research, and pharmacogenetics focuses on variation in the enzymes and transporters that determine the disposition of small molecule drugs, due consideration should be given to synonymous mutations.

Though increasing evidence shows that synonymous mutations are associated with human diseases, the underlying molecular mechanisms in details of this association is rarely known. The results of various studies suggested a series of mechanisms for synonymous mutations in affecting the yields of active, correctly folded proteins and thus have an impact on physiological activity [10]. Recent evidence in the control of gene expression suggested that the dominance of regulation lies in translation level [11]. While it is generally accepted that codon bias contributes to translation efficiency by tuning the elongation rate of the protein synthesis process, the codon usage bias research is a potential promising perspective to understand the synonymous mutations' role in protein structure and function. Codon usage bias in many different organisms indicated that synonymous codons were under evolutionary pressure [4, 12]. The rapid advances in the knowledge in protein synthesis and folding have led to new evidence that synonymous mutations can lead to abnormal mRNA splicing associated with human diseases [6]. A recent fascinating evidence is discovered by Bartoszewski et al. that a synonymous

substitution in the cystic fibrosis transmembrane conductance regulator (CFTR), in which synonymous mutation from ATC to ATT encoding Isoleucine (Ile, I), leads to changes in the mRNA structure and subsequently a misfolded protein, which is associated with cystic fibrosis in humans. Corrected mRNA structure and higher protein levels were observed after returning synonymous mutations from ATT back to ATC [13].

In summary, analysis of synonymous codons and synonymous mutations can provide insights into a better understanding of factors influencing protein structures and potential function alterations. Previous computational studies explored the association between synonymous codons and protein secondary structures including α -helix, β -sheets, and coils, but a more comprehensive investigation on the structural relationship among amino acid residues encoded by synonymous codons has not been conducted yet. Additionally, there is a knowledge gap of how synonymous mutations influence the secondary structure and tertiary structure, and the subsequent protein function.

1.2 Introduction to Delaunay tessellation method in protein structure analysis

Delaunay tessellation is a widely used technique to assess structural and topological properties of proteins [14-31]. By using Delaunay tessellation, a protein is described as a set of points in three-dimensional space represented by (usually) α -carbon atoms in amino acid residues. Delaunay tessellation of a protein structure generates an aggregate of space-filling irregular tetrahedra called Delaunay simplices. The vertices of each simplex define objectively four nearest neighbor α -carbon atoms in amino acid residues. Throughout this dissertation, Delaunay tessellation is the most fundamental

method to assess and quantify the structural and topological properties of proteins. The detailed procedures to perform Delaunay tessellation are included in Chapter 2.

1.3 Introduction to computational mutagenesis methods

Saturation point mutagenesis is a tool for protein mutation-function analysis [21, 32-37]. In a given residue, all 19 amino acid or all 60 non-stop codon substitutions were introduced. The functional impact (e.g., fitness, activity) caused by each one of the amino acid or synonymous codon substitutions in proteins are experimentally determined. The mutagenesis scores can be used as outcomes to quantify the impact of mutations on protein fitness and activity. We estimated Delaunay potentials in mutants and wild type proteins, reflecting the structural impacts of mutations. We linked mutagenesis scores and topological variables with protein experimental fitness/activity data, aiming to predict how protein fitness/activity change when new mutations were introduced.

1.4 Overview of the work presented here

My objectives for this dissertation were to improve existed knowledge-based potential scores, identify and characterize abnormal secondary structure, and link protein structure and function by using Delaunay tessellation. Additionally, the potential scores were utilized in computational mutagenesis methods to predict the topological impacts of cancer-causing silent mutations compared to non-cancer-causing ones.

In Chapter 2, we created a dataset containing 10,220 individual protein chains with annotations of codons, amino acids, secondary structures, and α -carbon coordinates. We performed Delaunay tessellation on the proteins and calculated knowledge-based potential scores for different Delaunay simplex composition. The potential scores for

each simplex composition at the amino acid level were compared to a previous study, which had a smaller dataset, to test whether Delaunay tessellation can achieve reliable and reproducible results across different datasets. Additionally, knowledge-based potential scores were further estimated at the codon level as an extension of the previous studies. With the codon level potential scores, we further investigated whether cancer-causing silent mutations had different topological structures compare to non-cancer-causing silent mutations.

In Chapter 3, we investigated whether the previously established computational mutagenesis methodology can be generalized to another protein, β -lactamase, by taking advantage of the publicly available saturation mutagenesis fitness data of β -lactamase at both amino acid and codon level. We investigate β -lactamase at the amino acid level in details, and compare performance at the amino acid level vs. codon level. We also applied the methodology to more proteins with mutagenesis activity data. Additionally, we tested whether model built for one protein can be used to predict mutagenesis score for another protein in the same activity category. The methods were also applied to pairwise amino acid substitutions.

In Chapter 4, we developed a new method, “*ValgusHel*”, to classify α -helices into kinked, curved, and normal ones based on their geometry and topological properties. We validated the method by comparing the classification agreement with other available methods. This new method may serve as an additional method for researchers to identify kinked α -helices in the function-structure analysis.

CHAPTER 2 : FOUR-BODY POTENTIAL ANALYSIS USING DELAUNAY TESSELLATION AND THEIR APPLICATION IN CANCER-CAUSING SILENT MUTATION TOPLOGICAL ANALYSIS

2.1 Overview

In this chapter we created a dataset containing 10,220 protein chains with codon, amino acid, secondary structure annotations, and α -carbon coordinates. We performed Delaunay tessellation on proteins and obtained Delaunay simplices represented by both amino acid and codon compositions. We estimated knowledge-based potential for each simplex composition. The potential estimation for the same simplex composition at amino acid level in this study was compared to a previous study, which had a smaller dataset and analysis performed at only amino acid level. The result agreement would show whether Delaunay tessellation was able to produce reliable and consistent results across different datasets. We also explored factors impacting the potentials and improved the methods. The association between protein subgroups (enzymes, hydrolases) and potential estimations were explored. Knowledge-based potentials disturbance caused by cancer-causing vs. non-cancer-causing silent mutations were compared to reveal the impact of silent mutations on protein structure and function by using the computational mutagenesis methods.

2.2 Materials and methods

2.2.1 Protein sequences and structures database construction

A set of 10,220 proteins were obtained from the PDB using the PISCES web server by applying the following filtering criteria: (1) resolution ≤ 2.0 angstrom (\AA); (2)

crystallographic residual factor (or reliability factor or the R-value) ≤ 0.25 ; (3) maximum pairwise sequence identity $\leq 30\%$; (4) protein chain length 100-1,000 amino acid residues; (5) X-ray as experimental method. This set is abbreviated as *10220culled*.

DNA sequences encoding protein in the 10220culled were collected by using BLAST+ [38], a sequence similarity searching tool downloaded from the National Center for Biotechnology Information (NCBI) website. All nucleotide sequence data, with entries from all traditional divisions of GenBank (the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences), European Molecular Biology Laboratory (EMBL), and DNA DataBank of Japan (DDBJ) were downloaded from NCBI blast database. Only the accession number, the start and end position of the matched nucleotide sequence in the database was returned by TBLASTN, but not the sequence itself. Biopython was used as the extraction tool for DNA [39]. All protein sequences read from the 5' end (the phosphoryl end) to the 3' end (the hydroxyl end) in the collected dataset. Nucleotide sequences containing character 'Y' (indicate 'Cytosine (C)' or 'Thymine (T)'), 'N' (indicate 'any'), or 'R' (indicate 'Adenine (A)' or 'Guanine (G)') were removed due to uncertainty from the dataset. No or very few missing residues is required in Delaunay tessellations. Therefore, we found the longest common substring (DNA match protein sequence) and kept those having ≥ 20 residues. Codons for all residues were also annotated in proteins in the 10220culled. Most proteins have many tblastn results, we selected those with protein and DNA in the same organism and lowest e-values. Individual protein chain secondary structures in the 10220culled were collected from DSSP. The 10220culled is a dataset containing 10,220 codon and amino acid

labeled individual protein chain, secondary structure annotations, and α -carbon coordinates.

2.2.2 Secondary structure preference of synonymous codons

After creating a dataset which contains 10,220 individual protein chains with different level annotations (codon, α -carbon, secondary structure), we examined whether synonymous codons had different preference to form secondary structures.

Methionine and Tryptophan were excluded because they don't have synonymous codons. A total of 1,836,434 residues with DSSP annotations were collected. Secondary structures are defined by DSSP based on hydrogen bonding patterns, including H (4-turn helix or alpha-helix), G (3-turn helix or 3-10 helix), I (5-turn helix or pi-helix), B (isolated beta-bridge), E (extended strand in beta-sheet conformation), S (bend), T (turn) and C (coil). Secondary structures are classified in two ways: (1) simple 3 classes, which contains alpha-helix (H, G, I), beta-strands (B, E) and coils (S, T, C); (2) all 8 classes, H, G, I, B, E, S, T, C. In each way of classes, we got the observed number of each codon in different secondary structures. The expected number of each codon in different secondary structure were then calculated by using number of synonymous codons in the corresponding amino acid multiplied by the number of secondary structures in corresponding amino acid. For example, there are 33,687 of A₁ (A for Alanine), 158,788 of Alanine, and 73,430 α -helix formed by Alanine in the total dataset. The expected number of A₁ with α -helix is $33687/158788 \times 73,430 = 15578$. After calculating all expected number of each codon in different secondary structures, we performed a chi-square test to determine whether there was a statistically significant difference between

the expected number and the observed number for each of codon in different secondary structure.

2.2.3 Knowledge-based potential estimation

Pyhull [40, 41] was used to perform Delaunay tessellation [14, 27] on protein structure analysis based on the coordinates of each α -carbon atom, the first carbon atom that attaches to a functional group. The protein is described as a set of points in three-dimensional space represented by α -carbon atoms in amino acid residues. Delaunay tessellation of a protein structure generates an aggregate of space-filling irregular tetrahedra called Delaunay simplices. The vertices of each simplex define objectively four nearest neighbor α -carbon atoms in amino acid residues. Computation of Delaunay tessellation was performed on proteins in 10220culled. The Delaunay simplices were obtained with a 12 Å cutoff on all edges. No simplex with an edge length greater than 12 Å was counted. Information of each Delaunay simplex, including the combination of four amino acids positioned on four vertices, the residue number of each amino acid in the protein sequence, the distance of each two vertices by residue number, the length of six edges, the volume, the tetrahedrality, and secondary structure of each residue, were collected for further use. The frequency of each kind of simplex based on amino acid or codon residue compositions was calculated. The distribution of tetrahedrality and volume for each category was also calculated.

Simplices classification was done based on the way the peptide chain threads through them that was introduced in previous studies [14, 27]. There are five types of simplices. In type 0, none of the four residues in the simplex are consecutive. In type 1,

there is one and only one pair of amino acid residues is consecutive. In type 2, two pairs of residues are consecutive and these two pairs are separated in the sequence. In type 3, three residues are consecutive and the fourth is a distant one. In type 4, all four residues in the simplex are consecutive in the protein primary sequence.

Volume and tetrahedrality are two geometrical parameters being used to describe the tetrahedra. Tetrahedrality is a quantitative measure of the degree of distortion of the Delaunay simplices from the ideal tetrahedron, the equation to calculate the tetrahedrality shows below:

Equation 2-1

$$T = \frac{\sum_{i>j} (l_i - l_j)^2}{15\bar{l}^2},$$

where l_i is the length of the i -th edge, and \bar{l} is the mean length of the edges of the given simplex.

Log-likelihood is a knowledge-based potential based on Boltzmann's principle: frequently observed states corresponding to low energy states of the system [42]. Log-likelihood was calculated to represent the preference of a simplex exists in naturally occurring proteins. Log-likelihood of simplices from different categories was defined as:

Equation 2-2

$$q_{ijkl} = \log_{10} \frac{f_{ijkl}}{p_{ijkl}},$$

where f_{ijkl} is the observed frequency of simplices with amino acid or codon types i, j, k and l at their vertices; p_{ijkl} was defined as:

Equation 2-3

$$p_{ijkl} = ca_i a_j a_k a_l,$$

where a_i , a_j , a_k , and a_l are the observed frequencies of the individual amino acid or codon types; and combinatorial factor c is defined as:

Equation 2-4

$$c = \frac{4!}{\prod_i^n (t_i!)},$$

where n is the number of distinct residue types in a quadruplet and t_i is the number of amino acids or codons of type i . All simplices identified in this study were ranked by the log-likelihood.

2.2.4 Comparison of amino acid level knowledge-based potential results with a previous study

This dissertation is built on the study by Taylor et al. [43]. In their work, x-ray structures of 1,417 non-homologous protein chains were obtained from the PISCES web server. There were no missing Carbon-alpha coordinates, resolution $\leq 2.2 \text{ \AA}$, crystallographic R-factor ≤ 0.23 , and maximum pairwise sequence identity $\leq 30\%$ in their dataset. Because PDB contains more entries since then (122,021 x-ray entries in 2017 vs. 34,249 in 2006), this dissertation includes a larger dataset, the 10220culled. The larger sample size in this dissertation may reveal new knowledge of knowledge-based potentials by comparing results in these two studies. Additionally, the knowledge-based potentials were also calculated at codon level, an extension of only at amino acid level in Taylor et al [28].

We explored the effect of restricting simplex edge length, simplex compositions with low number of simplices, and sequence bias on the correlation between potentials in this study and Taylor et al. Fast Low-Probability Subsequences (fLPS) locates sequences with high bias, where certain abnormal high frequencies of amino acids may make potential estimation less accurate.

2.2.5 Examining potentials distribution difference between synonymous codons

After calculating potential scores for all simplex composition, we would like to know if synonymous codons form different potential score patterns. We looked at simplices with the codon pattern of

$$P_{x_i} = \{x_i y_i z_i t_i\}$$

where P is the potential, x_i is the i -th synonymous codon of amino acid X and y_i, z_i, t_i are i -th synonymous codons of amino acid Y, Z, T , and i (2, 3, 4, 6) is the number of synonymous codons depending on the specific amino acid. X is different from Y, Z, T while Y, Z, T may be the same amino acid. We performed z-tests to examine whether the mean potentials are the same between the potential distributions on any two of x_i in $x_i y_i z_i t_i$. For example, Alanine has 4 synonymous codons, and we performed t-test on the 6 pairs among $A_1 y_i z_i t_i, A_2 y_i z_i t_i, A_3 y_i z_i t_i$, and $A_4 y_i z_i t_i$. To serve as a control, we additionally compared the potential means of $x_i y_i z_i t_i$ and $m_i y_i z_i t_i$, where X and M are different amino acids.

We also test double synonymous codons distribution. We performed similar analysis on

$$P_{x_i x_i} = \{x_i x_i y_i z_i\}$$

where P is the potential, x_i is the i -th synonymous codon of amino acid X and y_i, z_i are i -th synonymous codons of amino acid Y, Z , and i (2, 3, 4, 6) is the number of synonymous codons depending on its specific amino acid. X is different from Y, Z while Y, Z may be the same amino acid. Then we performed z-tests to examine whether the mean potentials are the same between the potential distributions on any two of x_i in $x_i x_i y_i z_i$. For example, Alanine has 4 synonymous codons, and we performed t-test on the 6 pairs among $A_1 A_1 y_i z_i, A_2 A_2 y_i z_i, A_3 A_3 y_i z_i$, and $A_4 A_4 y_i z_i$. To serve as a control, we additionally compared the potential means of $x_i x_i y_i z_i$ and $m_i m_i y_i z_i$, where X and M are different amino acids.

2.2.6 Investigating protein structure and function change by silent mutations through knowledge-based potentials

2.2.6.1 Dataset creation

For those silent mutations causing cancers, we aim to study whether we can detect the protein structure changes in terms of Delaunay tessellation knowledge-based potentials by using computational mutagenesis methods. We selected 15 genes whose mutations are known to cause human cancers (Table 2-1). We downloaded the protein structure data from PDB and secondary structure from DSSP. We searched each protein in PDB 1D Coordinate Service and obtained the corresponding position on genome and the codon for each amino acid residue. PDB provides genome position information for most amino acids in the dataset. We kept a few amino acids without genome position in the Delaunay tessellation to keep protein structure integrity, but they were not used in the computational mutagenesis part.

Table 2-1 Proteins studied for silent mutations causing cancers

Gene	PDB ID	Method	Resolution (Å)	Number of residues	Residue range	PDB missing residues
BRAF	5ITA.A	X-ray	1.95	250	448-718	598-615, 628-630
BRCA1	1JM7.A	NMR		103	1-103	
	4IGK.A	X-ray	1.75	252	1646-1859	1816-1818
CDK4	2W96.B	X-ray	2.30	264	4-295	42-47, 239-260
CHEK2	1GXC.A	X-ray	2.70	116	92-207	-
	2CN5.A	X-ray	2.25	281	210-504	255-268
EGFR	3QWQ.A	X-ray	2.75	613	2-614	-
	5UG9.A	X-ray	1.33	280	702-985	749-752
EP300	3BIY.A	X-ray	1.70	317	1287-1664	1520-1580
	3T92.A	X-ray	1.50	90	7-96	-
	5BT3.A	X-ray	1.05	113	1049-1161	-
FHIT	1FIT.A	X-ray	1.85	124	2-147	82, 107-126, 135
FLT3	1RJB.A	X-ray	2.10	133	572-710	649-654
HRAS	121P.A	X-ray	1.54	166	1-166	
MLH1	4P7A.A	X-ray	2.30	303	3-336	86-97, 301-319
PDGFRA	5GRN.A	X-ray	1.77	112	584-695	-
PRNP	4KML.A	X-ray	1.50	109	117-225	-
PTEN	1D5R.A	X-ray	2.10	268	14-281	-
RET	2IVS.A	X-ray	2.00	284	713-1012	828-843
TP53	1TUP.A	X-ray	2.20	196	94-289	-

2.2.6.2 Four-body statistical potentials and computational mutagenesis

We tessellated each protein and obtained the knowledge-based potentials by following the methods described in 2.2.2 *Knowledge-based potential estimation*. By following the definitions proposed by Masso et al. (2019) [32], we added up the potential scores of all the Delaunay simplices in the proteins. For each residue, a residue environment score (RES) is the sum of potential scores from all simplices this given residue belongs to. The number of simplices a residue belongs to varies. We ordered the collective RES scores for all amino acid positions in each protein to create a potential profile as described by previous studies [32]. Residual score quantifies the relative change in sequence-structure compatibility in mutations

We calculated residue scores (RS) and residue profiles by following equations below:

Equation 2-5

$$RS = p_m - p_n$$

where RS is residue score, p is potential, p_m is potential of mutants, and p_n is potential of native protein.

2.2.6.3 Protein variant feature vectors

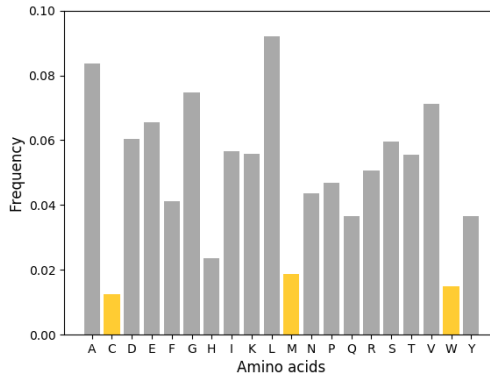
We collected those silent mutations that may cause human cancer from SynMICdb [44]. SynMICdb provides locations, nucleotide of both wild type and mutation, and a SynMICdb score to measure the likelihood of a silent mutation that have a function impact [44]. We matched local profile with SynMICdb score with same wild type and mutation of each residue, resulting in a total of 208 matches. Each of the 208

matches have residue information, residual score, and SynMICdb score. We examined the correlation between residue score and SynMICdb score. We created a control dataset including 413 residues whose synonymous codon mutations likely do not cause cancer. For example, if A1 to A2 mutation in a gene was cancer-causing in the dataset, we deduced that A1 to A3 and A1 to A4 were possibly not cancer-causing and they were included as controls. We compared the residual score distribution in these two datasets. We also created 5 balanced combine datasets, each contains 208 cancer-causing and 208 non-cancer-causing local profile. We use random forest and naïve Bayes to make 5-fold cross-validation and would like to see by using local profile, if we could differentiate synonymous mutation that cause or not cause cancer.

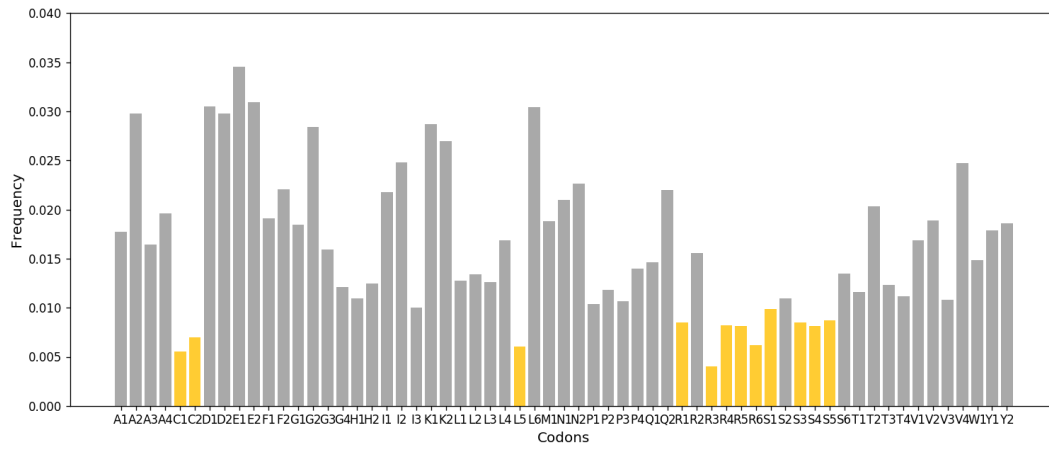
2.3 Results and Discussions

2.3.1 Knowledge-based potential results

Frequencies of amino acids in the 10220culled dataset was shown in Figure 2-1. While Alanine (A), Leucine (L) and Glycine (G) had the highest frequencies, Cysteine (C), Methionine (M) and Tryptophan (W) had the lowest ones.



(A)



(B)

Figure 2-1 Frequency of amino acids in 10220culled dataset.

Distribution's pattern of volume and tetrahedrality across the five types of simplices were presented in Figure 2-2. The distribution curve of the type 4 simplices had the sharpest and narrow peak, *i.e.* the lowest volume and lowest distortion of tetrahedrality but the highest density. This result was expected because it is common in most proteins that four consecutive residues are nearest neighbors to each other. Different

types of simplices showed different distributions of volume and tetrahedrality, and a possible explanation was that they were correlated with the conventional secondary structure assignment of contained residues [45].

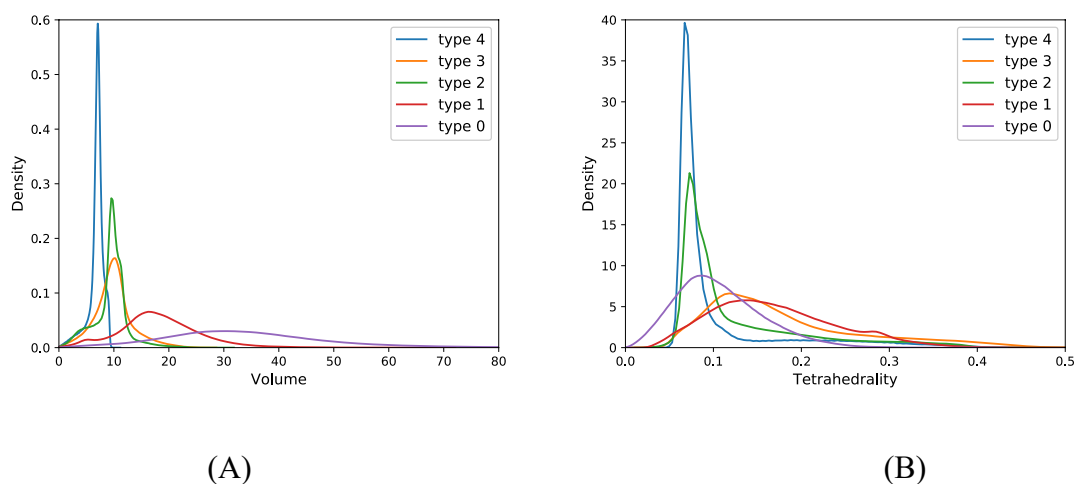


Figure 2-2 The distribution of volume and tetrahedrality of simplices in different simplex types.

Knowledge-based potential scores, or log-likelihood ratios, of each simplex composition at amino acid level and at codon level was presented in Figure 2-3. At amino acid level, we observed 8,854 simplex compositions, only one theoretical simplex composition (WWWW, Tryptophan, Trp) was not found (Figure 2-3(A)). This is not

surprising considering the frequency of natural occurring W is low and its side chain is long. The 10 simplices with the highest potentials were CCCC, CCCH, CCCW, CCCM, CCCG, CCCF, CCCS, CCCY, CCCN, and CCCR. Disulfide bridges could be a reason for high prevalence of Cysteine (Cys, C). At codon level, a total of 600,995 simplex compositions were observed, contrasting with a theoretically maximum number of 635,276. We observed similar trend in knowledge-based potential distributions at the amino acid level and codon level.

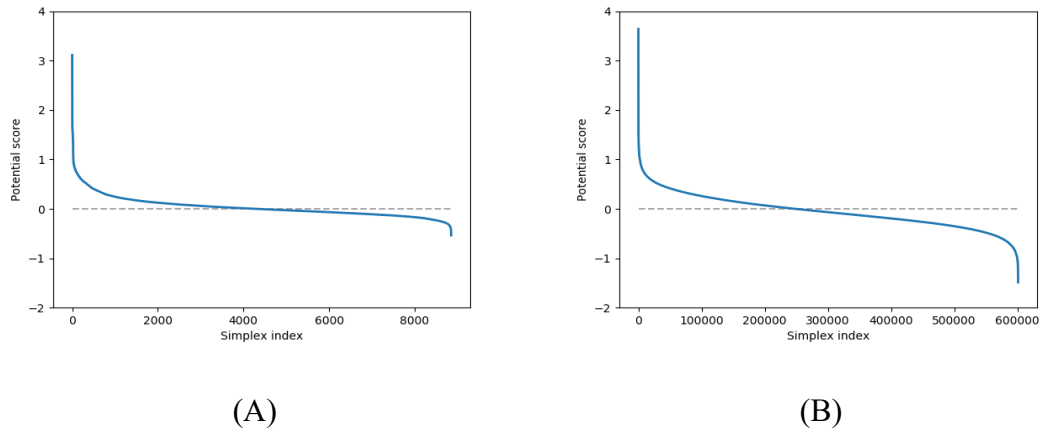
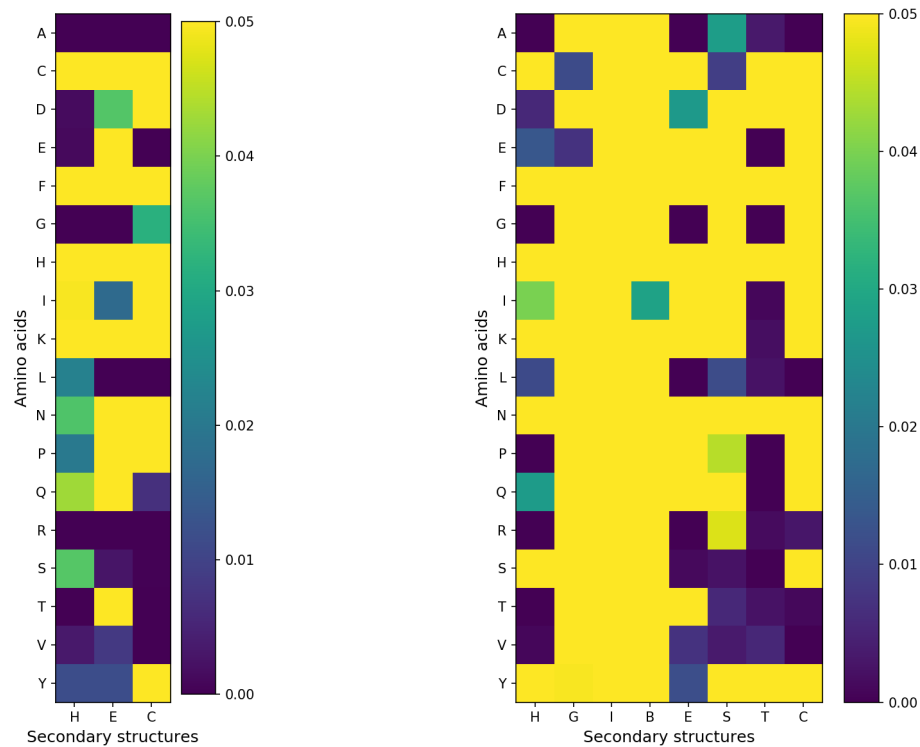


Figure 2-3 Log-likelihood ratio of Delaunay simplices at (A) amino acid level and (B) codon level. Simplex compositions were ranked from high potentials to low potentials on X-axis.

2.3.2 Secondary structure preference of synonymous codons results

When structures are categorized into 3 simple classes (H, E, and C), we found that 31 out of 54 AA-structure showed statistically significant different observed vs. expected synonymous codon number distributions, with p-values <0.05 in the Chi-square tests (Figure 2-4A). When structures are categorized into 8 classes, 47 out of 144 AA-structure showed statistically significant different observed vs. expected synonymous codon number distributions (Figure 2-4B). The results suggested that some synonymous codons had different preference to form secondary structures. In Figure 2-4B, synonymous codons have less degree of preference in G, I, B, S and C but higher degree of preferences in Alanine and Threonine. I (pi-helix) may need to be excluded from the results since the number of codons is very low (average is only 6 in different codons compared to 4,446 in other secondary structures).



(A). 3-classes secondary structure

(B). 8-classes secondary structure

Figure 2-4 Chi-square test results of the observed and expected number of synonymous codons in 18 amino acids in different secondary structures. P-values ≥ 0.05 are highlighted in yellow.

2.3.3 Amino acid level potential estimations comparing with a previous study

Out of 8,855 possible types of simplices, 8,854 (99.99%) were observed in 10220culled in this study and 8,851 (99.95%) in Taylor's dataset (2006) [28]. The mean counting of each simplex composition is 909 in 10220culled vs. 160 in Taylor's. The overall potential of all simplex compositions in 10220culled is similar to Taylors'

(0.0332 vs. 0.0258). We found that the knowledge-based potentials estimated for the same simplex composition in 10220culled were highly correlated (slope of 1.00, R-square of 0.90) with those in Taylor's (2016) (Figure 2-5 (A)). We explored why the correlation R-square was only 0.90.

Table 2-2 lists the potentials of top 10 simplex compositions deviated most from the regression line in Figure 2-5 (A). There was a greater number of simplices in the example simplex compositions in the 10220culled compared to that in Taylor's study. The density plots in Figure 2-5 (B) confirms this observation. Restricting simplex compositions with a low number increased the correlation R-square between the potentials in 10220culled and Taylor's study [28]. The potential estimations may not be accurate when there are only a few simplices in the simplex composition. By removing simplex compositions with cutoffs (1 to 1000), the correlation R-squares increased continuously at counting of simplices cutoffs 1-600 and plateaued at cutoffs 600-1000. However, the higher the cutoffs, the fewer simplices left for each simplex composition. The larger sample size in 10220culled may bring more accurate estimation of potentials for simplices, which is an improvement compared to Taylor's study.

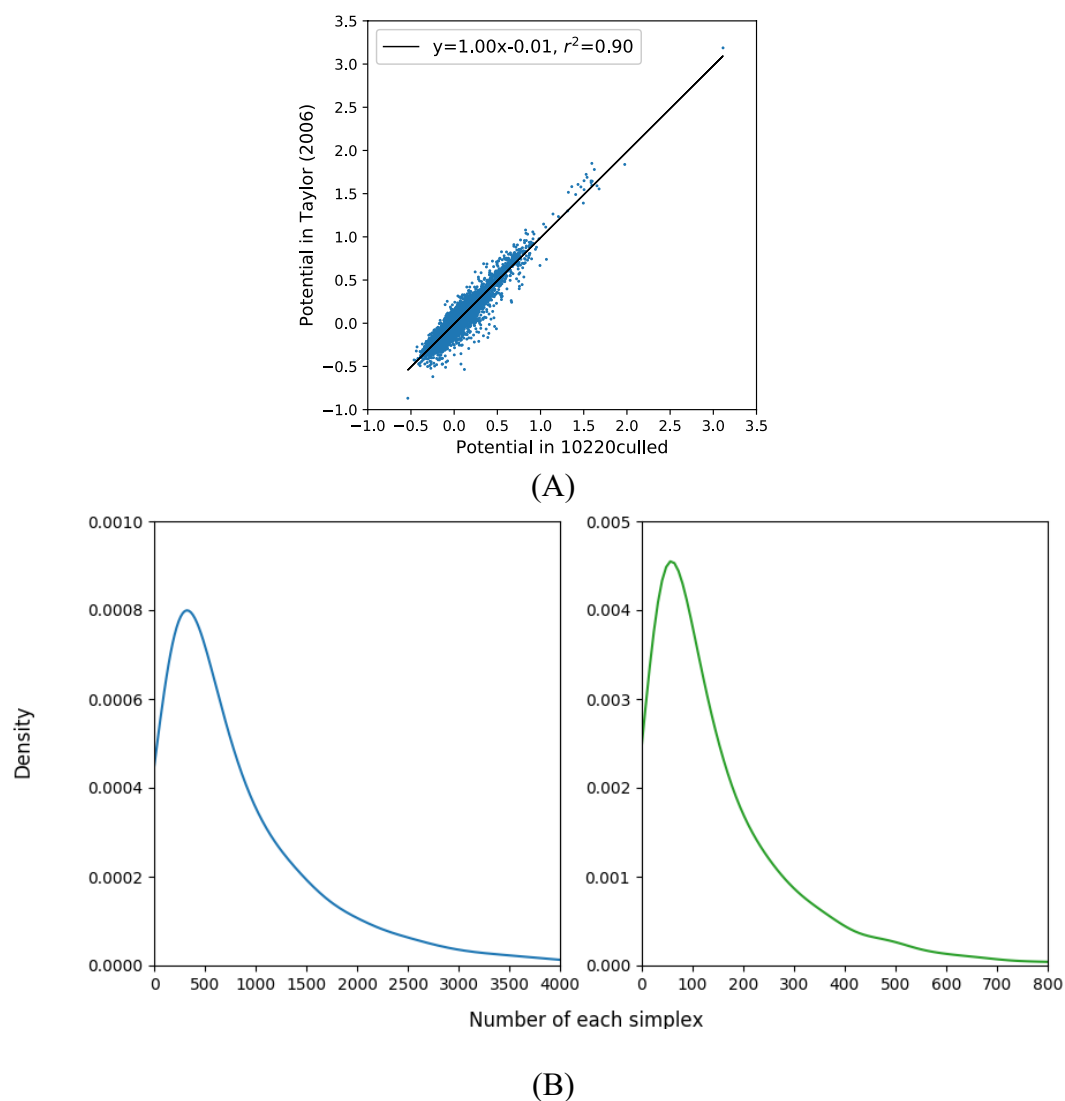


Figure 2-5 Knowledge-based potential results comparison from 10220culled dataset and Taylor's (2006) dataset. (A) Scatter plot showing the correlation of estimated potentials for the same simplex composition between the two studies. A regression line was shown with estimated a correlation slope with an R-square. (B) Density plots of each simplex

composition. Number of each simplex is the number of simplices in each simplex composition represented by amino acid residues.

Table 2-2 Knowledge-based potential score examples in 10220culled dataset in this dissertation and in the dataset in a previous study

Dataset	10220culled in this dissertation		Taylor (2006)	
Simplex	Knowledge-based potential score	N	Knowledge-based potential score	N
CNNW	0.119	45	-0.535	2
CIKW	0.079	136	-0.506	7
MQWW	0.390	36	-0.164	2
CDMW	0.079	49	-0.470	3
DMMM	0.489	40	-0.061	3
MMTW	0.402	71	-0.115	5
MMMT	0.468	35	-0.034	3
DHWW	0.285	58	-0.210	3
CIWW	0.339	33	-0.144	2
KMMQ	0.169	103	-0.290	9

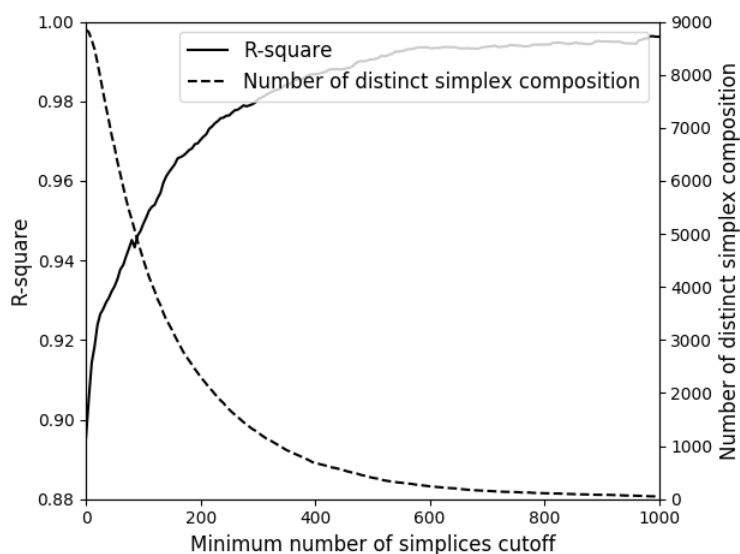


Figure 2-6 Distribution of number of simplices in 10220culled and Taylor's dataset. R-square in Y-axis (left) is the correlation R-square between potentials estimated in 10220culled and potentials estimated in Taylor's study for the same simplex composition. Number of scatter potentials in Y-axis (right) is the number of unique simplex compositions. Least number of simplices in X-axis is the cutoff to restrict simplex compositions with a number of simplices no less than the cutoffs.

Restricting simplex sequence bias also impacted the correlation R-square of potential estimations in 10220culled and Taylor's study (Figure 2-7). We found that by removing simplices with fast Low Probability Subsequences (fLPS) $p\text{-value} < 10^{-16}$, the correlation R-square increased. When fLPS was fixed, restricting simplex compositions to have at least 20, 40, 60, 80, or 100 simplices brought monotonically higher R-square values compared to no restriction. Figure 2-8 shows the impact of simplex edge length

cutoffs and restricting simplex compositions on the correlation R-square. We found that the combination of higher simplex edge length cutoffs and higher minimal number of simplices cutoffs resulted in higher correlation R-square.

Overall, we found that the knowledge-based potential estimations for the same simplex compositions in 10220culled dataset in this dissertation highly agreed with those in Taylor's study, suggesting the method is reliable in achieving stable results across different datasets. Additionally, the larger sample size in 10220culled may improve the potential estimations for some simplex compositions that only have few observations in Taylor's study. Restricting fLPS p-values, simplex compositions with low numbers, and simplex edge lengths can increase the correlation R-squares between potentials for the same simplex compositions in 10220culled and the ones in Taylor's study.

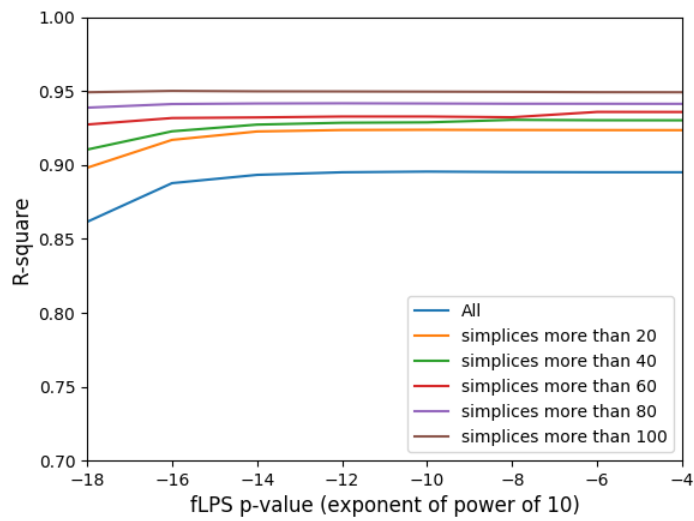


Figure 2-7 Impact of fLPS p-value and least number of simplices on the correlation R-square between potentials for the same simplex composition in 10220culled and Taylor's study. Y-axis is fLPS p-value in exponent of power of 10. Different color shows the effect of removing simplex compositions containing simplices less than N (0, 20, 40, 60, 80, or 100).

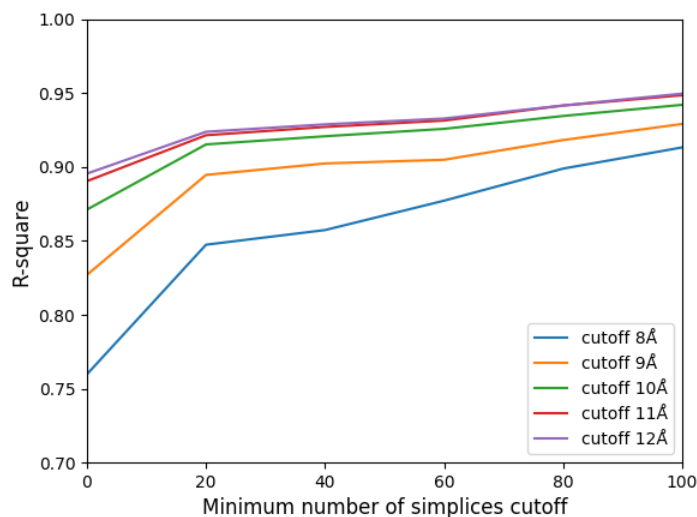


Figure 2-8 Impact of minimum number of simplices cutoff and Delaunay simplex edge length cutoffs on the correlation R-square between potentials for the same simplex composition in 10220culled and Taylor's study. Cutoff unit is angstrom (Å). Minimum number of simplices in X-axis is the cutoff to restrict simplex compositions with a number of simplices no less than the cutoffs.

2.3.4 Learning curves of potential estimations in the *10220culled* dataset

We demonstrated the effect of the training set size in correlation R-square between potential estimations in 10220culled and the ones in the training set (Figure 2-9). At the amino acid level, we found that a relatively small training set (approximately 1,500 proteins) yields a correlation R-square of 0.9. When the sample size further increased to greater than 3000, the correlation R-square plateaued out. At the codon level, however, the correlation R-square increased continuously when training set size increased from 1000 to 8000 proteins, reaching an R-square of around 0.8 in the end. A

much larger training set at the codon level was needed to yield the same correlation R-square of 0.9 at the amino acid level. This result suggested that our dataset of 10,220 proteins was not enough in sample size to analyze knowledge-based potentials at the codon level. Additionally, there are 8,855 theoretical simplex compositions at amino acid but 635,276 at the codon level. We obtained a total of 8,049,726 simplices with a mean number of simplices in each simplex composition is 909 at amino acid level but only 13 at the codon level. There were many simplex compositions had very few simplices at the codon level. The potential estimations at the codon level may be less accurate than those at amino acid level. It's an important next step in the future when more protein structural data are available in PDB.

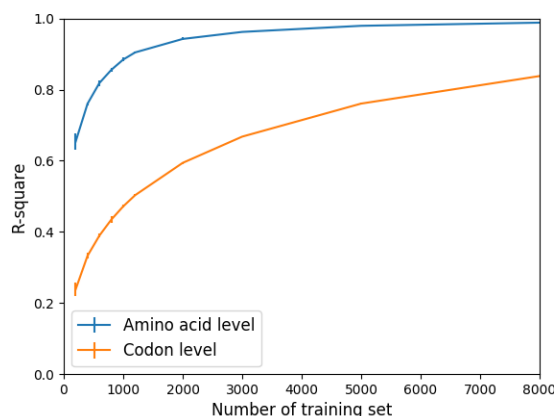


Figure 2-9 Learning curve of number of training set. R-square is the correlation R-square between mean knowledge-based potentials in 10220culled dataset and the ones in the training set.

2.3.5 Potentials distribution difference between synonymous codons

Figure 2-10 showed the z-tests results. A total of 79 out of 87 (90.8%) pairs had p-values < 0.05 , suggesting there was statistically significant difference of mean potentials between two potential distributions. The results may indicate synonymous codons were associated with different potential distribution. The comparison results of the potential means of the any two simplex compositions are shown in Figure 2-11, where 1659 out of 1743 (95.2%) comparisons with p-value < 0.05 . The results 90.8% and 95.2% were similar in Figure 2-10, suggesting that the impacts of synonymous codons may in the similar degree of those between codons of different amino acids.

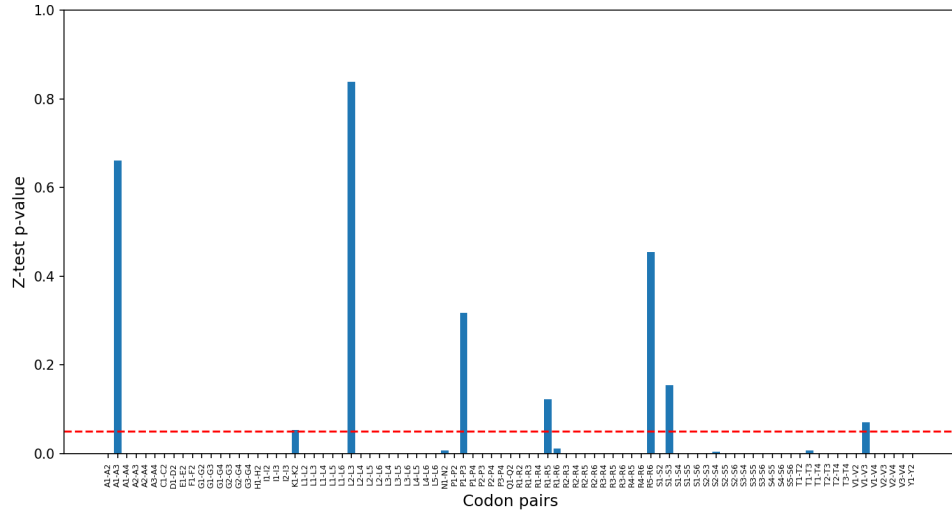


Figure 2-10 Z test results of the mean potentials of synonymous codon pairs of $x_i y_i z_i t_i$ and $x_i y_i z_i t_i$, where x_i is the i -th synonymous codon of amino acid X and y_i, z_i, t_i are i th synonymous codons of amino acid Y, Z, T, and i (2, 3, 4, 6) is the number of synonymous codons depending on its specific amino acid. X is different from Y, Z, T while Y, Z, T may be the same amino acid. P-values of 0.05 is highlighted in the red dashed line.

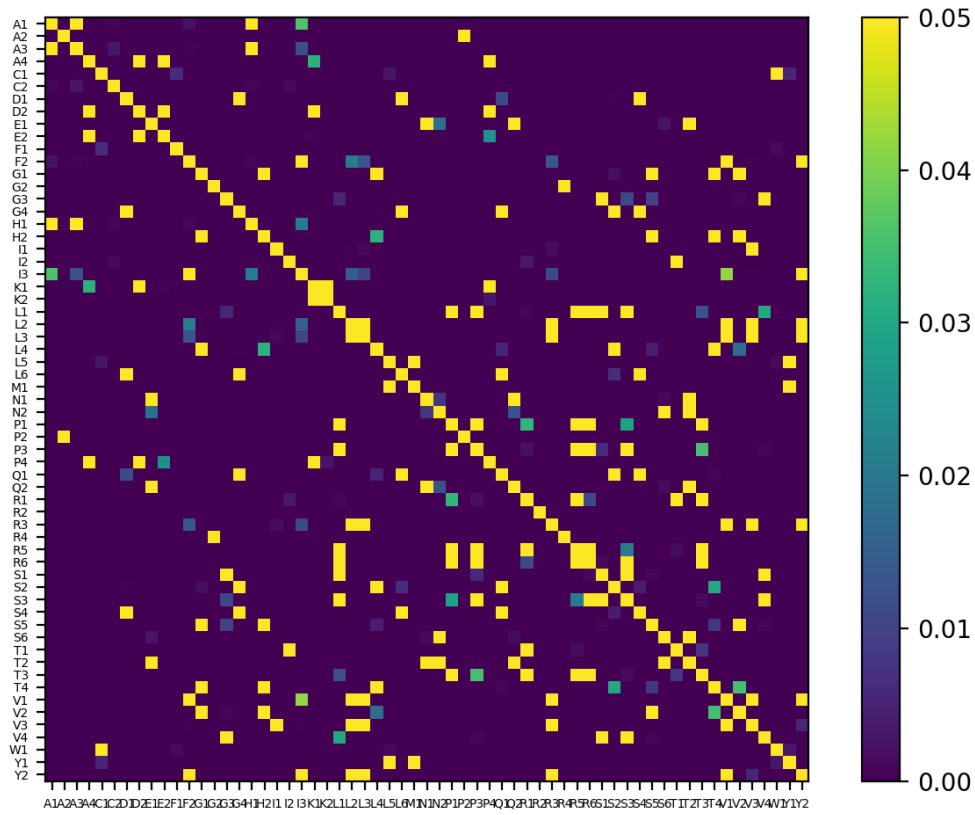
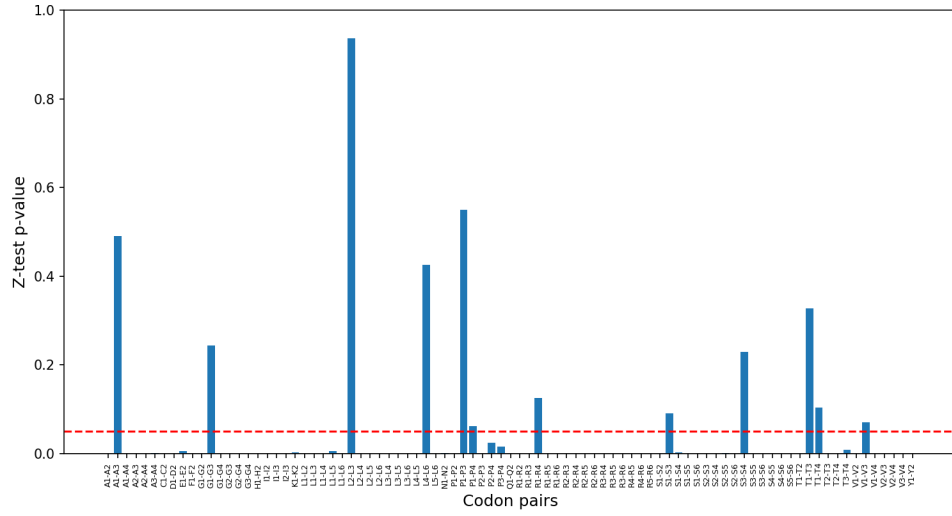
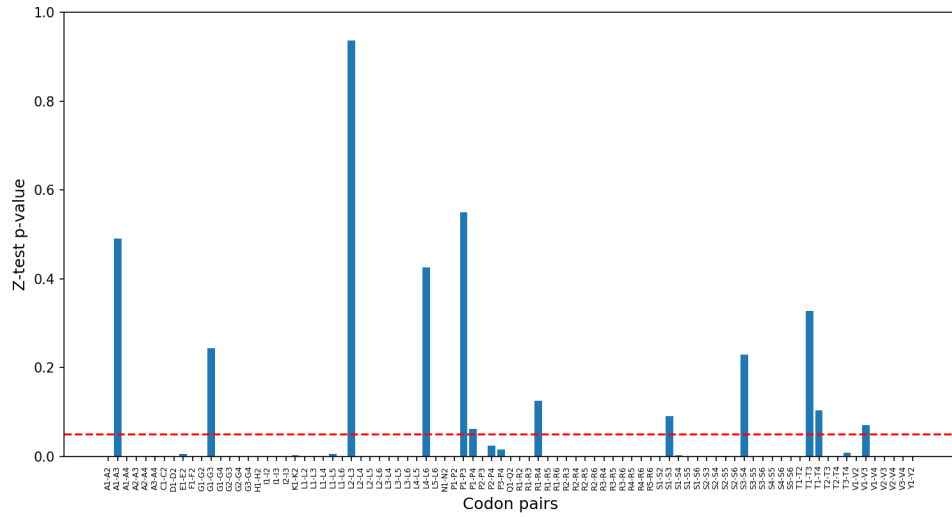


Figure 2-11 Z test results of the mean potentials of synonymous codon pairs of $x_i y_i z_i t_i$ and $m_i y_i z_i t_i$, where x_i and m_i is the i th synonymous codon of amino acid X and M respectively, and y_i, z_i, t_i are i th synonymous codons of amino acid Y, Z, T, and i (2, 3, 4, 6) is the number of synonymous codons depending on its specific amino acid. X/M is different from Y, Z, T while Y, Z, T may be the same amino acid. X and M can be different or same amino acid. P-values of 0.05 is highlighted in the red dashed line.

When comparing double codon simplex in synonymous codon pairs, we presented the z-test results in



. A total of 75 out of 87 (86.2%) synonymous pairs of $x_i x_i y_i z_i$ and $x_i x_i y_i z_i$ had different mean potentials (z-test, p-value < 0.05) in



, and 1581 out of 1643 (96.2%) all pairs with $x_i x_i y_i z_i$ and $m_i m_i y_i z_i$ had different mean potentials (z-test, p-value < 0.05) in Figure 2-13.

Comparing single and double codon simplex in synonymous codon pairs, we found that results are relatively consistent: among those 8 synonymous codon pairs which had similar potential means (p-value >0.05), 5 of them (A₁-A₃, L₂-L₃, P₁-P₃, S₁-S₃, V₁-V₃) had the same results in double codon pairs.

Triple codon simplices are also tested, however, due to number of counting in codons, the results may not accurate.

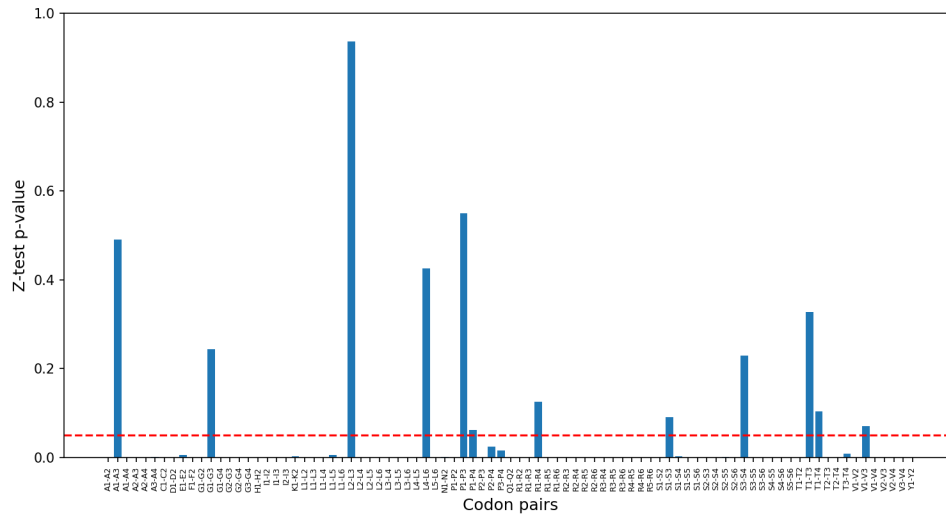


Figure 2-12 Z test results of the mean potentials of synonymous codon pairs of $x_i x_i y_i z_i$ and $x_i x_i y_i z_i$, where x_i is the i -th synonymous codon of amino acid X and y_i, z_i are i -th synonymous codons of amino acid Y, Z, and i (2, 3, 4, 6) is the number of synonymous codons depending on its specific amino acid. X is different from Y, Z while Y, Z may be the same amino acid. X and M can are different amino acids. P-values of 0.05 is highlighted in the red dashed line.

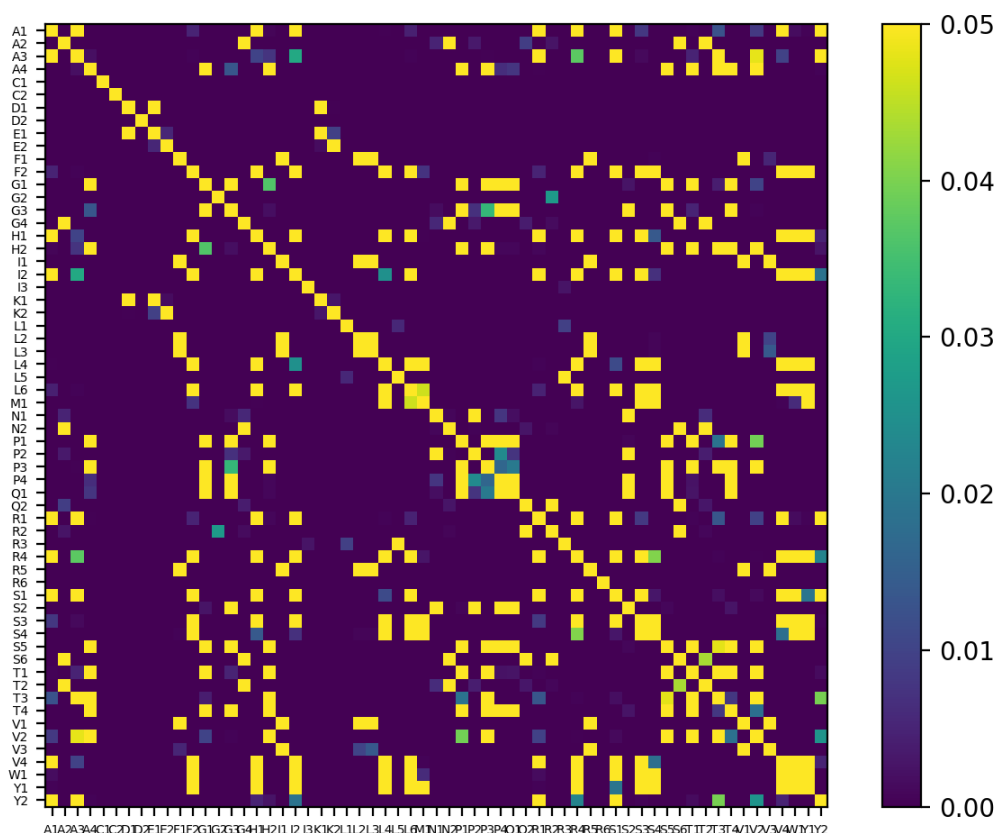
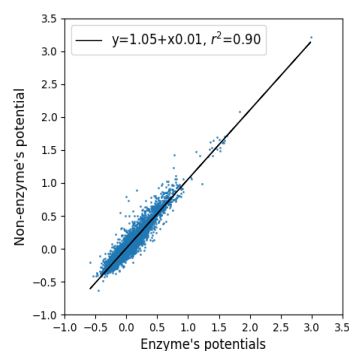


Figure 2-13 Z test results of the mean potentials of synonymous codon pairs of $x_i x_i y_i z_i$ and $m_i m_i y_i z_i$, where x_i and m_i is the i -th synonymous codon of amino acid X and M, respectively, and y_i, z_i are i -th synonymous codons of amino acid Y, Z, and i (2, 3, 4, 6) is the number of synonymous codons depending on its specific amino acid. X is different from Y, Z while Y, Z may be the same amino acid. P-values of 0.05 is highlighted in the red dashed line.

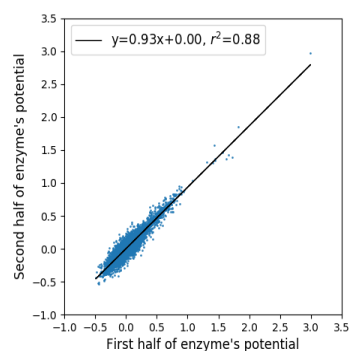
2.3.6 Potentials comparison among protein subgroups

The knowledge-based potentials of the 3,667 enzyme proteins and 6,553 non-enzyme proteins are shown in Figure 2-14. Their potentials showed a strong positive correlation with a R-square of 0.90 in Figure 2-14(A) at the amino acid level, and a

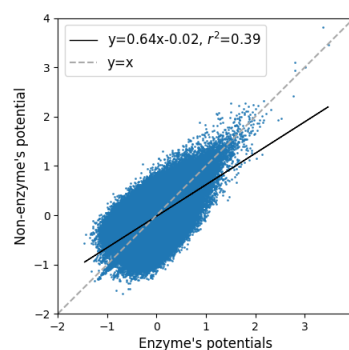
positive correlation with a much lower R-square of 0.35 in Figure 2-14(C) at the codon level. We observed that potentials in the first half and second half enzyme proteins had a strong correlation of 0.93 with a R-square of 0.88 in Figure 2-14(B), indicating potentials within the enzyme class were similar. By applying the same methods, we found that the potentials of the 1,843 hydrolase proteins and 8,377 non-hydrolase proteins showed a similar trend in Figure 2-15. The results suggested that protein subgroups (enzymes vs. non-enzymes, hydrolases vs. non-hydrolases) may have minimal influence on knowledge-based potential estimations at the amino acid level. The protein subgroups, however, may have an impact on the potential estimations at the codon levels, or the results are simply caused by the inaccurate estimation of potentials at codon levels. We cannot draw conclusions at this time.



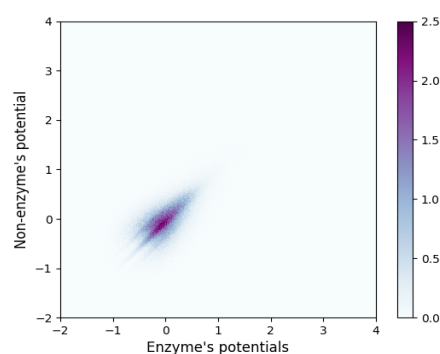
(A)



(B)



(C)



(D)

Figure 2-14 Potential comparison between enzymes vs. non-enzymes at (A) amino acid level, (C) codon level, (D) codon level (heatmap). Enzymes are random split into half and their potential scatter plot shows in (B).

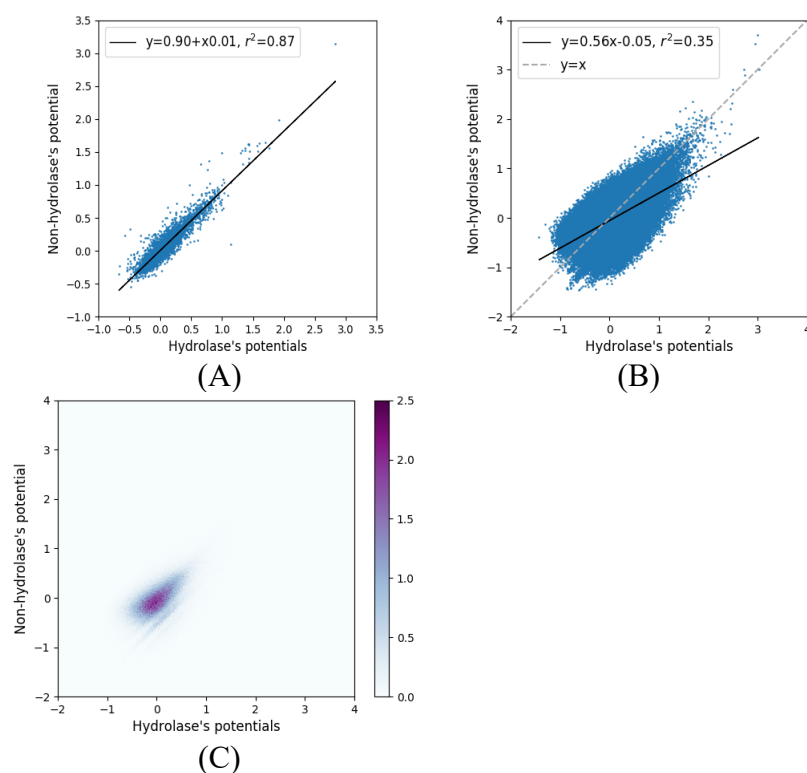
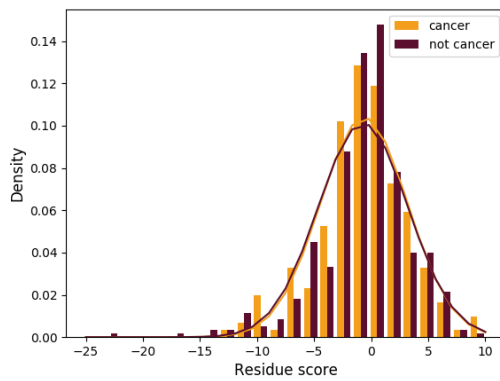


Figure 2-15 Potential comparison between hydrolases vs. non-hydrolases at (A) amino acid level, (B) codon level and (C) codon level (heatmap)

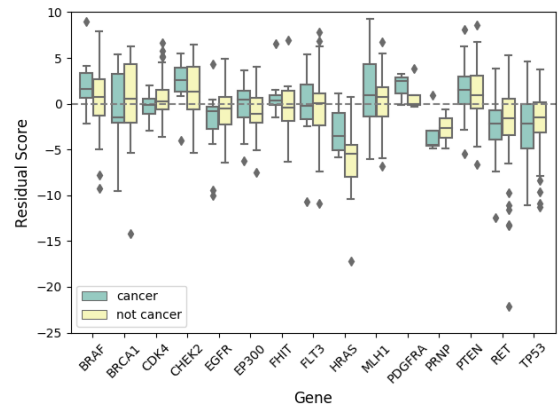
2.3.7 Linking protein structure and functions by silent mutations

Figure 2-16 (A) demonstrates the density distribution of residual scores by using Delaunay tessellation from cancer causing silent mutations and non-cancer-causing silent mutations. Overall, the distribution of residual scores was similar. Residual score of 0 suggests the mutation causes no disturbance in local structures around the mutation site, which is expected as mostly silent mutations cause no function change in proteins. When residual score was 0, cancer-causing silent mutations had a lower density compared to

non-cancer-causing silent mutations. This result may suggest cancer causing mutations caused more structure disturbance than non-causing mutations. Cancer distribution slightly shift to left, but overall, we did not see significant different between two distributions. When we looked at individual genes, *HRAS* gene had the most distinct residual score distribution between cancer vs. non-cancer-causing silent mutations in Figure 2-16(B).



(A)



(B)

Figure 2-16 Distribution of residue scores between cancer causing silent mutations vs. non-cancer-causing silent mutations in (A) histogram with all gene combined and (B) boxplot for each individual gene.

Table 2-3 reports the BAR results of Naïve Bayes and Random Forest using 5-fold cross-validation on WEKA. In the 5 balanced datasets that each contained 208 cancer causing and 208 non-cancer-causing silent mutations, we found similar BARs, all slightly above 0.5 by using Naïve Bayes or Random Forest. When we shuffled the silent mutations to cancer causing or non-cancer causing, the BARs were lower and closer to 0.5. The results suggested that the signal of residual scores from cancer causing silent mutations was low, not enough to differentiate from the residual scores from non-cancer-causing silent mutations. In the scatter plot, we found that the correlation between residual score and SynMICdb scores was not correlated ($R^2=0.01$) (Figure 2-17).

Table 2-3 BAR results of Naïve Bayes and Random Forest using 10-fold cross-validation (WEKA)

Method	Data processing	Dataset1	Dataset2	Dataset3	Dataset4	Dataset5
Naïve Bayes	Original	0.577	0.536	0.574	0.552	0.555
	shuffled	0.497	0.495	0.508	0.492	0.486
Random Forest	Original	0.558	0.489	0.516	0.475	0.514
	shuffled	0.486	0.459	0.459	0.516	0.475

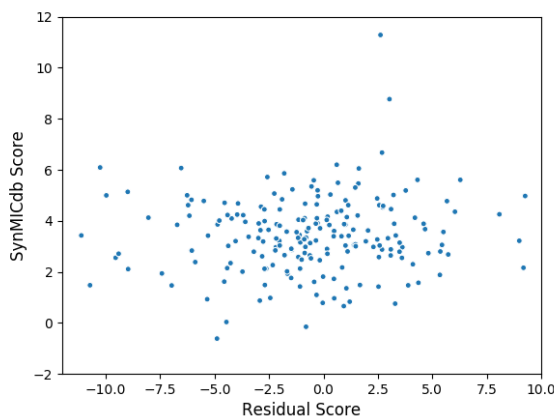


Figure 2-17 Correlation between residual scores and SynMICdb scores.

2.3.8 Conclusions

In summary, we have created a novel dataset containing 10,220 protein chains with codon, amino acid, secondary structure annotations, and α -carbon coordinates. We performed Delaunay tessellation on the proteins and obtained knowledge-based potentials at both amino acid and codon levels. For each simplex composition at amino acid level, potential estimations in this study highly agreed with the results in a previous study, validating Delaunay tessellation performs reliably in different datasets. The larger sample size in this study improved the potential estimations for some simple compositions that only had few simplices observed. Restricting simplex edge length and sequence bias also improved the potential estimation. Protein subgroups of enzymes, hydrolases influenced potential estimation at the codon level but not at the amino acid level. Learning curves showed a much larger sample size was required to achieve the same correlation R-square

at codon level compared to amino acid level. Including more available protein structures to perform codon level potential analysis is an important next step. Cancer-causing synonymous mutations and non-cancer-causing synonymous mutations in our dataset had similar residual scores. We did not observe different protein topological properties by using local profiles between them.

CHAPTER 3 : EXPLORING FITNESS AND ACTIVITY OF PROTEIN MUTANTS WITH COMPUTATIONAL MUTAGENESIS AND MACHINE LEARNING TECHNIQUES

3.1 Overview

This study explores the implementation of the computational mutagenesis methods from Masso et al. (2019) [46] to examine the generalizability of the methods in different proteins and aims to extend the methods in a novel angle to perform the analysis at the codon level. We used the computational mutagenesis technique based on a four-body statistical knowledge-based potential [33, 47] to examine the structural impacts of mutation variants of β -lactamase, which had available solved structure data and experimentally fitness data. β -lactamase (PDB ID: 1ZG4) is a type of enzyme (EC 3.5.2.6) produced by some bacteria that is responsible for their resistance to β -lactam antibiotics such as penicillin, cephalosporins, cephamycin and carbapenems [48]. The fitness dataset of β -lactamases protein at both the amino acid and codon levels were obtained from Mehlhoff et al. (2020) [49]. β -lactamases protein contains 263 amino acids. We have access to fitness scores for all 19 single residue β -lactamases mutants at positions 26-290 except for position 239 and 253, which are missing residues in PDB. For all $263 \times 19 = 4,997$ β -lactamases variants, both residual and local profile vectors were calculated. Additionally, the fitness scores of single residue substitution mutations at the codon level are also available. The structural changes and corresponding experimentally fitness score changes caused by single amino acid or codon substitution on protein structure were linked. Random forest classification and regression machine

learning algorithms were used to predict fitness based on structures. We used cross-validation to evaluate model performance, and conduct control experiments for evaluation of statistical significance. Built from Masso et al. (2019), this study explores the sensitivity and specificity of computational mutagenesis at a codon level and investigates the generalizability of the model on a different protein as additional analysis. The results will add knowledge on how to study and predict fitness and structure changes caused by mutations in proteins.

A previous study used machine learning models to predict function determining or buried residues through the analysis of saturation mutagenesis techniques using 12 published deep mutagenesis scanning datasets. The publicly available datasets bring an exciting opportunity for us to apply the computational mutagenesis on them. Additionally, the mutation's impact was measured in activity score in different selections rather than fitness scores in β -lactamases. We will also examine whether the model built for one protein can be applied to other proteins. The results may bring important knowledge on the application of computational mutagenesis techniques on protein activities in a more diverse protein library.

3.2 Materials and Methods

3.2.1 Four-body statistical potential estimation using Delaunay tessellation in β -lactamase

We performed Delaunay tessellation on the β -lactamase protein at both amino acid and codon levels by following the methods described in 2.2.2 *Knowledge-based potential estimation*. By following the definitions proposed by Masso et al. (2019) [32],

we added up the potential scores of all the Delaunay simplices in the proteins. For each residue, a residue environment score (RES) is the sum of potential scores from all simplices this given residue belongs to. The number of simplices a residue belongs to varies. We ordered the collective RES scores for all amino acid positions in each protein to create a potential profile as described by previous studies [32].

3.2.2 Computational mutagenesis on β -lactamase

For a single amino acid residue substitution mutant in the β -lactamase protein, the knowledge-based potentials of its substitution mutations to the other 19 amino acid residues were calculated. A few key definitions are adapted from the work by Masso et al. (2019) [46] in this chapter. Residual score is the difference of total potential between the of mutant and the native protein, representing the relative change in sequence-structure compactivity caused by the mutation. Residual profile is the difference of total potential profiles (component-wise subtraction of vectors) between the mutant and the native protein. The Comprehensive Mutation Profile (CMP) score is the mean residual scores of all 19 amino acid substitutions, representing the average impact of single residue substitution mutations on the structural compatibility of the protein sequence [32].

We compared the residual scores results in different amino acid subgroups. Based on their physicochemical properties, the 20 amino acid types were divided into six subsets as (A, S, T, G, P), (D, E, N, Q), (R, K, H), (F, Y, W), (V, L, I, M), (C) [50]. Amino acid substitution within a subset is categorized as Conservative (C) substitutions and the rest are categorized as Non-Conservative (NC) substitutions [32]. Additionally,

residues were divided into polar (C, G, H, N, Q, S, T, W, Y), apolar (A, F, I, L, M, P, V), and charged (D, E, K, R) [32]. we examined the residual CMP scores across these categories.

3.2.3 β -lactamases variant feature vectors, machine learning, and model evaluation

We ran the models in two approaches. In the first approach, we used a supervised learning algorithm to build a predictive model which was a complex non-linear function of the inputs and produced outputs based on the training data. We created a 267-dimensional (267D, 20 amino acid rows \times 267 columns) residual profile for the β -lactamase protein, including 263 β -lactamase single residue substitution variants (columns #1-263), position number of the mutation (column #264), native residue (column #265), replacement residue (column #266), and the experimentally determined fitness value (column #267). The 267D residual profile was included in supervised machine learning algorithms to train predictive models where the fitness value (column#267) is the output attribute while all preceding components are input variants.

In the second approach, we created a 28-dimensional (28D, 20 amino acid rows \times 28 columns) local profile for the β -lactamases protein, including the same 27 input attributes (columns #1-27) as Masso et al. (2019) and the experimentally determined fitness value of β -lactamase protein (column #28). Specifically, columns #1-27 for each β -lactamase variant are: (#1) mutated position number; (#2) native amino acid; (#3) replacement amino acid; (#4) residue score; (#5-10) EP scores at the six nearest neighbors (sorted) from the residual profile; (#11-16) amino acid identities at the six nearest neighbors (sorted); (#17-22) difference in primary sequence numbers between

that of the mutated position and those of the six neighbors (sorted); (#23-24) mean volume and mean tetrahedrality for all tetrahedra in the tessellation of β -lactamase that share the C-alpha coordinate of the mutated residue as a vertex; (#25) tessellation-based location of the mutated position in the protein (surface, undersurface, or buried); (#26) the number of edge contacts it has with surface residue positions; (#27) the secondary structure at the mutated position [32]. Random forest classification and regression algorithms were implemented using Weka software package of machine learning tools [51, 52] and Python. The experimentally determined numerical fitness value (column #28) was the output and the variables (columns #1-27) were inputs in the regression model. Random forest classification algorithms require the output of the training set have categorical rather than numerical values, and the trained models can predict the classification outcomes. All 4,997 β -lactamases variants were grouped into the two equally sized fitness categories (*increased* versus *decreased* β -lactamases variant fitness relative to the native protein) by using the median fitness value as the cutoff.

We applied the Leave-One-Out Cross-Validation (LOOCV) and 10-fold Cross-Validation (10-fold CV) procedures to evaluate model performance. Using 10-fold CV, the β -lactamases variants were initially placed into ten disjoint subsets of equal size with the following steps: (1) one subset was retained (10% of the data), while a model was trained using combined β -lactamases variants from the other nine subsets (90% of the data); (2) a well-trained model is used to predict (known) β -lactamase variants in the retention set; (3) we repeat the process to allow each subset be held-out once and predicted. With LOOCV, each β -lactamase variant forms its own subset (i.e., singleton)

in the initial step before going through the same iterative process as the 10-fold CV. We categorize the suitability of β -lactamase variants relative to wild type β -lactamase as increased (P, positive) or decreased (N, negative). Random forest classifier predictions were evaluated by $Se = sensitivity = \frac{TP}{TP+FN}$, $Sp = specificity = \frac{TN}{TN+FP}$, and $PPV = positive predictive value = \frac{TP}{TP+FP}$, where TP is True Positives, TN is True negatives, FP is False Positives, and FN is False Negatives. Performance[37] measures including *Balanced Accuracy rate* = $BAR = 0.5 \times (Se + Sp)$, area (AUC) under the receiver operating characteristic (ROC) curve, and Matthew's correlation coefficient (MCC). In random forest regression predictions, Pearson's correlation coefficient (r) between the experimentally determined fitness values and the predicted ones and the root mean square error (RMSE) were calculated. Furthermore, by using the median fit as a threshold, all random forest regression model fit values (actual values and predicted values) are converted into categories in order to calculate the classification performance indicators given above, and then the results are compared with the results of the random forest model.

3.2.4 Computational mutagenesis methods on protein activities

A recent study used machine learning models to predict function determining or buried residues through the analysis of saturation mutagenesis techniques [37]. The study made the 12 deep mutagenesis scanning datasets publicly available, and we extend this computational mutagenesis methodology to these datasets, which include individual datasets of proteins and provide mutation's impact on protein activities as shown in Table

3-1. For datasets included, the residue numbers ranged from 52 to 261 with single residue substitution mutations ranging from 988 to 4,959 (Table 3-1).

We performed same machine learning techniques as described in 3.2.3. We noticed that the protein activity selection can be categorized (e.g., antibiotic resistance, Ubiquitin ligase activity). After building the prediction models, we interchangeably applied the model built for one protein to predict activity scores for another protein in the same category (e.g., antibiotic resistance), evaluating the generalizability of models.

Table 3-1 Protein mutation activity datasets

Dataset	PDB ID	# of residue s	# of single residue substitution s	Host	Selection	Citation
BRCA1-E3	1JM7.A	99	1358	<i>S. cerevisiae</i>	Ubiquitin ligase activity	Starita et al. (2015) [53]
BRCA1-Y2H	1JM7.A	99	1359	<i>S. cerevisiae</i>	Binding activity (Y2H)	Starita et al. (2015) [53]
CcdB	3VUB. A	98	1627	<i>E. coli</i>	Toxin activity	Adkar et al. (2012) [54]
Gal4	3COQ. A	57	1083	<i>S. cerevisiae</i>	Transcription factor activity	Kitzman et al. (2015) [55]
GB1	1PGA. A	52	988	<i>Streptococcus sp. group G</i>	IgC-Fc binding	Olson et al. (2014) [56]
Hsp90	2CG9. A	202	3814	<i>S. cerevisiae</i>	Chaperone activity	Mishra et al. (2016) [57]

Dataset	PDB ID	# of residue s	# of single residue substitution s	Host	Selection	Citation
KKA2	1ND4. A	253	4807	<i>E. coli</i>	Antibiotic resistance	(Melnikov et al. (2014)) [58]
NUDT1 5	5LPG. A	154	2692	<i>E. coli</i>	Abundance and drug sensitivity	Suiter et al. (2020) [59]
PSD95	1BE9.A	83	1577	<i>E. coli</i>	Ligand binding	McLaughlin et al. (2012) [60]
Ras	5P21.A	163	3097	<i>E. coli</i>	Antibiotic resistance	Bandaru et al. (2017) [61]
TEM1	1ZG4.A	261	4959	<i>E. coli</i>	Antibiotic resistance	Stiffler et al. (2015) [62]
Ubiquitin	1UBQ. A	72	1154	<i>S. cerevisiae</i>	Ubiquitin ligase activity	Roscoe et al. (2013) [63]

Note: Mutagenesis libraries were partly obtained from Bhasin et al. [37]. One dataset in Bhasin et al., Pab1 (RRM domain) using 1CVJ, was not included due to many unmatched residues between PDB and experimentally determined activity data.

3.2.5 Computational mutagenesis methods on double mutations

Olson et al. (2014) published their work quantifying the effect of all pairwise mutations in the IgG-binding domain of protein G (GB1), and they made the data available [56]. We applied the same computational mutation genesis methods described in 3.2.3 and 3.2.4 on this pairwise mutation dataset.

3.3 Results and discussions

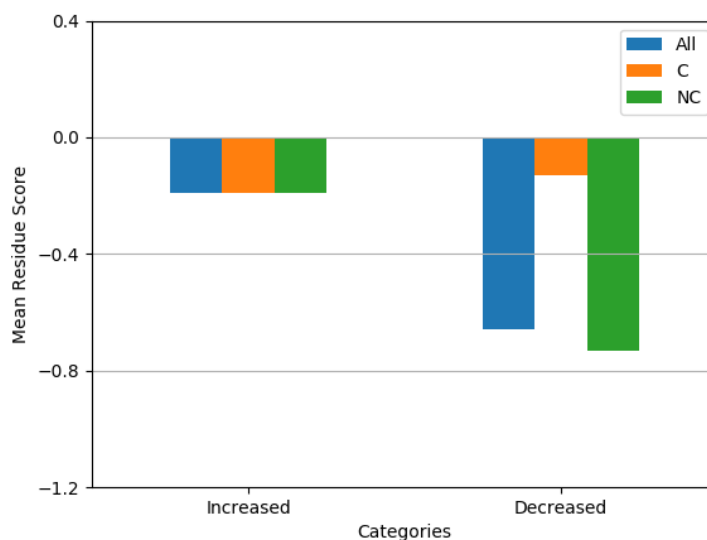
3.3.1 Fitness scores summary of β -lactamases

Residue scores were calculated for all 263 β -lactamases variants upon each of 19 possible amino acid residue replacements at position 26-290 (except 239 and 253) in the β -lactamases structure. We grouped the mutants based on their experimental fitness categories (i.e., decreased versus increased relative to the median fitness of β -lactamases protein of -0.0126 at amino acid level and -0.0148 at codon level) and residual scores were averaged in each the above grouping category.

We found that overall variants had negative mean residual score regardless of increased or decreased experimentally determined fitness score category (

Table 3-2). This is different from the results from the study by Masso et al. (2019) where the increased category had a positive mean residual score and the decreased category had a negative score [32]. We could, however, tell that the mean residual scores were different in the increased vs. decreased categories in this study as their confidence intervals did not overlap. Possible reasons included: (1) relatively high missingness of the experimentally determined fitness score of β -lactamases: 106 out of 5,260 (2.0%) missing at amino acid level, 938 out of 16,043 (5.8%) missing at codon level; (2) the impact of the threshold value to equally divide the 4,997 variants into decrease and increase groups. There is significant difference in the mean fitness scores in the Increased vs. Decreased variants in student's t-test (-0.19 vs. -0.66, p-value < 0.0001). After subgrouping the variants into conservative (C)/non-conservative (NC) amino acid [50] substitutions of the native residues, we found that non-conservative (NC) amino acid substitutions are strong drivers for the structure-function relationship in both the increased and decreased fitness score category, although both C and NC substitutions have impacts in the same direction. The results were as expected since conservative amino acids are more similar to each other and their substitution within the conservative category would bring limited impact. In contrast, we would expect to see more impact of a non-conservative amino acid substitution.

Table 3-2 Beta-lactamases protein structure-function relationship. All refers to the collection of all 4997 β -lactamase variants with experimental fitness data. C/NC is a subset of these variants and represents conservative/non-conservative amino acid substitutions of natural residues. The data in the right table is the average of the residual scores for the relevant subset of mutants. All numbers in parentheses on the graph or table row/column headers are counts of the total number of mutants in the subset.



Category (number)	Increased (2499), mean 95% CI	Decreased (2498), mean 95% CI
All(4997)	-0.26 (-0.31, -0.21)	-0.53 (-0.59, -0.47)
C(827)	-0.14 (-0.22, -0.07)	-0.09 (-0.18, 0.01)
NC(4170)	-0.29 (-0.35, -0.23)	-0.59 (-0.65, -0.52)

3.3.2 Residual scores distinguish between categories of β -lactamases amino acids

We found a strong inverse correlation ($r^2=0.78$) between CMP and RES in Figure 3-1. The polarity of the native amino acid did not impact the correlation between CMP and RES (see no obvious cluster in the same polarity subgroups). There was a linear relationship between residual CMP score and RES among hydrophobic residues, suggesting substitutions among hydrophobic residues may have a higher impact on protein topological properties. These results are consistent with the Ras protein results by Masso et al. (2019) [32].

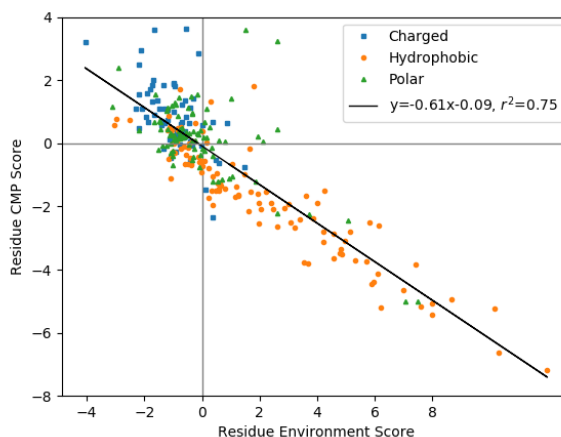


Figure 3-1 CMP scores for the β -lactamases amino acids vs. their residue environment scores. At each amino acid position, the 19 β -lactamase variants were categorized into hydrophobic, polar, and uncharged groups.

3.3.3 Machine learning models for predicting β -lactamases variant fitness

Table 3-3 reports the Leave-one-out cross-validation (LOOCV) performance results at the amino acid level. When using the Random Forest Classification (RFC),

LOOCV performance on β -lactamases variant datasets at the amino acid level has a 0.72 sensitivity and 0.75 specificity in residue profile classification, a 0.73 sensitivity and 0.75 specificity in local profile classification. Overall, the model performance metrics (Sensitivity, Specificity, PPV, BAR, MCC, and AUC) are quite comparable using the residual profiles vs. the local profiles (Table 3-3). The array presented in Figure 3-2 clearly visualized the RFC LOOCV predictions obtained for all 4,997 β -lactamases variants in the residual profiles dataset. When using Random Forest Regression (RFR), LOOCV performance on β -lactamases variant datasets at the amino acid level has a r of 0.64 and RMSE of 0.05 in residue profiles and r of 0.67 and RMSE of 0.05 in local profiles. Again, the model performance metrics (r , RMSE, BAR, MCC) are quite comparable using the residual profiles vs. the local profiles in the RFR.

Control datasets are created by randomly shuffling the experimentally determined fitness output for the β -lactamases variants in each original dataset. The modeling performance dropped significantly by using the shuffled data compared to the original variants data in both residual profiles or local profiles in random forest classification. The AUC dropped from 0.80 to 0.50 in residual profiles and from 0.70 to 0.51 in local profiles, suggesting the models are performing no better than random guessing (AUC=0.5).

Table 3-3 LOOCV performance on β -lactamases variant data sets at amino acid level.

Model	Profile	Processed	Sensitivity	Specificity	PPV	BAR	MCC	AUC
RFC	Residual	Original	0.72	0.75	0.76	0.73	0.46	0.80
		Shuffled	0.50	0.50	0.51	0.50	0.00	0.50
	Local	Original	0.73	0.75	0.75	0.74	0.48	0.80
		Shuffled	0.52	0.52	0.53	0.52	0.03	0.51
Model	Profile	Processed	r	RMSE	-	BAR	MCC	-
RFR	Residual	Original	0.64	0.05	-	0.70	0.40	-
		Shuffled	-0.01	0.07	-	0.55	0.07	-
	Local	Original	0.67	0.05	-	0.71	0.41	-
		Shuffled	0.00	0.07	-	0.52	0.02	-

Note: RFC = Random Forest Classification; RFR = Random Forest Regression; PPV=positive predictive value; BAR=balanced accuracy rate; MCC=Matthew's correlation coefficient; AUC= area under the curve; RMSE=root mean squared error.

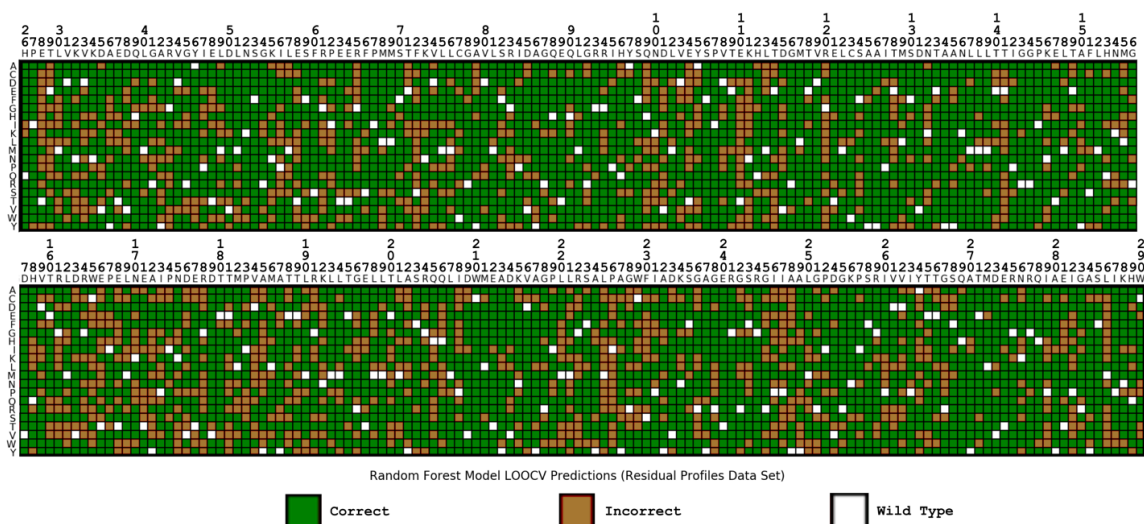


Figure 3-2 Random forest classification leave-one-out cross-validation (LOOCV) prediction array for all 4,997 β -lactamases variants (Residual Profiles data set). Collectively, these predictions yield the performance summary data in the top row. Columns correspond to the β -lactamases amino acid positions, and rows represent the 19 different types of residue replacements with wild type. A β -lactamases variant is labeled correct (green) if its experimental and predicted fitness categories are identical; otherwise, the variant is labeled incorrect (brown).

We evaluated the significance of β -lactamases prediction performance. Figure 3-3 illustrates that Random Forest Regression predicted fitness has a correlation ($r=0.67$) with the β -lactamases variant experimentally determined fitness values. The 0.67 r value in this study is lower than the r value of 0.79 from the study by Masso et al. (2019) [46]. Besides the inherent differences in β -lactamases protein vs. β -lactamases protein as the materials, different sources/quality of experimental fitness data may influence the r value differences as well. For the missing values in the experimentally determined fitness values in β -lactamases, we replaced them with the mean of the available fitness values

from all other mutations at the same position. Additionally, the collateral fitness effects we observed were associated with TEM-1/ preTEM-1 aggregation, improper signal sequence cleavage, impaired release of the mature protein from the membrane, incorrect disulfide-bond formation, induction of stress-response pathways, and pleiotropic changes in cell phenotype [61].

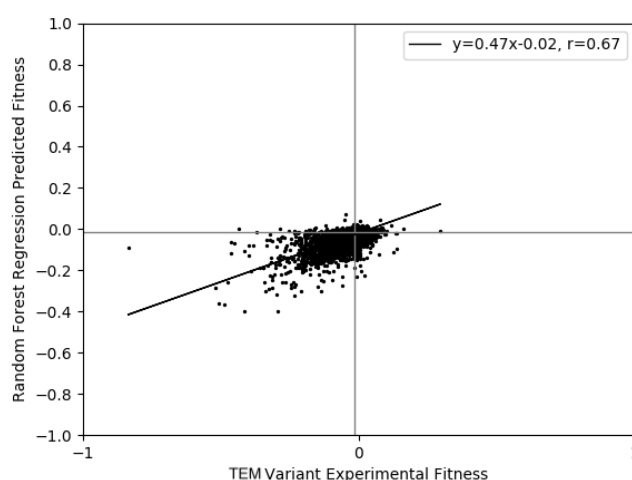


Figure 3-3 Evaluating the significance of β -lactamases prediction performance. The scatter plot compares the tree regression leave-one-out cross-validation (LOOCV) predicted values obtained for β -lactamase variability values (using a local feature data set) with their experimental measurements.

Table 3-4 reports the Leave-one-out cross-validation (LOOCV) performance on β -lactamases variant data sets at codon level. Using Random Forest Classification, LOOCV

performance on β -lactamases variant datasets at the codon level has a 0.71 sensitivity and 0.70 specificity in residue profile classification, a 0.71 sensitivity and 0.70 specificity in local profile classification. Higher model performance was obtained in residual profiles compared to shuffled residual profiles, and in local profiles compared to shuffled local profiles, suggesting the importance of signals encoded in the β -lactamases variant input attributes for effectively determining the fitness values. The array presented in Figure 3-4 clearly demonstrates the random forest LOOCV predictions obtained for all 16,043 β -lactamases variants in the Residual Profiles dataset. Using Random Forest Regression, LOOCV performance on β -lactamases variant datasets at the codon level has a r of 0.56 in residue profiles and r of 0.59 in local profiles. We found that the sensitivity and specificity for random forest classification and regression were lower at codon level compared to amino acid level, respectively in Table 3-3 and

Table 3-4. The results may suggest that additional input at the codon level did not improve the model performances in predicting the experimentally determined fitness.

Linking back to results in 2.3.3 *Learning curves of potential estimations in the 10220cullled dataset*, where we found that the knowledge-based potential estimations may be less accurate due to inadequate sample size, we think this may contribute to the low sensitivity and specificity at codon level. We cannot draw conclusions for this finding at this time.

Table 3-4 LOOCV performance on β -lactamases variant data sets at codon level.

Model	Profile	Processed	Se	Sp	PPV	BAR	MCC	AUC
RFC	Residual	Original	0.71	0.70	0.69	0.71	0.41	0.77
		Shuffled	0.51	0.51	0.53	0.51	0.02	0.51
	Local	Original	0.70	0.69	0.69	0.70	0.40	0.77
		Shuffled	0.50	0.50	0.51	0.50	0.00	0.50
Model	Profile	Processed	r	RMSE	-	BAR	MCC	-
RFR	Residual	Original	0.56	0.05		0.69	0.37	-
		Shuffled	-0.01	0.07		0.51	0.03	-
	Local	Original	0.59	0.05		0.69	0.38	-
		Shuffled	-0.01	0.07		0.51	0.01	-

Note: RFC = Random Forest Classification; RFR = Random Forest Regression;
Se=sensitivity; Sp=specificity; PPV=positive predictive value; BAR=balanced accuracy
rate; MCC=Matthew's correlation coefficient; AUC= area under the curve; RMSE=root
mean squared error.

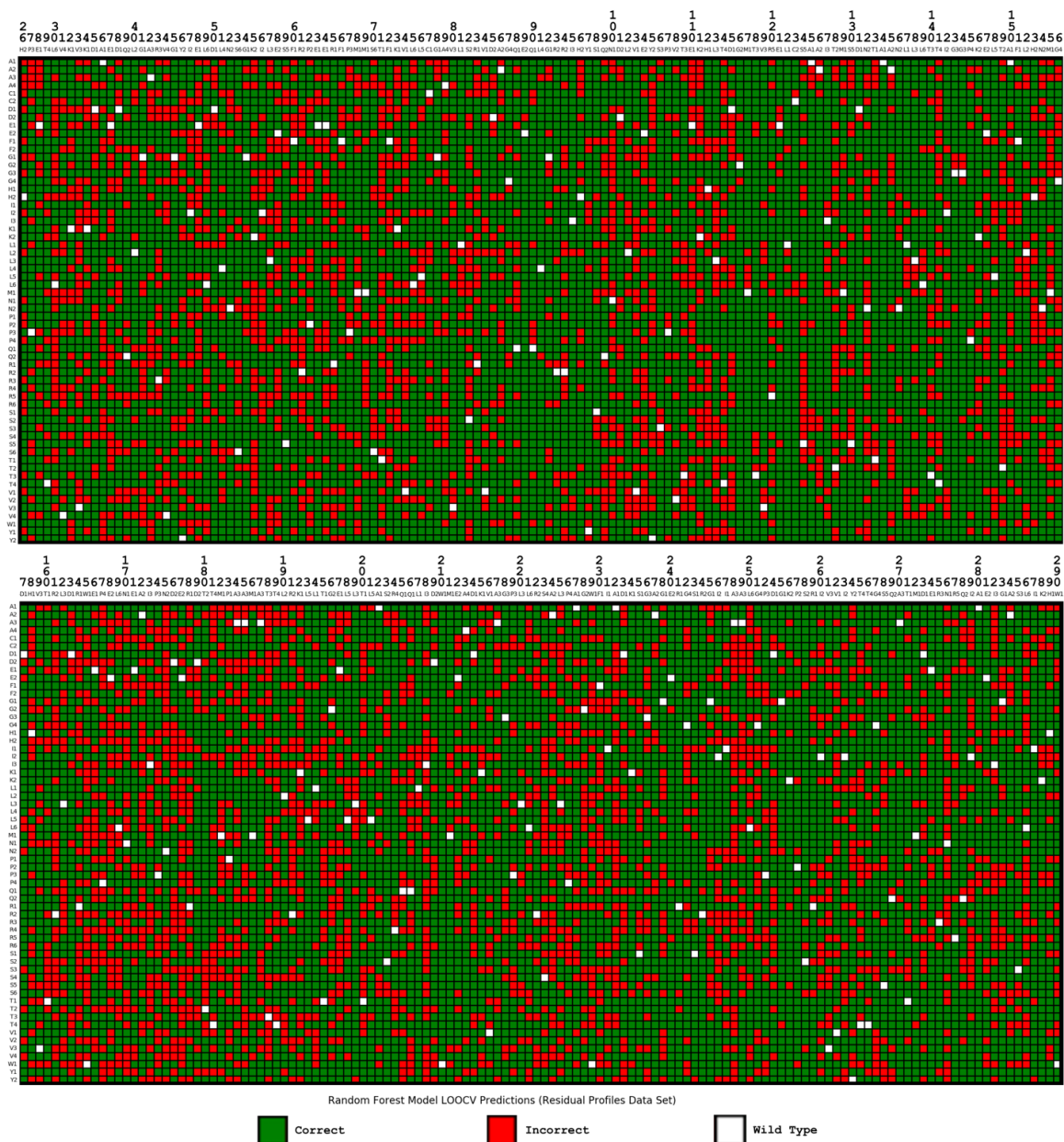


Figure 3-4 Random forest classification leave-one-out cross-validation (LOOCV) prediction array for all 16,043 β -lactamases variants (Residual Profiles data set) at codon level. Collectively, these predictions yield the performance summary data in the top row. Columns correspond to the β -lactamases amino acid positions, and rows represent the 61 different types of residue replacements with codon notation. A β -lactamases variant is labeled correct (green) if its experimental and predicted fitness categories are identical; otherwise, the variant is labeled incorrect (red).

We found that β -lactamases variants for which polar amino acids replaced polar, apolar, or charged amino acid were predicted in a similar accurate level, with a BAR range 0.69-0.75 (Table 3-5). In general, the performance of our model was similar across amino acid substitutions categorized into polar, apolar, or charged categories.

Table 3-5 Mean random forest leave-one-out cross-validation (LOOCV) prediction performance (β -lactamases variant Local Profiles) based on side chain polarities of the native and new amino acids at the mutated position.

New/native	Polar			Apolar			Charged		
	BAR	MCC	%	BAR	MCC	%	BAR	MCC	%
Polar	0.73	0.46	0.14	0.74	0.49	0.12	0.75	0.50	0.07
Apolar	0.72	0.44	0.20	0.70	0.40	0.13	0.73	0.47	0.09
Charged	0.70	0.39	0.12	0.73	0.46	0.09	0.69	0.39	0.04

Note: BAR=balanced accuracy rate; MCC=Matthew's correlation coefficient.

We found that predictions for mutations at residue positions on the protein undersurface were found to be the most accurate relative to those that are more buried. The predictions for mutations at residue positions in β -lactamases coils and helices were most accurate when compared to strands (

Table 3-6).

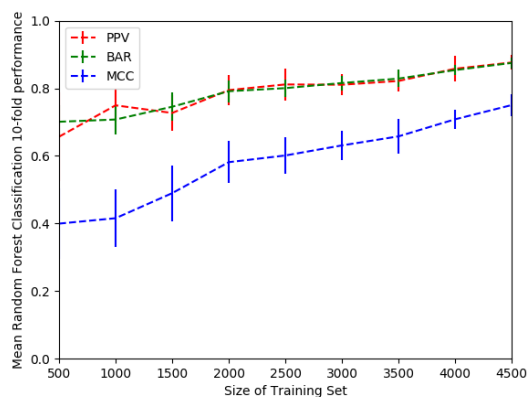
Table 3-6 Mean random forest leave-one-out cross-validation (LOOCV) prediction performance (β -lactamases variant Local Profiles) based on depth and secondary structure.

		BAR	MCC	%
Depth	Buried	0.54	0.08	0.03
	Undersurface	0.73	0.45	0.41
	Surface	0.72	0.44	0.57
Secondary structure	Strand	0.68	0.37	0.17
	Helix	0.73	0.46	0.44
	Coil	0.74	0.47	0.39

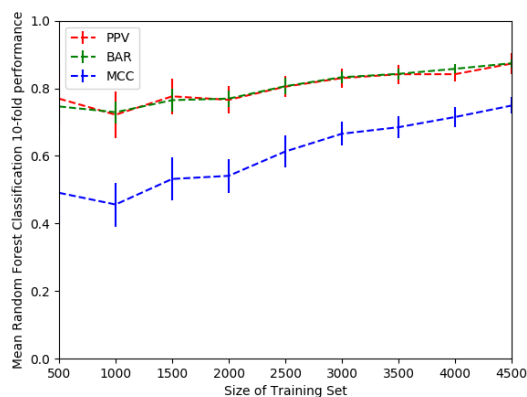
Note: BAR=balanced accuracy rate; MCC=Matthew's correlation coefficient.

3.3.4 Learning curves exploration in β -lactamase models

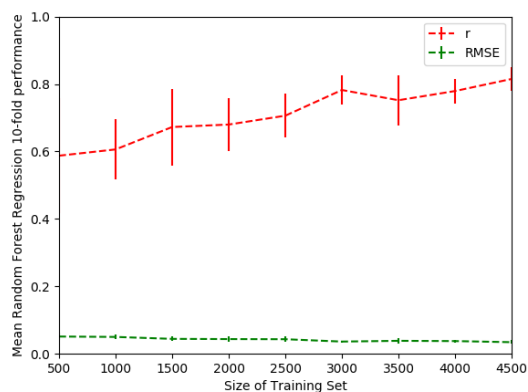
Learning curves demonstrated the relationship between model performance and the size of training set using both the residual profiles and local profiles (Figure 3-5). We found that the mean Random Forest Classification performance including PPV and BAR in 10-fold CV increased as the training set increased from 500 to 2,500 but plateaued after approximately 2500 in both the residual profiles and local profiles. MCC tend to continuously increase with greater training set size. R value increases slightly while RMSE stayed largely consistent with greater training set size. However, the largest dataset we collect from the original data is 4,998, which indicate if we have more data, the precision might go higher.



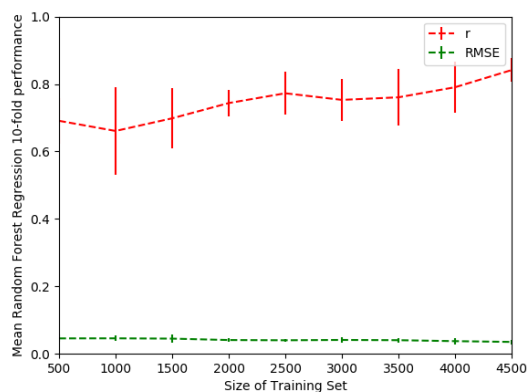
(A)



(B)



(C)



(D)

Figure 3-5 Learning curves. The plots reveal the degree to which performance is improved as the number of TEM variants in the training set is increased. Each point represents the average over ten runs of 10-fold CV, and the error bars indicate the standard deviation. Plots were generated by using both types of TEM variant data sets (Residual Profiles and Local Profiles feature vectors) with both random forest classification and tree regression. PPV=positive predictive value; BAR=balanced accuracy rate; MCC=Matthew's correlation coefficient; AUC= area under the curve; RMSE=root mean squared error.

3.3.5 Inclusion of deep learning exploration

To examine if deep learning algorithm improves the results, we performed Artificial Neural Network (ANN) to the data using Tensorflow Keras. The basic architecture we used is the feed-forward neural network (FFNN). There were many hyper-parameters need to be tuned before performing. These parameters were: number of neurons in the input and hidden layers, number of hidden layers, dropout rate, optimizer, learning rate, regularizer, loss function, activation function, epoch and batch size. The basic idea is to change one hyper-parameter at one time and keep the others, plot both training and cross-validation accuracies, choose the point where both training and cross-validation have good performance. The final hyper-parameters used in classification: (1) input layer has 40 neurons, 3 hidden layers each has 80 neurons; input and hidden layer both use ReLU function; output layer use softmax function; (2) kernel regularizer for each layer: L1 regularization penalty= $1e-5$; (3) loss function: Sparse Categorical Crossentropy; (4) optimizer: Adam, learning rate= $5e-4$; (5) Epoch: 100; (6) Batch: 15. Regression use same hyper-parameters, except the loss function use mean absolute error. We applied the obtained tuning hyper-parameters in the ANN model. We compared the ANN result with random forest classification with 10-fold cross validation.

Table 3-7 and

Table 3-8 report the model performance with 10-fold cross-validation using either random forest classification (Table 3-7) /random forest regression (

Table 3-8) or ANN models. Overall ANN models did not improve the model performance.

To investigate if synonymous codons provide additional information in creating the model, we randomly shuffled synonymous codons for each of the 20 amino acids for 1000 times. For example, Alanine has 4 codons: A1, A2, A3 and A4, and we shuffled all the Alanine codons but kept other local profile variables unchanged. The results are presented in Table 6, row 4. The random forest classification sensitivity increased from 0.70 at original codon level to 0.71 at the synonymous codons shuffled level while the specificity stayed the same of 0.70. RFC at the codon level did not achieved better model performance than at the shuffled codon level suggests that synonymous codons do not provide additional information helpful in predicting the β -lactamases protein fitness by using topological models.

Table 3-7 Comparison of 10-fold cross-validation prediction performance (β -lactamases variant Local Profiles) at the amino acid or codon level using Random Forest Classification (RFC) or Artificial Neural Networks (ANN).

Level	Amino acid					Codon				
Metric	Sensitivity	Specificity	PPV	BAR	MCC	Sensitivity	Specificity	PPV	BAR	MCC
RFC	0.73	0.72	0.71	0.72	0.45	0.70	0.70	0.69	0.70	0.40
ANN	0.73	0.71	0.70	0.72	0.44	0.72	0.69	0.67	0.71	0.42
RFC shuffling	-	-	-	-	-	0.70	0.69	0.68	0.70	0.40

Note: PPV=positive predictive value; BAR=balanced accuracy rate; MCC=Matthew's correlation coefficient. RFC = random forest classification; ANN = Artificial Neural Network.

Table 3-8 Comparison of 10-fold cross-validation prediction performance (β -lactamases variant Local Profiles) at the amino acid or codon level using Random Forest Regression (RFR) or Artificial Neural Networks (ANN).

Level		Amino acid				Codon				
Metric	r	RMSE	BAR	MCC	r	Metric	RMSE	BAR	MCC	r
RFR	0.67	0.05	0.71	0.41	0.67	0.59	0.05	0.69	0.39	0.59
ANN	0.62	0.05	0.73	0.45	0.62	0.58	0.05	0.70	0.40	0.58

Note: RFC = Random Forest Classification; ANN = Artificial Neural Network;
BAR=balanced accuracy rate; MCC=Matthew's correlation coefficient.

3.3.6 Mutagenesis on protein activities

Table 3-9 reports the 10-fold cross-validation results in Random Forest Regression (RFR). We observed moderately high r values in 0.70-0.88, where GB1 dataset yielded the highest r value. When the experimentally determined activity score was rescaled, we obtained the same r values but different RMSE, most of which were reduced. This result indicates that rescaling minimally impacts the activity score. Rescaling transformed the data and made most datapoints lie within the normal distribution with a mean of zero, narrowing down the range of the original data. Rescaled data are more comparable and they bring more easily interpreted results.

Table 3-9 10-fold cross-validation, random forest regression results

Dataset	Reported		Rescaled	
	r	RMSE	r	RMSE
BRCA1-E3	0.70	0.30	0.70	0.15
BRCA1-Y2H	0.77	0.15	0.77	0.22
CcdB	0.70	1.28	0.71	0.18
Gal4	0.71	2.01	0.71	0.28
GB1	0.88	0.93	0.88	0.17
Hsp90	0.83	0.21	0.83	0.19
KKA2	0.68	0.41	0.68	0.34
NUDT15	0.80	0.22	0.80	0.15
PSD95	0.79	0.25	0.79	0.19
Ras	0.79	0.19	0.79	0.28
TEM1	0.80	0.62	0.80	0.24
Ubiquitin	0.76	0.22	0.76	0.20

Note: RMSE = Root-Mean-Square Deviation. Mutagenesis libraries were obtained from Bhasin et al [37].

After rescaling the activity score and categorized them into increased or decreased categories by comparing with the median values, we found that the 10-fold cross-validation results stayed high. Missing result in the CdcB dataset was due to excessive zeros in the dataset.

Table 3-10 10-fold cross-validation, random forest classification, median, rescaled

	Sensitivity	Specificity	PPV	BAR	MCC	PRC
BRCA1-E3	0.78	0.79	0.80	0.79	0.57	0.86
BRCA1-Y2H	0.67	0.69	0.70	0.68	0.35	0.74
CcdB	-	-	-	-	-	-
Gal4	0.84	0.81	0.80	0.83	0.65	0.89
GB1	0.88	0.89	0.89	0.88	0.76	0.95
Hsp90	0.79	0.81	0.81	0.80	0.60	0.87
KKA2	0.78	0.77	0.77	0.77	0.55	0.82
NUDT15	0.81	0.79	0.82	0.80	0.60	0.77
PSD95	0.77	0.79	0.80	0.78	0.57	0.87
Ras	0.80	0.82	0.82	0.81	0.61	0.89
TEM1	0.80	0.81	0.81	0.80	0.60	0.88
Ubiquitin	0.81	0.82	0.81	0.81	0.63	0.90

Note: PPV=positive predictive value; BAR=balanced accuracy rate; MCC=Matthew's correlation coefficient.

We found that overall models were not generalizable within datasets containing proteins evaluated for the same activity categories in Table 3-1. In the antibiotic resistance category, using TEM1 as training set and Ras as the test dataset resulted in the highest sensitivity of 0.65. In the ubiquitin ligase activity category, using BRCA1-E3 as the training dataset and UB as the test dataset brought a sensitivity of 0.56. Overall, the sensitivity, specificity, PPV etc. criteria suggested that model generalization was not good.

Table 3-11 Model performance by using different training and test datasets in two protein activity categories

Activity	Training dataset	Test dataset	Model performance					
			Sensitivity	Specificity	PPV	BAR	MCC	AUC
Antibiotic resistance	Ras	KKA2	0.56	0.55	0.47	0.55	0.11	0.58
	Ras	TEM1	0.63	0.59	0.52	0.61	0.21	0.66
	KKA2	Ras	0.55	0.60	0.73	0.57	0.14	0.59
	KKA2	TEM1	0.58	0.64	0.74	0.61	0.21	0.64
	TEM1	Ras	0.65	0.61	0.56	0.63	0.26	0.66
	TEM1	KKA2	0.64	0.59	0.50	0.62	0.23	0.66
Ubiquitin ligase	UB	BRCA1-E3	0.52	0.56	0.73	0.54	0.08	0.57
	BRCA1-E3	UB	0.56	0.55	0.42	0.55	0.11	0.59

Note: PPV=positive predictive value; BAR=balanced accuracy rate; MCC=Matthew's correlation coefficient; AUC = area under curve.

3.3.7 Mutagenesis on double residue mutations

We found that the computational mutagenesis methods resulted in high sensitivity, specificity, PPV *etc.* metrics in model performance by using either residual profile or local profile as the outcomes in Table 3-12. The shuffled profiles served as a control group had close to 0.5 sensitivity, specificity, and PPV, validating the method is working properly. We ran sensitivity analysis by including only 1/10, 1/100, and 1/000 random samples of the total 270,990 mutation instances in Table 3-12, and we found relatively lower performance metrics including sensitivity, specificity, PPV *etc.* compared to results in Table 3-12. However, even when the only 271 instances were included, the sensitivity, specificity, PPV *etc.* metrics were still high. The effect of

including different number of instances on model performance was demonstrated in Figure 3-6, where 30k-40k was large to bring a sufficiently high model performance. These results suggested that the created model can predict the impacts of double residue mutations on protein fitness well.

Table 3-12 10-fold cross validation, median, classification, instance =270,990

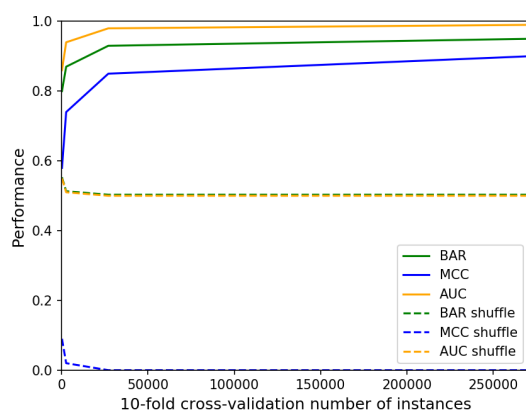
Profile	Process	Sensitivity	Specificity	PPV	BAR	MCC	AUC
residual	original	0.96	0.95	0.95	0.95	0.90	0.99
	shuffled	0.49	0.51	0.47	0.50	-0.00	0.50
local	original	0.95	0.94	0.94	0.94	0.89	0.99
	shuffled	0.49	0.51	0.46	0.50	-0.00	0.50

Note: PPV=positive predictive value; BAR=balanced accuracy rate; MCC=Matthew's correlation coefficient; AUC = area under curve.

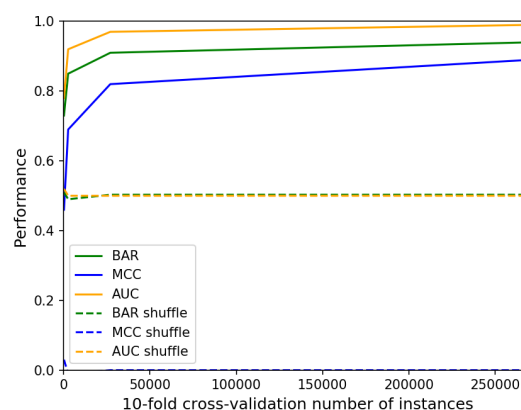
Table 3-13 10-fold cross validation, median, classification, instance number varies

instance	Profile	Process	Sensitivity	Specificity	PPV	BAR	MCC	AUC
27,099	residual	original	0.92	0.94	0.94	0.93	0.85	0.98
		shuffled	0.50	0.50	0.49	0.50	-0.00	0.50
	local	original	0.90	0.92	0.91	0.91	0.82	0.97
		shuffled	0.50	0.49	0.51	0.50	-0.00	0.50
2,710	residual	original	0.86	0.88	0.89	0.87	0.74	0.94
		shuffled	0.51	0.50	0.53	0.51	0.02	0.51
	local	original	0.85	0.84	0.84	0.85	0.69	0.92
		shuffled	0.46	0.53	0.41	0.49	-0.01	0.50
271	residual	original	0.77	0.83	0.88	0.80	0.58	0.86
		shuffled	0.54	0.55	0.49	0.55	0.09	0.55
	local	original	0.72	0.74	0.74	0.73	0.46	0.78
		shuffled	0.51	0.52	0.50	0.51	0.03	0.52

Note: PPV=positive predictive value; BAR=balanced accuracy rate; MCC=Matthew's correlation coefficient; AUC = area under curve.



(A)



(B)

Figure 3-6 Model performance by including different number of instances in the 10-fold cross-validation. (A) Using residual profile as inputs. (B) Using local profile as inputs.

We found that the residual score for mutation 1 and that for mutation 2 had a very low Pearson's correlation coefficient of 0.12. Fitness for mutation 1 and fitness mutation 2 also had a very low correlation coefficient of 0.03. We also found the correlation between fitness of mutation 1&2 was moderately positively correlated with fitness for mutation 1 or fitness for mutation 2. The results are as expected. Only one mutation residual score or fitness score cannot predict the fitness score of double mutations. Topological property changes can be used in predicting the fitness scores of double mutations by using the computational mutagenesis methods.

Table 3-14 Residual scores, mutation fitness scores, and double mutation fitness score Pearson's correlation coefficients

	Residual score for mutation 1	Residual score for mutation 2	Fitness for mutation 1	Fitness for mutation 2	Fitness for mutation 1&2
Residual score for mutation 1	-	-	-	-	-
Residual score for mutation 2	0.12	-	-	-	-
Fitness for mutation 1	0.00	0.17	-	-	-
Fitness for mutation 2	0.29	0.02	0.03	-	-
Fitness for mutation 1&2	0.23	0.15	0.60	0.70	-

We found that each variable was weakly correlated to the fitness class by using local profile or residual profile in Figure 3-7. Considering the high model performance in

Table 3-12, we think the results in this figure suggest that there was no dominating effect of any variables in predicting fitness class but each variable contributed a small part in the prediction. It's useful to include all these variables as model inputs for a reliable model prediction on protein activity.

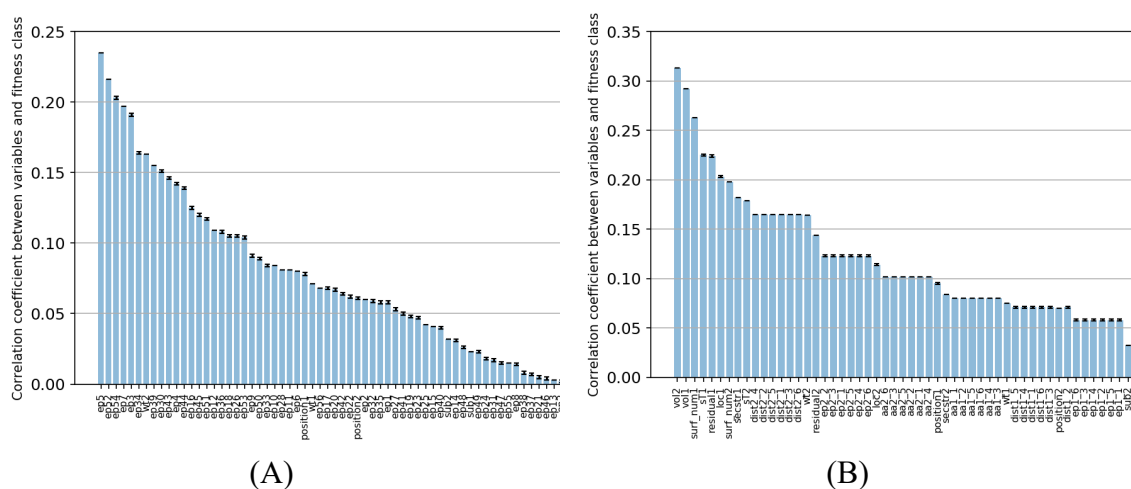


Figure 3-7 Pearson's correlation coefficients between variables and fitness class at (A) residue profile and (B) local profile.

3.4 Conclusions

We applied the machine learning techniques on various proteins which had experimentally determined fitness or activity scores by using the saturation mutagenesis methods. We extended the prediction at both amino acid level and codon level, although

amino acid level prediction had better model performance. By using computational mutagenesis, local profile can be used to predict the impact of mutations on fitness or activity in different proteins with good model performance. Model performance improved in double mutation protein activity score prediction. Model created for one protein had low performance when being applied to a different protein, suggesting low generalizability. The results may be useful to quantify and predict protein fitness and activity changes caused by mutations in different proteins.

CHAPTER 4 : VALGUSHEL, A NOVEL METHOD TO IDENTIFY AND CHARACTERIZE KINKED ALPHA-HELICES

4.1 Overview

We develop a new approach, “ValgusHel”, to identify kinked α -helices in proteins based on their geometrical (ValgusHel-geometry) and topological structures (ValgusHel-topology). We compare the agreement between this two-part method and characterize kinked helices after classification. In ValgusHel-geometry, in an α -helix with at least 9 residues ($i, i+1, \dots, i+8$), we calculated the helix angle formed by a helix axis in residues $i, i+1, \dots, i+5$ and a second helix axis in residues $i+3, i+4, \dots, i+8$. Based on the calculated helix angles, the center residues ($i+4$) and 9-residue α -helix fragments were classified as kinked ($>30^\circ$), curved ($19-30^\circ$) or normal ($\leq 19^\circ$) residues or sequences, respectively. Entire α -helices were classified as kinked, curved, or normal helices based on the highest helix angles estimated. We measured structure alignment deviations of kinked, curved and normal α -helices from a model ideal helix using Root Mean-Square Deviation (RMSD). We examined the residues frequency and calculated sequence similarity score within and across the kinked, curved and normal helices groups. DSSP variables of residues within the α -helices or residues characteristics of surrounding environment around α -helices were included in Random Forest Classification to predict the presence of kinked/normal α -helices. In ValgusHel-topology, we tessellated the protein structures into Delaunay simplices and explored the association between the distribution of types of simplices within a helix and the kinked, curved or normal helix

classification determined by ValgusHel-geometry. We used N-gram in different sequence fragment to determine helix classification (kinked, curved, normal) based on Delaunay simple type stacking within a helix. We compared the agreement of helix classification by using ValgusHel-geometry and ValgusHel-topology. In our data set, a total of 2,621 (0.94%) residues were classified as kinked residues, 17,536 (6.31%) as curved residues, and 257,853 (92.75%) as normal residues. Among kinked and curved residues, mean RMSD monotonically increase when the calculated helix angles increase from 15-46°. Within group sequence similarity scores show highest sequence diversity within kinked sequences group. Both DSSP variables within α -helices and residue characteristics in surrounding environment of α -helices predicts helix angles and thus α -helices classifications with relatively high sensitivity and specificity. By using the distribution of types of Delaunay simplices, we were able to identify kinked helices at both sensitivity and specificity equal to 0.75. N-gram based on the distribution of types of Delaunay simplices is useful to identify kinked helices, and a combination of fragment length ($m \geq 12$) and N-gram ($N \geq 8$) may result in an acceptable identification result. The helix classification agreement was as high as 0.76 between ValgusHel-geometry and ValgusHel-topology. ValgusHel can be used to identify and characterize kinked and curved α -helices. ValgusHel-topology may bring better results and higher consistency compared to using geometric method alone. This method may be useful in future studies to examine structure-function relationships in kinked α -helices.

4.2 Introduction

α -helix is the most common type of protein secondary structure [64, 65]. Previous studies have found α -helix do not always have a straight helix axis [66-69]. Hall et al. found that 44% of a total of 405 transmembrane helices were kinked [70]. Blundell et al. found that most α -helices were curved [67]. Evidence shows that solvent induced distortions [67], peptide bond distortions [68], or proline residues [71] are possible reasons for kinked α -helices. The underlying causes of helical kinks have not been fully revealed. The points provided by the helical kinks easily allow for conformational changes and structural variations, so they usually have important functions in proteins [72-75]. Evidence shows kinks in α -helices have important biological functions. For example, Law et al. (2016) found that changes in kinks could be related to the binding of agonists or antagonist in G-protein-coupled receptors (GPCRs) in receptor activation through conformational change [76]. Kinks in transmembrane helices provide various sizes, shapes, and electrostatic properties of ligand binding pockets in different GPCR subfamilies [75]. Studies found that a kinked α -helix could function as a funnel in ion-channels [77, 78]. However, helical kinks are often neglected in the discussion and not being annotated or misleadingly annotated in Protein Data Bank (PDB) [79].

Some algorithms have been developed to identify kinks in helices. HELANAL-Plus sorts helices into kinked, curved, or linear helices via estimated bend angles formed by two helix axes, each formed by a set of four consecutive α -Carbon atoms within the helix [66, 80]. Prokinked is a highly specific protocol to evaluate proline induced distortions in helices via determining helical axes [81], and it has limited sensitivity to

detect kinks introduced by non-proline residues [66]. MC-HELAN detects and characterizes helical kinks through a Monte Carlo approach determined helical axes. In MC-HELAN, a helix will not be characterized if the algorithm is not converging to a single position [66]. Comparison of algorithms sensitivity to detect kinked helices between HELANAL-Plus and MC-HELAN showed relatively low agreement. For example, in the same data set containing 842 helices, HELANAL-Plus sorted 275 (33%) helices into kinked helices while MC-HELAN sorted 516 (61%) helices into kinked helices [66]. Kinked Finder fits a cylinder over six residues with five residue overlapping and characterized a helix as 'kink' if helix axes angles $>20^\circ$ [82]. AH^AH, a web based survey investigating how humans provide a different to determine if protein α -helices are kinked, provides another perspective to compare the kinked α -helices identification with the above computer algorithms [83]. In this background, additional algorithms to identify kinked helices are needed to provide more knowledge on helical kinks structure and their characterization.

Most studies investigated kinked α -helices inside the environment of the α -helices. Evidence shows that proline is a powerful sequence signature of helical kinks [70, 79, 82, 84-87] although up to two-thirds of kinked helices do not have proline [70, 79]. Hall et al. found that Ser, Thr, and Gly are common in kinked helices [70]. Langelaan et al. observed changes in prevalence of other polar residues within helices [79]. Sequence preferences in kinked helices have been used to predict the presence of kinked helices (e.g., TMKink) [87]. Langelaan et al. found that initial attempts to predict membrane protein kinks using only the protein sequence were unsuccessful [69].

Whether or how the surrounding amino acid residues impact the kinked α -helices has been largely unexplored. Not limited to the residue prevalence within α -helices but including nearby residues around the helical kinks may improve the characterization of helical kinks. Kinks are generally conserved despite changes in sequence, which indicates that the characterization of the kinked helix should consider both local sequence effects and more global interactions with neighboring helices [76, 84].

Kinked helices can be viewed as topological changes from normal helices with straight axes. Examining the differences of topological characteristics between kinked and normal helices may be useful to identify and characterize kinked helices. Delaunay tessellation has been used to define each amino acid residue's nearest neighbors in three-dimensional (3D) protein structures [43]. Each amino acid is abstracted to a point, and the protein structure is then tessellated to form a set of non-overlapping, irregular, space-filling tetrahedra, whose vertices form a Delaunay simplex [43]. The four residues connected by a simplex edge are considered nearest neighbors [43]. However, to the best of our knowledge, Delaunay tessellation has not been used to identify and study kinked α -helices' nearest neighbor residues. In a pilot study, we found that among simplices type 211 [43] whose four residues are all within α -helices, 97.5% are formed by residues i , $i+3$, $i+4$, $i+7$, where i is the residue number. Considering on average 3.6 residues form a helix turn, i and $i+7$ are not likely to become nearest neighbors in a standard helix but possible if the helix is bent. We further visualized these locations in PyMOL software and found there was a bend in the middle of these α -helices. We hypothesized that the bend may decrease the distance between residue i and residue $i+7$ so that there may

become nearest neighbors. Additionally, the distribution of Delaunay simplex types represents structure characteristics of α -helices, and we are interested to see if this distribution is associated with helix angle of α -helices and whether the distribution can be used to predict the helix angle thus helix classification (kinked or normal).

We propose the “ValgusHel” method, which includes both a geometry and a topological method, to identify and characterize kinked α -helices. ValgusHel-geometry calculates helix angles and then identify and classify α -helices based on helix angles. Based on the classification results, we examined the sequence and structure characteristics within and across different α -helices groups based on helix angles. ValgusHel-topology uses topological characteristics obtained from Delaunay tessellations to identify and characterize kinked α -helices. More topological characteristics of kinked α -helices were explored. We further explore the agreement in identifying kinked α -helices between ValgusHel-geometry and ValgusHel-topology in this study.

4.3 Materials and Methods

4.3.1 ValgusHel Method Definition

“ValgusHel” is a collection of a geometry method and a topological method to identify and characterize kinked α -helices.

4.3.1.1 Part 1: ValgusHel-geometry

“ValgusHel-geometry” is the method of dividing an α -helix into a set of six consecutive amino acid residues, three amino acid residues overlapping sextuplet. Each sextuplet forms a cylinder observation unit. In an α -helix containing n ($n \geq 9$) amino acid residues, we included a set of six consecutive amino acid residues $i, i+1, i+2, \dots, i+5$, to

form a cylinder j within the α -helix. The last three amino acid residues, $i+3$, $i+4$, and $i+5$, together with amino acid residues $i+6$, $i+7$, and $i+8$, are included to form the next cylinder $j+1$. We calculated the helix angle between the cylinder j 's axis and the cylinder $j+1$'s axis (Figure 4-1). Since the amino acid residue $i+4$ is at the midpoint of the nine consecutive amino acid chain (residue i to $i+8$), we assigned the helix angle to this midpoint amino acid residue $i+4$. We categorized amino acid residue $i+4$ as a “normal residue” if its helix angle is $\leq 19^\circ$, a “curved residue” if the helix angle is $19-30^\circ$, a “kinked residue” if the helix angle is $>30^\circ$. The sequence containing residues i to $i+8$ was categorized as “normal sequence”, “curved sequence”, and “kinked sequence” based on the center residue $i+4$ helix angles. In a whole α -helix sequences that may contain multiple ValgusHel-geometry observation units, we categorized it as normal α -helix (with all helix angles $\leq 19^\circ$), a curved α -helix (with the greatest helix angles $19-30^\circ$), or a kinked α -helix (at least one helix angle $>30^\circ$).

In “ValgusHel-geometry”, we chose the number of “6” and “3” in a balance of observation units and detection sensitivity after trials of different combinations in the pilot study. If a cylinder is formed by many residues (e.g., 10), we may be able to find bent axis but not be able to accurately locate the bend position. On the other hand, if a cylinder is formed by few residues (e.g., 4), the cylinder axis may not be accurate. Five overlapping residues lead to greatest number of observation units while zero overlapping residue leads to smallest number of observation units. With five overlapping residues, we are concerned that only α -helices with extreme kinked or bend can be detected, which led us to use three overlapping residues.

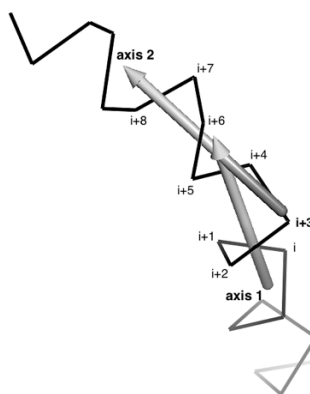


Figure 4-1 Illustration of defining a helix angle using ValgusHel-geometry. Axis 1 is the axis of the cylinder formed by residue i , $i+1$, ..., $i+5$ and axis 2 is the axis of the cylinder formed by residue $i+3$, $i+4$, ..., $i+8$. Residue $i+4$ is the center amino acid residue in residues i to $i+8$, and the calculated helix angle between axis 1 and axis 2 is annotated on it. The fragment containing residue i to $i+8$ is classified as normal α -helix (helix angle $\leq 19^\circ$), kinked α -helix (helix angle $> 30^\circ$), or curved α -helix (helix angle $19-30^\circ$).

4.3.1.2 Part 2: ValgusHel-topology

4.3.1.2.1 Part 2.1: t-numbers

“ValgusHel-topology” is the method to identify kinked α -helices based on the topological characteristics by using Delaunay tessellation. Each amino acid residue was represented by a point, which is collocated with the α -carbons of an amino acid residue. The method further tessellated the protein structure to form a set of non-overlapping, irregular, and space filling tetrahedrons, whose vertices form a Delaunay simplex [43]. Residues contained in a simplex edge are considered nearest neighbors [43]. We divided protein sequences into 9-residue fragments and performed Delaunay tessellation on them. We classified the obtained Delaunay simplices into five types by following the

classification method based on the way the main amino acid sequence chain threads through the simplices proposed by Taylor et al. (2015).[43] Then the obtained Delaunay simplices were represented by simplex type # (0, 1, 2, 3, or 4). We want to know the number of simplices of each type that a residue belongs to. Previous studies called these sums *t-numbers* [14, 27]. For example, if a residue is a vertex in five type 0 simplices, six type 1 simplices, seven type 2 simplices, eight type 3 simplices, nine type 4 simplices, its *t-numbers* are $t(0)=5$, $t(1)=6$, $t(2)=7$, $t(3)=8$, and $t(4)=9$. A residue can participate in different number of simplices with various simplex type distributions. For each residue, combining its *t-numbers* with the helix angles and thus helix classification (kinked, curved, normal) determined by the “ValgusHel-geometry” method, we ran Random Forest Classification in the following format:

Equation 4-1

$$HC \sim t(0) + t(1) + t(2) + t(3) + t(4)$$

where HC is helix classification.

We tested if there were associations between the residue’s *t-numbers* and helix angle classification, and if we could predict helix angle classification based on the *t-numbers*.

4.3.1.2.2 Part 2.2: N-gram

In a protein sequence, we ranked all obtained simplices based on the containing residue number and gave the simplices a stacking index from 1 to n, where n is the total

number of simplices. Next, we annotated each stacking index with simplex type number (0, 1, 2, 3, 4). Therefore, a sequence was represented by a set of simplex type numbers. For example, an example sequence fragment contained 9 simplices, and the sequence's simplex type stacking index is 421421421 in Table 4-1.

Table 4-1 Simplex type stacking index in an example sequence

Stacking index	Simplex composition represented by AA	Vertex1 residue#	Vertex2 residue#	Vertex3 residue#	Vertex4 residue#	Simplex type
1	TGLI	100	101	102	103	4
2	TGIV	100	101	103	104	2
3	TIVE	100	103	104	107	1
4	GLIV	101	102	103	104	4
5	GLVS	101	102	104	105	2
6	GVSK	101	104	105	108	1
7	LIVS	102	103	104	105	4
8	LISL	102	103	105	106	2
9	LSLE	102	105	106	109	1

We proceed with N-gram approach. For each N value (N = 3, 4, 5, ..., 20), we created a N-gram frequency table within normal, kinked, and curved helices within the whole data set, respectively. Then we reformatted the frequency table to reflect the frequency normal, kinked, and curved helices within each combination in each N-gram.

We tried different combinations of fragment length m (8 to 20) and N -gram with N (3 to 20). Take $m=9$ and $N=5$ as an example, we divided a helix sequence by 9 consecutive simplex indexes identified in the previous step (e.g., 421421421). Next, we select 5 consecutive simplices out of the 9-residue fragment, resulting in 5 possible 5-gram: 42142, 21421, 14214, 41242, and 21421. We calculated the scores of normal, curved, and kinked, respectively, as below:

Equation 4-2

$$S_s = \prod_{i=1}^5 f_{i,s}$$

where S is score, f is frequency, which is the calculated frequency from the 5-gram table from the data set, s is the helix structure of normal, curved, or kinked, and i is one of the 5 possible 5-gram. The stacking type for this fragment is determined by the structure type with the highest score. For the example sequence above, normal helices have the highest score, and the 9-residue fragment 421421421 is annotated as N, where N is Normal.

4.3.2 Data set preparation

We created a data set containing 8,826 individual protein chains and 46,604 α -helices for α -helix classification by using ValgusHel-geometry and ValgusHel topology. The dataset is downloaded from PDB with the following criteria: sequence identity < 30%, resolution higher than 2 Angstrom, r-factor ≤ 0.25 , sequence length between 100-1000. This dataset is abbreviated as 8826culled.

4.3.3 Kinked α -helices identification using ValgusHel geometry method

To compare the method sensitivity of ValgusHel-geometry to MC-HELAN, and HELANAL-Plus, we used a published data set containing 140 proteins and 887 α -helices from the MC-HELAN server [88]. Kinked residues identified by the MC-HELAN method are already labeled. In this MC-HELAN data set, we identified kinked α -helices on each protein chain using the ValgusHel-geometry and the HELANAL-Plus method. We compared the kinked α -helices identification results from the ValgusHel-geometry, MC-Helena, and HELANAL-Plus methods.

There are a total of 177 helices with annotations of kinked, curved or straight classifications on AHAH [83]. We performed ValgusHel-geometry and HELANAL-Plus on this AHAH data set to compare the results agreement.

Next, for each observation unit (9 consecutive amino acid residues), we calculated a Root-Mean-Square Deviation (RMSD) by comparing its structure to a model ideal helix which has a straight axis in structure alignment. A lower RMSD suggests higher similarity between the structure of interest and the ideal alpha-helix model. We examined the relationship between RMSD and calculated helix angles.

We applied the ValgusHel-geometry method to classify residues and helices into normal, curved, and kinked residues and helices, respectively by following the criteria described above. We also categorized the nine non-overlapping sequence (e.g., residues i to $i+8$) into normal, curved, or kinked sequences based on their corresponding center residue helix angle. We compared the sequence similarity by using the BLOSUM62 penalty table within and across different sequence normal/curved/kinked groups [89].

4.3.4 Association between sequence and structure in kinked α -helices

Sequences are classified as normal, curved, or kinked sequences based on helix angles as described above. We used Jalview to cluster sequences by using average distance with BLOSUM62 to build a hierarchical tree [90]. Since Jalview limits number of input entries, we randomly selected and input 15,000 out of 95,210 sequences. The six clusters with the greatest number of sequences were included for further analysis. We examined the helix angles within each of the six clusters.

We identified all coils and beta-strands secondary structures with greater or equal to 9 consecutive residues in our data set [91]. Because equal number of sequences between two sequence comparison groups is required, and the kinked sequence group has the lowest sequence number ($N=752$), we randomly selected 752 out of 56,826 sequences in the normal sequence group, 752 out of 37,632 sequences in the curved sequence group, and 752 out of 2,755 of the coiled-coils and out of 38,402 of beta-strands to compare, respectively. Within each group, one sequence is compared to the rest 751 sequences, producing $C_{752}^2 = 282,376$ sequence similarity scores, which are further processed to produce an overall mean similarity score within the group. When comparing two sequence groups, each sequence in group 1 is compared to all the sequences in group 2, producing $752 \times 752 = 565,504$ sequence similarity scores, which are further processed to produce an overall mean similarity score between group 1 and group 2.

We were interested to see whether kinked helices were due to a kink position in the center of the fragment or the whole fragment is kinked. To explore the kink position within the 9-residue fragment described above, in each of the 9-residue sequence

(N=775), we removed the center residue. A pair of two helix fragments, each containing four-residues, were obtained from each 9-residue sequence. We did structure alignment of each pair of the 4-residue helix fragments with a model ideal helix and added up the two RMSD from two fragments.

To find whether Leucine was associated with kinked helices, we also identified coiled-coils and leucine zippers among all the kinked helices by using the methods by Lupas et al. (1991) [91]. We summarized the number of kinked helices that belong to leucine zippers or transmembrane proteins.

We examined whether identified kinked helices would be predicted as helices instead of coils by Jpred 4 [92]. We included all 9-residue kinked alpha-helix fragments and calculated trees using Average distance BLOSUM62 on Jalview [93]. We picked 5 clusters with most similar sequences while allowing each cluster to include 20 to 39 sequences. Next, we found sequence motif for each cluster using MEME [94]. We used the motif regular expression to scan against protein sequence databases of Swiss-Prot and TrEMBL in ScanProsite [95, 96]. In the hit results, we randomly picked 50 hits for each cluster if the hit number is greater than 50. We used Jpred 4 to predict secondary structures of clusters 1-4, where cluster 5 was not included due to low hit number. The 9-residue fragments were matched with corresponding secondary structures predicted by Jpred 4. For each residue (1 to 9), we calculated the frequency of different secondary structures and their average confidence scores.

4.3.5 Prediction of normal, curve, and kinked residues based on variables (DSSP or Delaunay tessellation) in Random Forest Classification

We were interested to see whether using the ‘ACC’, ‘TCO’, ‘KAPPA’, ‘A’, ‘PHI’, ‘PSI’ variables from DSSP would predict the classification of normal vs. non-normal (curve and kink) residue defined by ValgusHel-geometry. Based on the results of ValgusHel-geometry, we coded the center residues $i+4$ as 1 for non-normal (curve and kink) residues and 0 for normal residues. We ran Random Forest Classification (RFC) models to predict whether residues are normal or non-normal using the ‘ACC’, ‘TCO’, ‘KAPPA’, ‘A’, ‘PHI’, ‘PSI’ variables as independent variables. While DSSP variables are within the helices, we also explored whether the surrounding residues to the normal/curve/kinked residues are useful in predicting the helix angles defined by ValgusHel-geometry. Using Delaunay tessellation, we tessellated the entire α -helices and located the nearest neighbor residues to them. No simplices with edge length >12 angstrom was included. Residue distance and edge length, along with amino acid type, sequence secondary structure, amino acid locations (buried or on surface), are included as independent variables in the RFC models to predict the helix angle and thus the classification of normal vs. non-normal residues.

4.3.6 Kinked α -helices identification using ValgusHel-topology method

We performed the ValgusHel-topology method on our data set with a combination of fragment length ($m=8$ to 20) and N-gram ($N=3$ to 20). We annotated each fragment with N (Normal), K (Kinked), or C (Curved) based on the results.

We created a confusion matrix with the number of kink, curved, and normal helices identified by ValgusHel-geometry and ValgusHel-topology on our data set. We calculated sensitivity, specificity, and Balanced Accuracy Rate (BAR) to evaluate the agreement.

4.4 Results and discussion

4.4.1 Characteristics of calculated helix angles

In the 8826culled dataset, we identified 46,615 α -helices by using the ValgusHel-geometry method. We annotated a total of 278,010 residues with helix angles and their distribution is shown in Figure 4-2. There are 257,853 (92.75%) normal residue, 17,536 (6.31%) curved residue and 2,621 (0.94%) kinked residues. There are 34,575 (74.17%) normal α -helices, 10,827 (23.23%) curved α -helices, and 1,213 (2.60%) kinked α -helices.

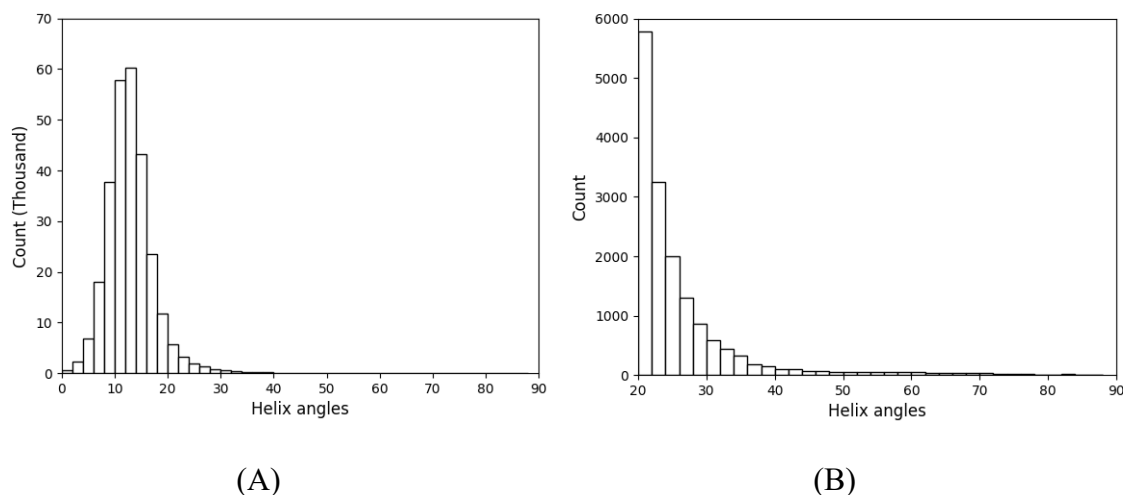


Figure 4-2 Helix angles distribution by using ValgusHel method. A total of 278,010 helix angles were estimated in Our data set (Figure 2A). Figure 2B shows the zoom in results when helix angles $\geq 20^\circ$.

Comparing with a model ideal α -helix which forms a cylinder with straight axis and only by Alanine. Using six residues to form a cylinder, we found that even a model ideal helix has a helix angel of 11.91° . RMSD increase monotonically when the calculated helix angles increase from approximately $14-46^\circ$ and fluctuated when the calculated helix angles increased from approximately $46-70^\circ$ (Figure 4-3). The dip of RMSD occurred when the helix angle was approximately $12-14^\circ$. We further divide non-normal helices, fragments and residues into curve (helix angle $19-30^\circ$) and kinked (helix angle $>30^\circ$). Figure 4-3 shows that higher helix angles defined by ValgusHel-geometry are associated more structure deviations from a model ideal helix with a straight axis,

bringing more validation evidence of the method. In Figure 4-3, the mean RMSD increased monotonically when the helix angles increased from 14-46°, agreeing with our assumption that greater helix angles are associated with more “bend” of the helix barrel and more deviation from the standard normal helix barrel. Theoretically, an α -helix has 3.6 residues per turn [97], and using 6 residues to define axis may lead to slightly different axis in a model ideal helix. In fact, we estimated that a model ideal helix has a helix angle of 11.9°. Therefore, we deduced that a helix angle close to 11.9°, either higher or lower, was more likely to be within a straight/normal helix. The fluctuation of RMSD when helix angles increased from 46-70° may be due to small sample size (Figure 4-3).

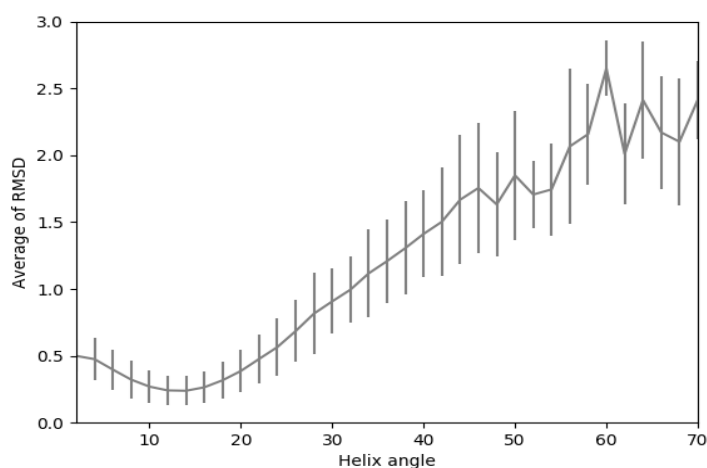


Figure 4-3 The relationship between Root Mean Square Deviation (RMSD) and calculated helix angles (°) by using ValgusHel-geometry. Error bars are standard deviations.

4.4.2 Validation of ValgusHel geometry method against HELANAL-plus, MC-HELAN, and AHAH

In the data set from the containing 887 α -helices from the MC-HELAN server. The kinked α -helix identification agreement was 45.5% between ValgusHel and HELANAL-Plus, 67.1% between ValgusHel and MC-HELAN, and 58.9% between HELANAL-Plus and MC-HELAN. The agreement among the three methods was 45.7% (Figure 4-4). Figure 4-5 demonstrates the α -helices identified as kinked ones by ValgusHel-geometry, HELANAL-plus, and AHAH in the data set provided on AHAH website, respectively. A total of 48.6% of all α -helices were identified as kinked α -helices by all the three methods. A total of 61.0% of all α -helices were identified as kinked ones by both ValgusHel-geometry and HELANAL-plus, and 65.5% by both ValgusHel-geometry and AHAH (Figure 4-5). Relatively fair percent of the kinked α -helices identified by ValgusHel-geometry were also identified by the MC-HELAN and HELANAL-Plus methods (Figure 4-4). Although there is no “gold standard” of defining a kinked α -helix, the relatively high agreement with MC-HELAN and HELANAL-Plus supports the validity of ValgusHel-geometry. The moderate agreement between AHAH and ValgusHel-geometry is comparable to the agreement between AHAH and HELANAL-Plus (Figure 4-6), suggesting ValgusHel-geometry brings equivalent alignment with human perspective in identifying kinked α -helices compared to HELANAL-Plus.

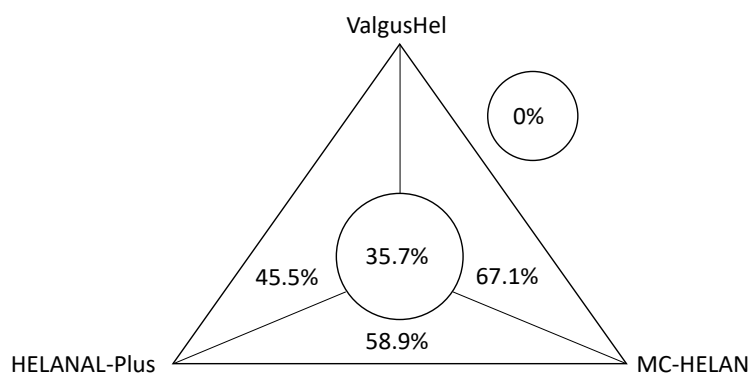


Figure 4-4 Number of α -helices categorized as kinked α -helices by using ValgusHel-geometry, HELANAL-Plus and MC-HELAN methods in a data set provided on the MC-HELAN server. The total number of α -helices was 887. The edges of the triangle represent the agreement of α -helices classified as kinked α -helices by two methods on the vertices. The circle in the middle represent the agreement of α -helices classified as kinked α -helices by the three methods.

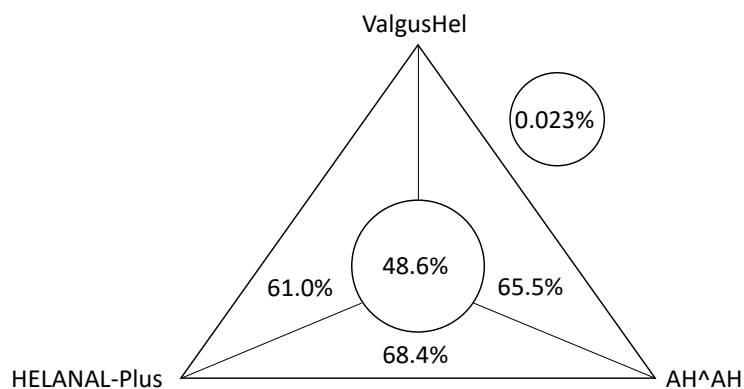


Figure 4-5 Percent of α -helices categorized as kinked α -helices by using ValgusHel-geometry, HELANAL-Plus and AHAH methods in the data set provided on AHAH website. The total number of α -helices was 177. The edges of the triangle represent the agreement of α -helices classified as kinked α -helices by two methods on the vertices. The circle in the middle represent the agreement of α -helices classified as kinked α -helices by the three methods.

4.4.3 Amino acid residues frequency in different α -helices groups

Figure 4-6 illustrates difference of amino acid frequencies between kinked residues/helices and normal residues/helices. We found that the difference frequency (%) is highest in Leucine (L), Glycine (G), Alanine (A) and lowest in Valine (V), Isoleucine (I), and Glutamic acid (E) at the residue level (Figure 4-6A). Interesting to observe L and V are on the top opposite site, though they are similar amino acids. Proline (P) had highest difference of frequency and A had the lowest one at the helices level (Figure 4-6B). In Figure 4-6B, the high helix angles annotated at prolines residues agrees with the previous studies suggesting that Proline (P) may be a powerful sequence signature of helical kinks [70, 79, 82, 84-87]. However, the low prevalence of Proline in kinked helices suggests Proline is not necessary in kinked helices. Yohannan et al. (2003) proposed mutation to proline initially induces kinks in transmembrane proteins, but the mutation was replaced by further mutation but the kinked structure stays [84].

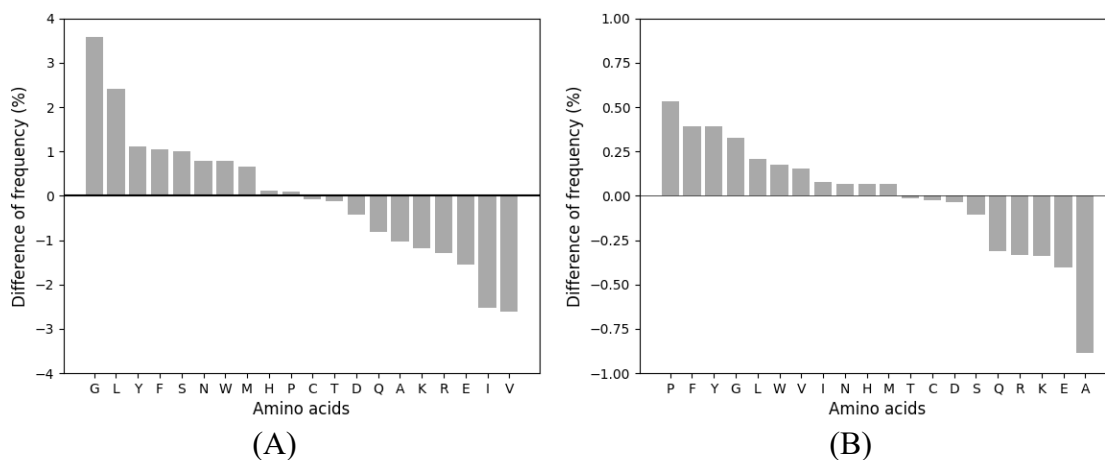


Figure 4-6. Difference of amino acid frequencies (%) in (A) all residues and (B) all helices. Difference of frequency (%)=frequency (%)_i in kinked residue/helix group-frequency (%)_i in normal residue/helix group. i (1 to 20) represents 20 natural amino acids.

4.4.4 Sequence similarity in different cluster and α -helices groups

Based on sequence similarity scores within and across sequence clusters from BLOSUM62, the distributions of sequence helix angles are similar across clusters (Figure 4-7). Table 4-2 reports the sequence identity and helix angles after clustering the sequences first. In the sequence identity degree between two sequences <0.5 group, the Pearson's correlation coefficient was 0.372 between the helix angles. A total of 33.6% of all pairs of sequences had the same helix angles. The results were higher (0.428 and 38.1%) in sequence identity degree between two sequences >0.5 group, where the sequences were more similar to each other, although the difference was relatively small.

Figure 4-7 shows different sequence cluster groups have similar helix angle distribution, suggesting sequence may not strongly correlated with helix angles. However, in Table 4-2, we observed higher Pearson's correlation coefficient in groups with sequence identify degree >0.5 compared to those with sequence identify degree < 0.5 (0.428 vs. 0.372). Combining results from Figure 4-7 and Table 4-2 together, we think that helix sequences may be correlated with helix angles but the correlations may not be strong.

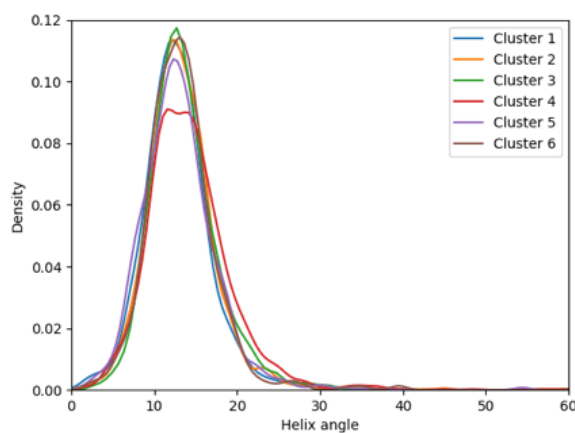


Figure 4-7 Density of sequence helix angles in each of the six sequence clusters. The calculated helix angles ($^{\circ}$) were obtained by using ValgusHel-geometry. Cluster 1-6 were top sequence clusters obtained from Jalview based on sequence similarity. Since Jalview limits number of input entries, we randomly selected and input 15,000 out of 95,210 sequences.

Table 4-2 Sequence identity and helix angles by clustering sequences

Sequence identity degree between two sequences	Number of sequences	Pearson's correlation coefficient	Percent of two sequences with same helix angles (%)
<0.5	30,396	0.372	33.6
>0.5	30,396	0.428	38.1

Table 4-3 reports the sequence similarity score of sequences comparing with an ideal α -helix. Sequences in the normal group (center residue helix angle $\leq 19^\circ$) had the highest similarity score of -6.814 (SD 6.683) while sequences in the kinked group (center residue helix angle $>30^\circ$) had the lowest similarity score of -8.132 (SD 6.807). Sequence similarity scores were higher in coiled-coils and β -strands than kinked sequences, which may suggest large diversity in kinked helices sequences. When we compared sequence similarity scores across groups, we found that similarity score between normal vs. curve (-6.986) and between normal vs. kinked (-7.956) are lower than that within normal (-6.814), suggesting differences existed in sequences between normal vs. curved sequence groups and between normal vs. kinked sequence groups, with the degree of difference greater in the normal vs. kinked comparison. In Table 4-3, the finding that sequence similarity scores are highest within normal groups and lowest within the kinked group suggest that sequences in kinked group were more diverse. There may be multiple patterns of sequences leading to the kinked structure. In the motif regular expression patterns, we observed some differences among kinked vs. normal helices (Figure 4-8, Figure 4-9, and Table 4-5). We found Proline or Glycine positioning in close to center

locations may be a signal for kinked helices while Proline, W, and D on the two ends of the 9-residue sequence fragment may be a signal for normal helices. We found some interesting results in the secondary structures predicted by Jpred 4. The frequencies of helix were lower at residue number 6-9 among kinked helices compared to normal helices (Figure 4-10, Figure 4-11). We are unclear why this happened. One possible reason is that some kinked α -helices were categorized as coils by Jpred due to their similar structures. For example, a short helix connected to a coil may be identified as a single helix structure with a kink/curve.

Table 4-3 Mean sequence similarity score with SD of different sequence groups by clustering helix angles. N=752 in each group.

	Normal	Curve	Kink	β -strand	Coil
Normal	-6.814 (6.683)				
Curve	-6.986 (6.766)	-7.195 (6.851)			
Kink	-7.956 (6.559)	-7.982 (6.634)	-8.132 (6.807)		
β -strand	-7.976 (6.182)	-8.095 (6.198)	-8.788 (6.273)	-7.933 (6.535)	
Coil	-8.661 (6.084)	-8.796 (6.138)	-9.123 (6.205)	-9.374 (6.218)	-7.982 (6.702)

Note: SD = standard deviation. Significant difference was observed between any two comparison (p-value =0).

Table 4-4 reports the sequence identify score of sequences comparing with an ideal α -helix. Normal helices had the highest sequence identify score (0.073) while kinked helices and β -strand had the lowest score (0.066). Overall, the trend was similar in Table 4-3 and Table 4-4.

Table 4-4 Mean sequence identity score with SD of different sequence groups by clustering helix angles. N=752 in each group.

	Normal	Curved	Kinked	β -strand	Coil
Normal	0.073 (0.088)				
Curved	0.069 (0.086)	0.068 (0.085)			
Kinked	0.066 (0.083)	0.065 (0.083)	0.066 (0.084)		
β -strand	0.062 (0.081)	0.062 (0.080)	0.059 (0.079)	0.066 (0.083)	
Coil	0.056 (0.077)	0.056 (0.076)	0.055 (0.076)	0.055 (0.077)	0.067 (0.084)

Note: SD = standard deviation.

The clustering results and obtained motif regular expressions in kinked α -helices from average distance BLOSUM62 on Jalview are presented in Table 4-5. Among the 775 identified kinked α -helices, the motif regular expressions from the 5 clusters are also presented in Figure 4-8. We observed Proline (P) and Glycine (G) was dominate in the middle positions of 6 or 7. Clustering and motif regular expression results in normal α -helices are presented in Table 4-5 and Figure 4-9. A total of 8000 normal α -helix

fragments were included in the clustering, and only the randomly selected 8 clusters were shown in Figure 4-9.

Table 4-5 Clustering results of α -helix fragments (N=775)

α -helix	Clusters	Number of sequences	Motif	Number of hits
Kinked	1	20	LEA[LI]AP[LY][VI]D	17
	2	20	[IL]XEX[LM]E[KR]YV	50
	3	21	Y[LY]EK[HY]L[DE]E[YF]	18
	4	39	LAXXLXP[IL][LI]	40
	5	21	LLAEHGEEG	2
Normal	1	29	LEX[LI]Q[KQ][LI][IV]D	50
	2	40	LPE[LI]XE[ILA][LI]A	43
	3	29	E[AD]D[FL][VL]K[ILV][IL]N	47
	4	37	EEL[LV]K[KE]L[KE]E	39
	5	44	WDX[IA]XAX[LV]E	28
	6	25	[IV][LVI]X[DE]A[LI][KE][EA]A	35
	7	24	[DE][AE][AR][AE]A[LIV]XRW	41
	8	24	[EP]W[AL]KEILKQ	2

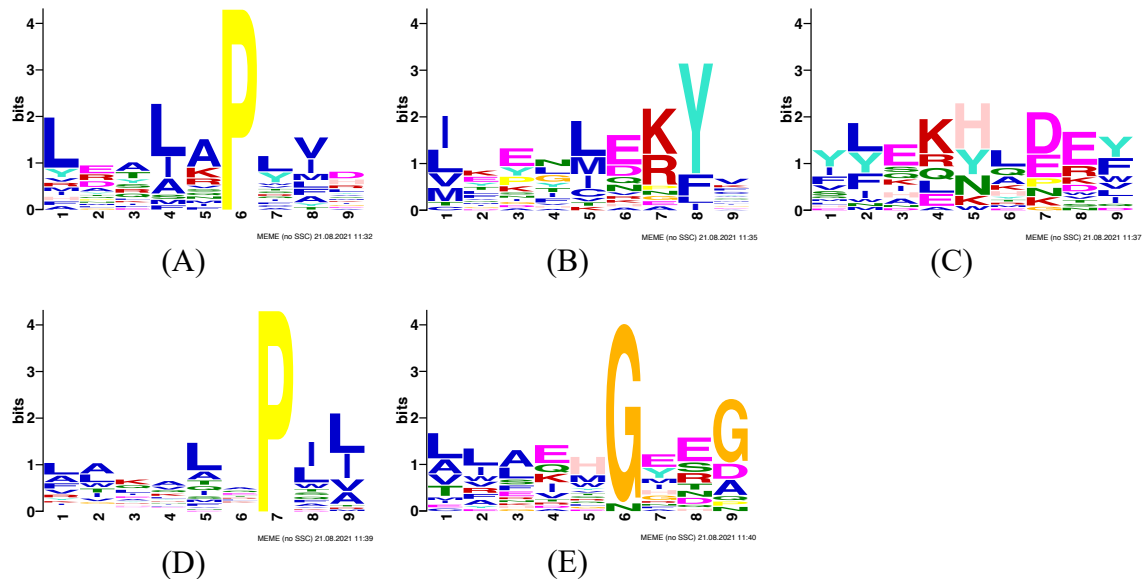


Figure 4-8 Kinked helix cluster motif logo. (A)-(E) are from clusters 1 to 5 in Table 4-5. Numbers in x-axis are residue number within the 9-residue fragment. Bit score in y-axis is measurement of certainty. Higher bit score indicates higher certainty to observe the amino acid at the residue position in x-axis.

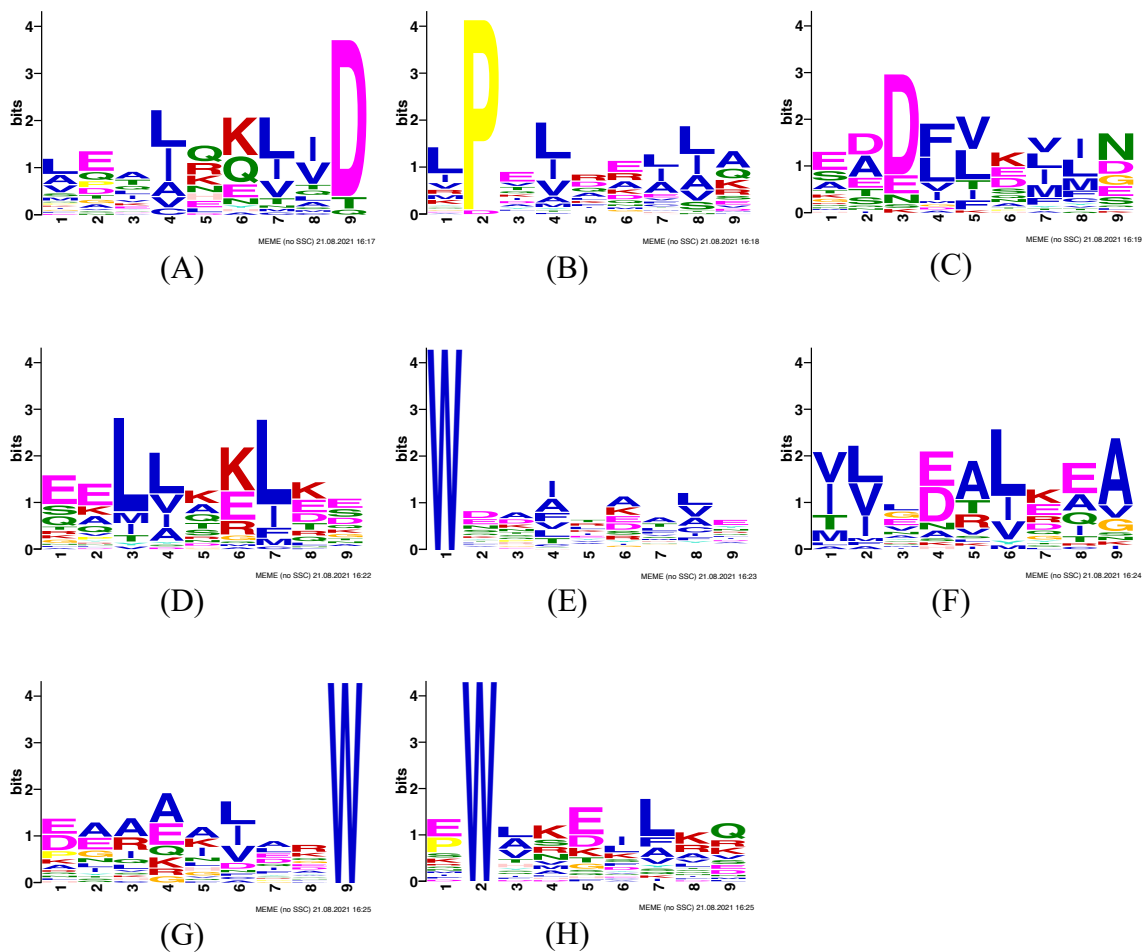


Figure 4-9 Normal helix cluster motif logo. (A)-(H) are from clusters 1 to 8 in Table 4-5. Numbers in x-axis are residue number within the 9-residue fragment. Bit score in y-axis is measurement of certainty. Higher bit score indicates higher certainty to observe the amino acid at the residue position in x-axis.

We obtained “hits” by searching the motifs in Table 4-5 against the protein sequence databases. The secondary structure predictions of these “hits” made by Jpred 4 showed different frequencies of helix at residues 5-9, where Jpred 4 predicted higher

frequencies of coils in kinked α -helices compared to normal α -helices (Figure 4-10). We further examined the helix frequency among normal vs. kinked α -helices at residues 1 to 9, and results are presented in Figure 4-11. Normal and kinked α -helices were predicted to have similar helix frequencies at residues 1-5 while normal α -helices had wider range of helix frequencies. However, at residues 6-9, we found kinked α -helices had lower predicted helix frequencies with wider range compared to normal α -helices, consistent with results in Figure 4-10.

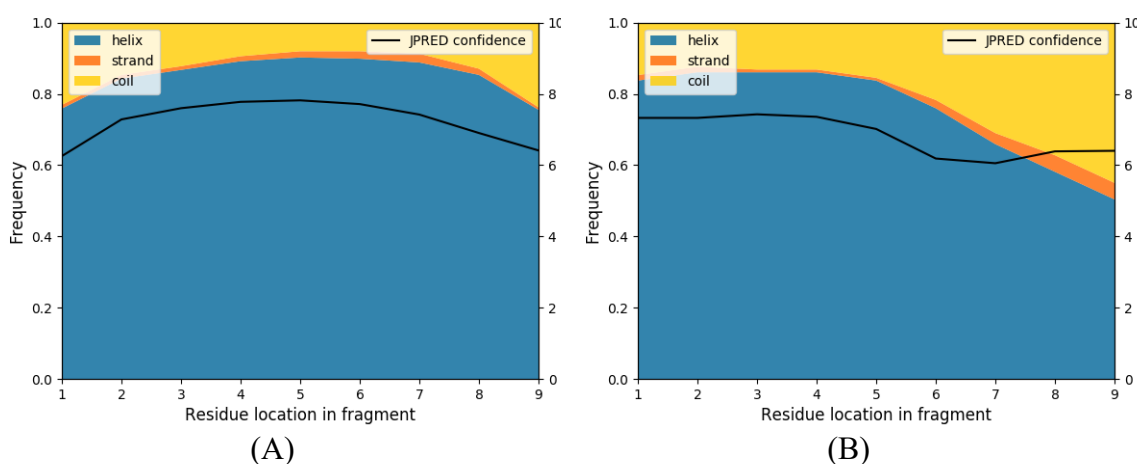


Figure 4-10 Predictions made by Jpred 4 among (A) normal helix clusters and (B) kinked helix clusters. Numbers in x-axis are residue number within the 9-residue fragment. Jpred 4 confidence uses the y-axis scale on the right (scale 0 to 10).

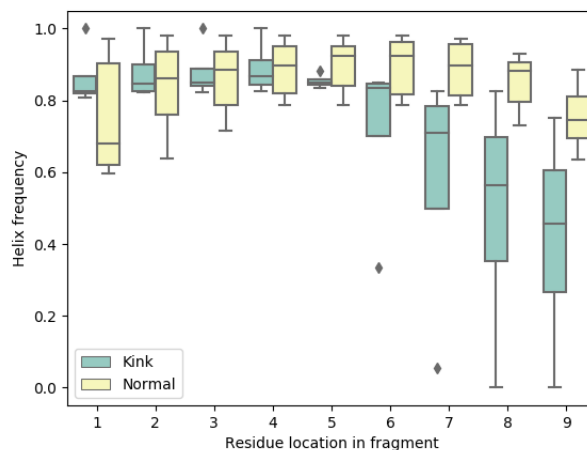
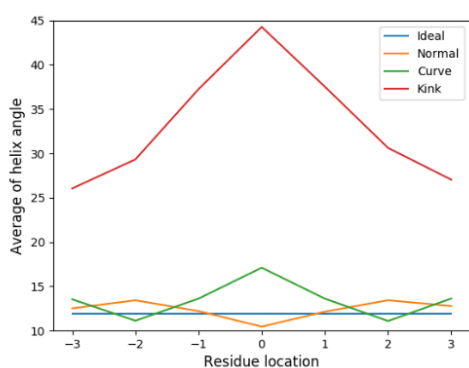


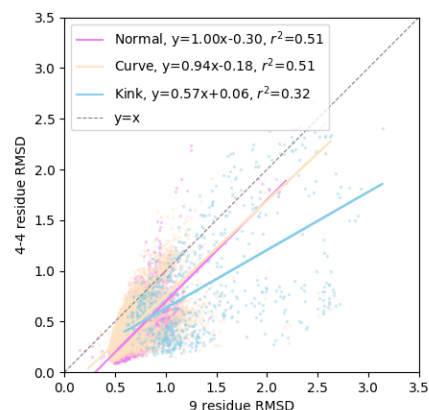
Figure 4-11 Boxplot of helix clusters among kink and normal helix clusters. Median, 25th percentile (Q1), 75th percentile (Q3), minimum (Q1 – 1.5IQR) and maximum (Q3 +1.5IQR) are shown. Outliers are shown as dots. IQR = interquartile range.

Figure 4-12A shows the average of helix angles at different residue locations among the ideal, normal, kinked, and curved α -helices, respectively (Figure 4-12A). Residue location 0 is the center residue in the 9-residue sequence fragment. We observed highest average of helix angles in kinked α -helices at residue 0 and lower helix angles as the residue number moves from 0 to ± 3 , suggesting the kinked position is more prevalent at center residue. The trend was similar in curved α -helices. The mean RMSD of the 9-residue sequence (N=775) after taking out the center residue are shown in Figure 4-12B-C. A total of 101 (13%) 9-residue sequences were divided into two short normal helices (sum RMSD<0.2), 669 (86%) were divided in one normal and one kinked helices (sum RMSD 0.2-0.5), and 5 (1%) were divided into two kinked helices (sum

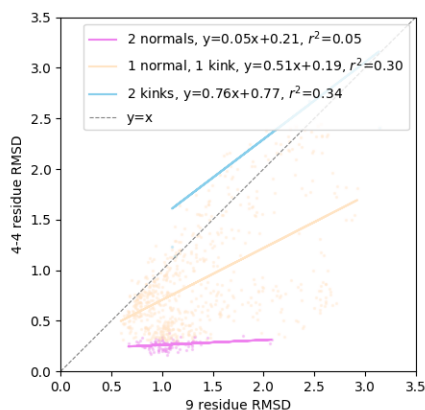
RMSD>0.5). In Figure 4-12B, we found that the sum of RMSD from two 4-residue fragment is lower than the RMSD from the 9-residue fragment, suggesting the center residue may be the kink position. In Figure 4-12C, we found that when the sequence was divided into two normal helix fragments, the RMSD dropped, which may suggest the center residue is the only kinked location. When the sequence was divided into one normal and one kinked helix fragments, RMSD also dropped, and a remaining kinked 4-residue helix fragment may suggest there are more than one kinked location within the 9-residue sequences. The two kinked 4-residue helix fragments group may suggest at least three kinked locations within the 9-residue sequences (Figure 4-12C). In Figure 4-12, we attempted to locate how many kinked positions within a 9-residue observation unit using ValgusHel-geometry. While Figure 4-12B suggests that kink position were likely at the center residue position in the 9-residue sequence fragment, Figure 4-12C suggestions there may be more kink positions except the center residue. A total of 87% kinked sequences have at least one 4-residue kinked fragments, suggesting more than one kinked position in the 9-residue sequence (Figure 4-12C).



(A)



(B)



(C)

Figure 4-12 The relationship between a 9-residue sequence RMSD and the sum of two 4-residue sequence RMSD. In the 9-residue sequence (N=775), the center residue is removed, producing two sequence fragments, each containing 4 residues. The 9-residue sequence is aligned and compared with a model ideal helix. (A) average of helix angles at different residue locations among ideal, normal, kinked, or curve α -helices. Residue 0 is the center residue of the 9-residue sequence fragment. (B) center residue 0 was removed from the 9-residue sequence fragment, resulting in two 4-residue fragments. The sum of the RMSD from the two 4-residue fragment is the y-axis. Color dots and lines represented the classification of the 9-residue sequence. (C) center residue 0 was removed from the 9-residue sequence fragment, resulting in two 4-residue fragments. The sum of the RMSD from the two 4-residue fragment is the y-axis. Color dots and lines represented the classification of the two 4-residue fragments.

We found a total of 230 coiled-coils and 18 leucine zippers out of 46,615 helices. The low number of coiled-coils and leucine zippers may suggest they are not the main reasons behind the kinked helices.

4.4.5 Kinked and normal α -helices prediction using DSSP and Local profile

Table 4-6 reports the random forest classification (RFC) or random forest regression (RFR) performance in predicting helix angles thus kinked helices by using the DSSP variables including N-H-->O, O-->H-N, N-H-->O, O-->H-N, TCO, KAPPA, ALPHA, PHI, PSI. The performance of RFC is good and balanced with a sensitivity of 0.89 and a specificity of 0.90. RFR had a r of 0.81 with a standard deviation of 0.02. In Table 4-6, our kinked α -helices classification results were well predicted by using the DSSP variables with a 0.89 sensitivity and 0.90 specificity, supporting that ValgusHel-geometry is a valid method in α -helices identification and classification.

Table 4-6 DSSP random forest 10-fold, evenly data set (n=15,487 for each class).

Model	Dataset	Precision	Recall	F-Measure	MCC	ROC Area
RFC	Original	0.890	0.890	0.890	0.781	0.951
	Shuffled	0.502	0.502	0.502	0.004	0.504
Model	Dataset	r	RMSE			
RFR	Original	0.799	6.042			
	Shuffled	-0.01	10.328			

Table 4-7 reports the RFC and RFR performance in predicting helix angles thus kinked helices by including residue distance and edge length of the nearest neighbors of the kinked residues. When using RFC, a lower sensitivity of 0.75 and a specificity of 0.75 were obtained compared to using DSSP variables in Table 4-6. RFR had a r of 0.568 with a RMSE of 7.714. Table 4-7 explores the use of residue neighbors' characteristics around the kinked residues to predict the identification of kinked and curved α -helices. The results in Table 4-7 suggest that nearest neighbors' information, along with the centered amino acid information, may be useful in predicting helix angles by using random forest classification or random forest regression. Unlike DSSP variables, the characteristics from nearest neighbor residues help us explore the how predictive the surrounding protein structure information are when predicting the kinked residues.

Table 4-7 Local profile random forest 10-fold, evenly data set (n=11,747 for each class).

Model	Dataset	Precision	Recall	F-Measure	MCC	ROC Area
RFC	Original	0.750	0.750	0.750	0.500	0.830
	Shuffled	0.5000	0.500	0.500	0.000	0.500
Model	Dataset	r	RMSE			
RFR	Original	0.568	7.714			
	Shuffled	-0.007	8.876			

Note: RFC = random forest classification, RFR = random forest regression, MCC = Matthew's correlation coefficient, RMSE = Root Mean Square Error.

4.4.6 Random forest classification using Delaunay simplex type descriptor (t-numbers)

Figure 4-13 shows simplex type distribution within α -helices. Simplex type #4 had the highest prevalence in normal, curved, and kinked helices, followed by simplex type 2 and 1. Simplex type 0 was only observed in kinked helices, and simplex type #3 was dominated by kinked helices. By using ValgusHel-topology, in Figure 4-13, we found that the types of Delaunay simplices had different distributions in kinked, curved, and normal helices. For example, most helices in simplex type #0 and #3 are kinked. We used these simplex type distribution (t-number) feature to train models with the helix classification of kinked, curved, or normal obtained from the ValgusHel-geometry part as the outcome.

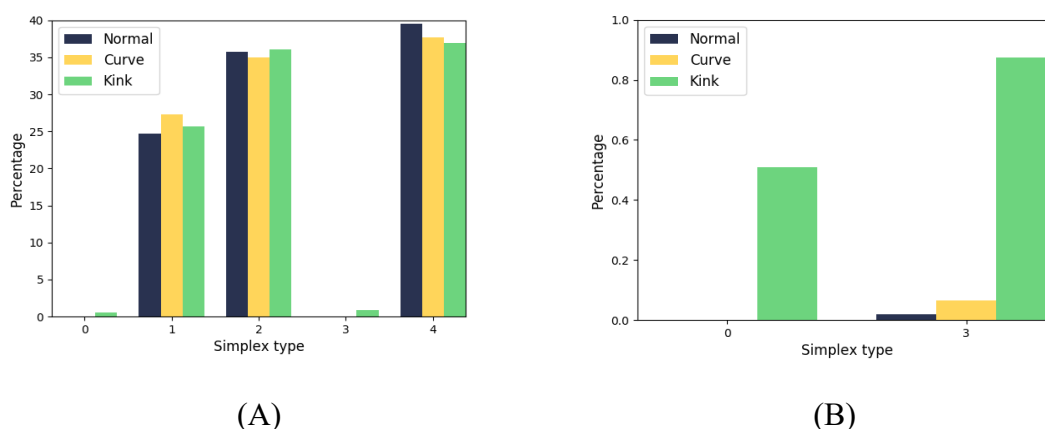


Figure 4-13 Percentage of simplex types in normal, curved, and kinked helices. Simplex types (0, 1, 2, 3, and 4) are adopted from Taylor et al. (2015). Figure 8(B) is a zoom in for Figure 8(A) to show details of percentage lower than 1%.

We found that using simplex type numbers (t-numbers) in a helix to predict whether the helix was normal, kinked or curved brought about reasonable results (Table 4-8). The sensitivity to detect kinked and curved helices were 0.78 and 0.79, respectively. The specificity to detect kinked helices was highest of 0.82. We further predict the helix classification based on the simplex type distribution within a helix. We found the model was sensitive to detect kinked helices with a sensitivity of 0.78 and curved helices with a sensitivity of 0.79 (Table 4-8). In terms of specificity, we found that the model had the highest specificity of 0.82 to detect kinked helices and 0.54 to detect curved ones (Table 4-8). The lower specificity to detect curved helices were expected, as it is difficult to differentiate between kinked and curved helices. Considering these criteria values are

acceptable, we think the distribution of types of Delaunay simplex may be useful to identify kinked, curved, or normal helices.

Table 4-8 10-fold performance results for each class using random forest classification

	Sensitivity	Specificity	BAR
Kinked	0.78	0.82	0.80
Curved	0.79	0.54	0.67
Normal	0.96	0.71	0.83

4.4.7 Agreement between ValgusHel-geometry and ValgusHel-topology methods using N-gram

Figure 4-14 demonstrates the agreement between topological and geometrical methods to identify normal, kinked, and curved helices. True Positive (TP) is the number of α -helices classified as kinked α -helices by both the ValgusHel-geometry and ValgusHel-topology. Sum is the number of total α -helices included. Higher True Positive (TP)/Sum indicates higher agreement between ValgusHel-geometry and ValgusHel-topology. The ValgusHel-topology N-gram results had high (~ 0.76) agreement with ValgusHel-geometry method results. The effect of fragment length (8 to 20) in combination with different N-grams (4 to 20) was also demonstrated in Figure 4-14. We compared the result agreement of ValgusHel-geometry part and topology part and

presented the results in Figure 4-14. We found that a combination of fragment length ($m \geq 12$) and N-gram ($N \geq 8$) may result in an acceptable True Positive/SUM. We observed the agreement is as high as 0.76 using this combination. The agreement was moderately high considering the agreement among other methods was relatively low (Figure 4-4, Figure 4-5). The subjectivity of kinked helices identification and the complexity in the structure of kinked helices were possible reasons for not very high agreement. The effect of different combinations of fragment length and N-grams was demonstrated in this figure, which also served as evidence for selecting m and N in future studies. Based on the results, we think a fragment length of 12 and N number of 8 may be a start point for future studies.

Overall, this study explores a new method to identify and study kinked helices geometrically and topologically. As the definition of “kinked” is still largely subjective, we will validate our approach with more published methods as more helix structures become available. We addressed the limitation of selecting “6” and “3” in ValgusHel-geometry in the Methods part. We tried different combinations of different residues forming cylinders and different overlapping residues but chose 6-3 mainly for the consideration of kinked detection sensitivity. We applied the new method in a relatively large data set and explored the association between kinked structures and sequences, and between kinked structures and its nearest 3D neighbor residues. ValgusHel-topology offers an opportunity to identify kinked or curved helices without relying on measuring geometry angles. Delaunay tessellation can be used to aid in the identification of normal, curved and kinked α -helices from a topological viewpoint, yielding comparable results to

the geometry method. Topological method may bring better results and higher consistency compared to using geometric method alone. Our method may be useful for future studies identifying and characterizing kinked helices, and revealing the function-structure relationship in kinked helices.

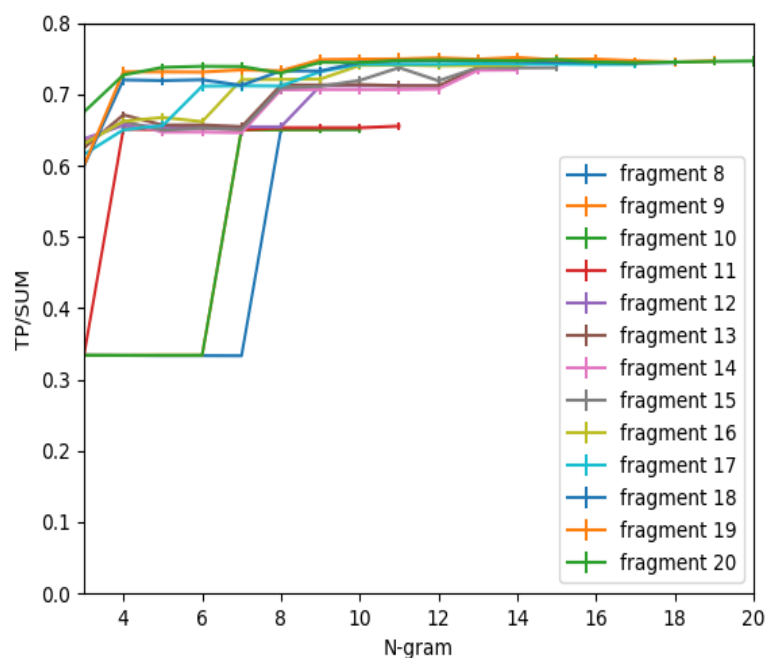


Figure 4-14 Agreement between topological and geometrical methods at different fragment length and N-grams. True Positive (TP) is the number of α -helices classified as kinked α -helices by both the ValgusHel-geometry and ValgusHel-topology. Sum is the number of total α -helices included. TP/SUM reflects the agreement between topological and geometry methods.

4.5 CONCLUSIONS

We develop a new method “ValgusHel” to identify and annotate kinked and curved α -helices by using both geometric and topological characteristics. Topological method may bring better results and higher consistency compared to using geometric method alone. 9-residue sequence pairs with higher sequence identity (>0.5) had higher helix angle similarity compared to those with lower sequence identity (<0.5), although the difference was small. Sequence similarity is lowest within kink, then curved, and highest in normal α -helices. “ValgusHel” may be useful in future studies to examine structure-function relationships in kinked α -helices.

CHAPTER 5 : CONCLUSIONS

This dissertation focuses on analysis of structural and topological variations in amino acids encoded by synonymous codons, and their application in predicting protein fitness/activity and α -helix secondary structure classification. In Chapter 2, we successfully created the 10220culled dataset containing 10,220 individual protein chains with codon, amino acid sequence, secondary structure, and α -carbon coordinates. Since there are no other similar datasets publicly available and this 10220culled dataset may be used by researchers for related protein structure-function research. Although Delaunay tessellation is adequately applied in a previous study (2006) [28] to examine the knowledge-based potential results, we re-applied the method to a much larger dataset and extended the analysis to codon level rather than amino acid level alone. Moreover, we explored how factors (e.g., simplex sequence bias, simplex edge length, low number of simplices in certain simplex compositions) impacted the potential estimation and improved the estimation through restricting these factors. While synonymous codons have been associated with some human diseases, the learning curve results may suggest that potential estimation at codon level to be less accurate due to inadequate sample size. These findings may be useful for future researcher in using Delaunay tessellation and considering performing analysis at codon level with more protein structure available in PDB in the future. We did not observe significant topological property differences in proteins caused by cancer-causing vs. non-cancer-causing silent mutations. In Chapter 3, tested whether the methodology of computational mutagenesis methods designed for Ras

protein was transferable to β -lactamase and other proteins, taking advantage of more proteins have saturation mutagenesis function data. While developing a machine learning model for each protein may result in better model performance, it is challenging for researchers studying proteins without created models. Thus, we examined whether model built for one protein can be used for another protein in the same activity category (e.g., antibiotic resistance). While this generalizability result is not good, the results can still serve as references for researchers to explore more variables to be included to improve the model in future works. The model performance was very good when predicting pairwise amino acid substitutions. In Chapter 4, the new *ValgusHel* method showed acceptable reliability and detectability to identify kinked, curved, and normal α -helices. Sequence similarity is lowest within kinked, then curved, and highest in normal α -helices. The *ValgusHel*-geometry and *ValgusHel*-topology had good agreement, and n-gram using protein topological properties may be more reliable than the geometry method alone. While the agreement in such classification is generally low in current available methods, this *ValgusHel* method may serve as an additional method to identify and annotate protein α -helices and contribute to the protein structure-function research field.

APPENDIX

Appendix Table 1 Codon labels used in Database based on Standard Genetic Codon Chart

1 st base	2 nd base												3 rd base
	T			C			A			G			
T	TTT	F1	Phe/F	TCT	S1	Ser/S	TAT	Y1	Tyr/Y	TGT	C1	Cys/C	T
	TTC	F2		TCC	S2		TAC	Y2		TGC	C2		C
	TTA	L1		TCA	S3		TAA	-	Stop	TGA	-	Stop	A
	TTG	L2		TCG	S4		TAG	-		TGG	W1	Trp/W	G
C	CTT	L3	Leu/L	CCT	P1	Pro/P	CAT	H1	His/H	CGT	R1	Arg/R	T
	CTC	L4		CCC	P2		CAC	H2		CGC	R2		C
	CTA	L5		CCA	P3		CAA	Q1	Gln/Q	CGA	R3		A
	CTG	L6		CCG	P4		CAG	Q2		CGG	R4		G
A	ATT	I1	Ile/I	ACT	T1	Thr/T	AAT	N1	Asn/N	AGT	S5	Ser/S	T
	ATC	I2		ACC	T2		AAC	N2		AGC	S6		C
	ATA	I3		ACA	T3		AAA	K1	Lys/K	AGA	R5	Arg/R	A
	ATG	M1	Met/M	ACG	T4		AAG	K2		AGG	R6		G
G	GTT	V1	Val/V	GCT	A1	Ala/A	GAT	D1	Asp/D	GGT	G1	Gly/G	T
	GTC	V2		GCC	A2		GAC	D2		GGC	G2		C
	GTA	V3		GCA	A3		GAA	E1	Glu/E	GGA	G3		A
	GTG	V4		GCG	A4		GAG	E2		GGG	G4		G

REFERENCES

1. Sauna, Z.E. and C. Kimchi-Sarfaty, *Understanding the contribution of synonymous mutations to human disease*. Nat Rev Genet, 2011. **12**(10): p. 683-91.
2. Osawa, S., et al., *Recent evidence for evolution of the genetic code*. Microbiol Rev, 1992. **56**(1): p. 229-64.
3. Kimura, M., *Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution*. Nature, 1977. **267**(5608): p. 275-6.
4. Chamary, J.-V., J.L. Parmley, and L.D. Hurst, *Hearing silence: non-neutral evolution at synonymous sites in mammals*. Nature Reviews Genetics, 2006. **7**(2): p. 98-108.
5. Denecke, J., et al., *An activated 5' cryptic splice site in the human ALG3 gene generates a premature termination codon insensitive to nonsense-mediated mRNA decay in a new case of congenital disorder of glycosylation type Id (CDG-Id)*. Hum Mutat, 2004. **23**(5): p. 477-86.
6. Cartegni, L., S.L. Chew, and A.R. Krainer, *Listening to silence and understanding nonsense: exonic mutations that affect splicing*. Nat Rev Genet, 2002. **3**(4): p. 285-98.
7. Aretz, S., et al., *Familial adenomatous polyposis: aberrant splicing due to missense or silent mutations in the APC gene*. Hum Mutat, 2004. **24**(5): p. 370-80.
8. Chen, R., et al., *Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association*. PLoS One, 2010. **5**(10): p. e13574.
9. Kimchi-Sarfaty, C., et al., *A "silent" polymorphism in the MDR1 gene changes substrate specificity*. Science, 2007. **315**(5811): p. 525-528.
10. Heal, J.R., et al., *Specific interactions between sense and complementary peptides: the basis for the proteomic code*. Chembiochem, 2002. **3**(2-3): p. 136-51.
11. Schwanhauss, B., et al., *Global quantification of mammalian gene expression control*. Nature, 2011. **473**(7347): p. 337-42.
12. Plotkin, J.B. and G. Kudla, *Synonymous but not the same: the causes and consequences of codon bias*. Nature Reviews Genetics, 2011. **12**(1): p. 32-42.
13. Bartoszewski, R.A., et al., *A synonymous single nucleotide polymorphism in DeltaF508 CFTR alters the secondary structure of the mRNA and the expression of the mutant protein*. J Biol Chem, 2010. **285**(37): p. 28741-8.
14. Singh, R.K., A. Tropsha, and Vaisman, II, *Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues*. J Comput Biol, 1996. **3**(2): p. 213-21.
15. Tropsha, A., et al., *Statistical geometry analysis of proteins: implications for inverted structure prediction*. Pac Symp Biocomput, 1996: p. 614-23.

16. Krishnamoorthy, B. and A. Tropsha, *Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations*. Bioinformatics, 2003. **19**(12): p. 1540-8.
17. Ilyin, V.A., A. Abyzov, and C.M. Leslin, *Structural alignment of proteins by a novel TOPOFIT method, as a superimposition of common volumes at a topomax point*. Protein Sci, 2004. **13**(7): p. 1865-74.
18. Roach, J., et al., *Structure alignment via Delaunay tetrahedralization*. Proteins, 2005. **60**(1): p. 66-81.
19. Bostick, D. and Vaisman, II, *A new topological method to measure protein structure similarity*. Biochem Biophys Res Commun, 2003. **304**(2): p. 320-5.
20. Liang, J., et al., *Analytical shape computation of macromolecules: II. Inaccessible cavities in proteins*. Proteins, 1998. **33**(1): p. 18-29.
21. Masso, M. and Vaisman, II, *Comprehensive mutagenesis of HIV-1 protease: a computational geometry approach*. Biochem Biophys Res Commun, 2003. **305**(2): p. 322-6.
22. Carter, C.W., Jr., et al., *Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations*. J Mol Biol, 2001. **311**(4): p. 625-38.
23. Tropsha, A., et al., *Simplicial neighborhood analysis of protein packing (SNAPP): a computational geometry approach to studying proteins*. Methods Enzymol, 2003. **374**: p. 509-44.
24. Wako, H. and T. Yamato, *Novel method to detect a motif of local structures in different protein conformations*. Protein Eng, 1998. **11**(11): p. 981-90.
25. Huan, J., et al., *Comparing graph representations of protein structure for mining family-specific residue-based packing motifs*. J Comput Biol, 2005. **12**(6): p. 657-71.
26. Huan, J., et al., *Accurate classification of protein structural families using coherent subgraph analysis*. Pac Symp Biocomput, 2004: p. 411-22.
27. Taylor, T., et al., *New method for protein secondary structure assignment based on a simple topological descriptor*. Proteins, 2005. **60**(3): p. 513-24.
28. Taylor, T.J., *Analysis of the structure and topology of real and model proteins using Delaunay tessellation*. 2006: George Mason University.
29. Poupon, A., *Voronoi and Voronoi-related tessellations in studies of protein structure and interaction*. Curr Opin Struct Biol, 2004. **14**(2): p. 233-41.
30. Vaisman, I.I., *Statistical and computational geometry of biomolecular structure*, in *Handbook of Computational Statistics*. 2012, Springer. p. 1095-1112.
31. Schlick, T. and H.H. Gan, *Computational Methods for Macromolecules: Challenges and Applications: Proceedings of the 3rd International Workshop on Algorithms for Macromolecular Modeling, New York, October 12–14, 2000*. Vol. 24. 2012: Springer Science & Business Media.
32. Masso, M., et al., *Fitness of unregulated human Ras mutants modeled by implementing computational mutagenesis and machine learning techniques*. Heliyon, 2019. **5**(6): p. e01884.

33. Masso, M. and I.I. Vaisman, *Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis*. Bioinformatics, 2008. **24**(18): p. 2002-2009.
34. Burks, E.A., et al., *In vitro scanning saturation mutagenesis of an antibody binding pocket*. Proc Natl Acad Sci U S A, 1997. **94**(2): p. 412-7.
35. Georgescu, R., G. Bandara, and L. Sun, *Saturation mutagenesis*. Methods Mol Biol, 2003. **231**: p. 75-83.
36. Fowler, D.M. and S. Fields, *Deep mutational scanning: a new style of protein science*. Nat Methods, 2014. **11**(8): p. 801-7.
37. Bhasin, M. and R. Varadarajan, *Prediction of Function Determining and Buried Residues Through Analysis of Saturation Mutagenesis Datasets*. Front Mol Biosci, 2021. **8**: p. 635425.
38. Camacho, C., et al., *BLAST+: architecture and applications*. BMC bioinformatics, 2009. **10**(1): p. 421.
39. Cock, P.J., et al., *Biopython: freely available Python tools for computational molecular biology and bioinformatics*. Bioinformatics, 2009. **25**(11): p. 1422-1423.
40. Qhull. 1995; Available from: www.qhull.org.
41. Pyhull 1.5.4 documentation. [cited 2021 September 5]; Available from: <https://pythonhosted.org/pyhull/>.
42. Sippl, M.J., *Knowledge-based potentials for proteins*. Current opinion in structural biology, 1995. **5**(2): p. 229-235.
43. Taylor, T., et al., *New method for protein secondary structure assignment based on a simple topological descriptor*. 2005. **60**(3): p. 513-524.
44. *Synonymous Mutations In Cancer database*. Available from: <http://synmicdb.dkfz.de/rsynmicdb/>.
45. Singh, R.K., A. Tropsha, and I.I. Vaisman, *Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues*. Journal of Computational Biology, 1996. **3**(2): p. 213-221.
46. Masso, M., et al., *Fitness of unregulated human Ras mutants modeled by implementing computational mutagenesis and machine learning techniques*. Heliyon, 2019. **5**(6): p. e01884.
47. Masso, M. and I.I. Vaisman, *AUTO-MUTE: web-based tools for predicting stability changes in proteins due to single amino acid replacements*. Protein Engineering, Design & Selection, 2010. **23**(8): p. 683-687.
48. *LACTB. E.COLI*. [cited 2021 May 10]; Available from: https://www.prospecbio.com/beta_lactamase.
49. Mehlhoff, J.D., et al., *Collateral fitness effects of mutations*. Proceedings of the National Academy of Sciences, 2020. **117**(21): p. 11597-11607.
50. Dayhoff, M.O., *A model of evolutionary change in proteins*. Atlas of protein sequence and structure, 1972. **5**: p. 89-99.
51. Frank, E., et al., *Data mining in bioinformatics using Weka*. Bioinformatics, 2004. **20**(15): p. 2479-2481.

52. Smith, T.C. and E. Frank, *Introducing machine learning concepts with WEKA*, in *Statistical genomics*. 2016, Springer. p. 353-378.
53. Starita, L.M., et al., *Massively Parallel Functional Analysis of BRCA1 RING Domain Variants*. *Genetics*, 2015. **200**(2): p. 413-22.
54. Adkar, B.V., et al., *Protein model discrimination using mutational sensitivity derived from deep sequencing*. *Structure*, 2012. **20**(2): p. 371-381.
55. Kitzman, J.O., et al., *Massively parallel single-amino-acid mutagenesis*. *Nat Methods*, 2015. **12**(3): p. 203-6, 4 p following 206.
56. Olson, C.A., N.C. Wu, and R. Sun, *A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain*. *Curr Biol*, 2014. **24**(22): p. 2643-51.
57. Mishra, P., et al., *Systematic Mutant Analyses Elucidate General and Client-Specific Aspects of Hsp90 Function*. *Cell Rep*, 2016. **15**(3): p. 588-598.
58. Melnikov, A., et al., *Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes*. *Nucleic Acids Res*, 2014. **42**(14): p. e112.
59. Suiter, C.C., et al., *Massively parallel variant characterization identifies NUDT15 alleles associated with thiopurine toxicity*. *Proc Natl Acad Sci U S A*, 2020. **117**(10): p. 5394-5401.
60. McLaughlin, R.N., Jr., et al., *The spatial architecture of protein function and adaptation*. *Nature*, 2012. **491**(7422): p. 138-42.
61. Bandaru, P., et al., *Deconstruction of the Ras switching cycle through saturation mutagenesis*. *Elife*, 2017. **6**: p. e27810.
62. Stiffler, M.A., D.R. Hekstra, and R. Ranganathan, *Evolvability as a function of purifying selection in TEM-1 β -lactamase*. *Cell*, 2015. **160**(5): p. 882-892.
63. Roscoe, B.P., et al., *Analyses of the effects of all ubiquitin point mutants on yeast growth rate*. *Journal of molecular biology*, 2013. **425**(8): p. 1363-1377.
64. Kendrew, J.C., et al., *A three-dimensional model of the myoglobin molecule obtained by x-ray analysis*. *Nature*, 1958. **181**(4610): p. 662-666.
65. Pauling, L., R.B. Corey, and H.R. Branson, *The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain*. *Proceedings of the National Academy of Sciences*, 1951. **37**(4): p. 205-211.
66. Kumar, P. and M. Bansal, *HELANAL-Plus: a web server for analysis of helix geometry in protein structures*. *Journal of Biomolecular Structure and Dynamics*, 2012. **30**(6): p. 773-783.
67. Blundell, T., et al., *Solvent-induced distortions and the curvature of α -helices*. *Nature*, 1983. **306**(5940): p. 281-283.
68. Barlow, D. and J. Thornton, *Helix geometry in proteins*. *Journal of molecular biology*, 1988. **201**(3): p. 601-619.
69. Langelaan, D., *Structural studies of apelin and its receptor as well as the characteristics and causes of membrane protein helix kinks*. 2013.
70. Hall, S.E., K. Roberts, and N. Vaidehi, *Position of helical kinks in membrane protein crystal structures and the accuracy of computational prediction*. *Journal of Molecular Graphics and Modelling*, 2009. **27**(8): p. 944-950.

71. Chakrabarti, P. and S. Chakrabarti, *C—H···O hydrogen bond involving proline residues in α -helices*. Journal of molecular biology, 1998. **284**(4): p. 867-873.
72. Harris, T., A.R. Graber, and M. Covarrubias, *Allosteric modulation of a neuronal K⁺ channel by 1-alkanols is linked to a key residue in the activation gate*. American Journal of Physiology-Cell Physiology, 2003. **285**(4): p. C788-C796.
73. Reddy, T., et al., *Structural and functional characterization of transmembrane segment IX of the NHE1 isoform of the Na⁺/H⁺ exchanger*. Journal of Biological Chemistry, 2008. **283**(32): p. 22018-22030.
74. Singh, R., et al., *Activation of the cannabinoid CB1 receptor may involve a W6.48/F3.36 rotamer toggle switch*. The Journal of peptide research, 2002. **60**(6): p. 357-370.
75. Katritch, V., V. Cherezov, and R.C. Stevens, *Structure-function of the G protein-coupled receptor superfamily*. Annual review of pharmacology and toxicology, 2013. **53**.
76. Law, E.C., et al., *Examining the conservation of kinks in alpha helices*. PloS one, 2016. **11**(6): p. e0157553.
77. Woolfson, D.N., R.J. Mortishire-Smith, and D.H. Williams, *Conserved positioning of proline residues in membrane-spanning helices of ion-channel proteins*. Biochemical and biophysical research communications, 1991. **175**(3): p. 733-737.
78. Suh, J.-Y., et al., *Unusually stable helical kink in the antimicrobial peptide—A derivative of gaegurin*. FEBS letters, 1996. **392**(3): p. 309-312.
79. Langelaan, D.N., et al., *Improved helix and kink characterization in membrane proteins allows evaluation of kink sequence predictors*. Journal of chemical information and modeling, 2010. **50**(12): p. 2213-2220.
80. Bansal, M., S. Kumart, and R. Velavan, *HELANAL: a program to characterize helix geometry in proteins*. Journal of Biomolecular Structure and Dynamics, 2000. **17**(5): p. 811-819.
81. Visiers, I., B.B. Braunheim, and H. Weinstein, *Prokink: a protocol for numerical evaluation of helix distortions by proline*. Protein engineering, 2000. **13**(9): p. 603-606.
82. Wilman, H.R., J. Shi, and C.M. Deane, *Helix kinks are equally prevalent in soluble and membrane proteins*. Proteins: Structure, Function, and Bioinformatics, 2014. **82**(9): p. 1960-1970.
83. *AH^AH Survey Results*. 2021 [cited 2021 August 24]; Available from: http://opig.stats.ox.ac.uk/webapps/ahah/php/experiment_results.php.
84. Yohannan, S., et al., *The evolution of transmembrane helix kinks and the structural diversity of G protein-coupled receptors*. Proceedings of the National Academy of Sciences, 2004. **101**(4): p. 959-963.
85. Cordes, F.S., J.N. Bright, and M.S. Sansom, *Proline-induced distortions of transmembrane helices*. Journal of molecular biology, 2002. **323**(5): p. 951-960.
86. Riek, R.P., et al., *Non- α -helical elements modulate polytopic membrane protein architecture*. Journal of molecular biology, 2001. **306**(2): p. 349-362.

87. Meruelo, A.D., I. Samish, and J.U. Bowie, *TMKink: a method to predict transmembrane helix kinks*. Protein Science, 2011. **20**(7): p. 1256-1264.
88. *The Rainey Lab - MC-HELAN*. [cited 2021 April 10]; Available from: <http://structbio.biochem.dal.ca/jrainey/MC-HELAN/documentation.html>.
89. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein blocks*. Proceedings of the National Academy of Sciences, 1992. **89**(22): p. 10915-10919.
90. Waterhouse, A.M., et al., *Jalview Version 2—a multiple sequence alignment editor and analysis workbench*. Bioinformatics, 2009. **25**(9): p. 1189-1191.
91. Lupas, A., M. Van Dyke, and J. Stock, *Predicting coiled coils from protein sequences*. Science, 1991: p. 1162-1164.
92. Drozdetskiy, A., et al., *JPred4: a protein secondary structure prediction server*. Nucleic Acids Res, 2015. **43**(W1): p. W389-94.
93. Waterhouse, A.M., et al., *Jalview Version 2—a multiple sequence alignment editor and analysis workbench*. Bioinformatics, 2009. **25**(9): p. 1189-91.
94. Bailey, T.L. and C. Elkan, *Fitting a mixture model by expectation maximization to discover motifs in biopolymers*. Proc Int Conf Intell Syst Mol Biol, 1994. **2**: p. 28-36.
95. Sigrist, C.J., et al., *New and continuing developments at PROSITE*. Nucleic Acids Res, 2013. **41**(Database issue): p. D344-7.
96. de Castro, E., et al., *ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins*. Nucleic Acids Res, 2006. **34**(Web Server issue): p. W362-5.
97. *Alpha-Helix Geometry Part. 2*. [cited 2021 April 16]; Available from: http://www.cryst.bbk.ac.uk/PPS95/course/3_geometry/helix2.html.

BIOGRAPHY

Shengyuan Wang received his Bachelor of Sciences in Biotechnology from China Agricultural University in 2013. He received his Master of Sciences in Pharmaceutical Sciences from North Carolina Central University in 2016.