

ALGEBRAIC TOPOLOGY AND GRAPH THEORY BASED
APPROACHES FOR PROTEIN FLEXIBILITY ANALYSIS AND B
FACTOR PREDICTION

By

David Bramer

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Mathematics - Doctor of Philosophy

2019

ABSTRACT

ALGEBRAIC TOPOLOGY AND GRAPH THEORY BASED APPROACHES FOR PROTEIN FLEXIBILITY ANALYSIS AND B FACTOR PREDICTION

By

David Bramer

Protein fluctuation, measured by B factors, has been shown to highly correlate to protein flexibility and function. Several methods have been developed to predict protein B factor as well as related applications. While many B factor methods exist, reliable B factor prediction continues to be an ongoing challenge and there is much room for improvement.

This work introduces a paradigm shifting geometric graph based model called the multi-scale weighted colored graph (MWCG) model. The MWCG model is a new computational algorithm that greatly improves the current landscape of protein structural fluctuation analysis. The MWCG model treats each protein as a colored graph where colored nodes correspond to atomic element types and edges are weighted by a generalized centrality metric. Each graph contains multiple subgraphs based on interaction types between graphic nodes. Protein rigidity is represented by generalized centralities of subgraphs. MWCGs predict B factors of protein residues and accurately analyze the flexibility of all atoms in a protein simultaneously. The MWCG model presented here captures element specific interactions across multiple scales and is a novel visual tool for identifying various protein secondary structures. This work also demonstrates MWCG protein hinge detection using a variety of proteins.

Cross-protein prediction of B factors has previously been an unsolved problem in terms of B factor prediction. Many proteins are difficult to crystallize, and for some it is likely impossible, so models that can cross predict protein B factor are absolutely necessary. Using

machine learning and the MWCG method, this work provides a robust cross protein B factor prediction using a set of known proteins to predict the B factors of a protein previously unseen to the algorithm. The algorithm connects different proteins using global protein features such as the resolution of the X-ray crystallography data. The combination of global and local features results in successful cross protein B factor prediction. To test and validate these results this work considers several machine learning approaches such as random forest, gradient boosted trees, and deep convolutional neural networks.

Recently, persistent homology has had tremendous success in biomolecular data analysis. It works by examining the topological relationship or connectivity of a group of atoms in a molecule at a variety of scales, then rendering a family of topological representations of the molecule. Persistent homology is rarely employed for analysis of atomic properties, such as protein flexibility analysis or B factor prediction. This work introduces atom specific persistent homology (ASPH) to provide a local atomic level representation of a molecule via a global topological tool. This is achieved through the construction of a pair of conjugated sets of atoms and corresponding conjugated simplicial complexes, as well as conjugated topological spaces. The difference between the topological invariants of the pair of conjugated sets is measured by Bottleneck and Wasserstein metrics and leads to an atom specific topological representation of individual atomic properties in a molecule. Atom specific topological features are integrated with various machine learning algorithms, including gradient boosting trees and convolutional neural network for protein thermal fluctuation analysis and blind cross protein B factor prediction.

Extensive numerical testing indicates the proposed methods provide novel and powerful graph theory and algebraic topology based tools for analyzing and predicting atom specific, localized protein flexibility information.

TABLE OF CONTENTS

LIST OF TABLES	vii
LIST OF FIGURES	xii
KEY TO ABBREVIATIONS	xv
Chapter 1 Overview	1
Chapter 2 Background	7
2.1 Computing Protein Flexibility and Dynamics	8
2.2 Data	11
Chapter 3 Multiscale Weighted Colored Graphs	13
3.1 Weighted colored graphs	13
3.2 WCG Centrality	15
3.3 Weighted Colored Graph Flexibility Analysis	16
3.4 Multiscale Weighted Colored Graph Flexibility Analysis	17
3.5 Parameterization	19
Chapter 4 Atom Specific Persistent Homology	22
4.1 Overview	22
4.2 Simplex & Simplicial Complex	24
4.3 Homology	25
4.4 Filtration & Persistence	27
4.5 Similarity and distance	27
4.6 Vietoris-Rips Complex	29
4.7 Atom Specific Persistent Homology & Element Specific Persistent Homology	29
Chapter 5 Machine Learning	34
5.1 Machine Learning Algorithms	35
5.1.1 Ensemble Methods	35
5.1.1.1 Random forest	36
5.1.1.2 Gradient boosted trees	37
5.1.2 Neural Networks	37
5.1.2.1 Convolutional Neural Network	38
5.1.3 Consensus methods	39
5.2 General Machine Learning Features	39
5.2.1 Global features	40
5.2.2 Local features	41
5.3 MWCG Features	42
5.3.1 Image-like MWCG Features	43

5.4	ASPH & ESPH Features	44
5.4.1	Image-like ASPH & ESPH Features	45
5.4.2	Cutoff Distance	47
5.5	Machine Learning Model Parameters	48
5.5.1	MWCG	48
5.5.1.1	Random Forest	48
5.5.1.2	Gradient Boosted Trees	49
5.5.1.3	Deep Convolutional Neural Network	49
5.5.2	ASPH & ESPH	51
5.5.2.1	Gradient Boosted Trees	52
5.5.2.2	Deep Convolutional Neural Network	52
5.6	Machine Learning Datasets	54
5.6.1	Training set and test set	54
Chapter 6	Workflow	56
Chapter 7	Results	59
7.1	Visualization of Element Specific Correlation Maps	59
7.2	Hinge Detection	59
7.3	MWCG	66
7.3.1	Validation	66
7.3.2	Fitting Results	67
7.4	Machine Learning Results	72
7.4.1	MWCG	72
7.4.1.1	Efficiency comparison	73
7.4.1.2	Machine learning performance	74
7.4.1.3	Relative feature importance	80
7.5	ASPH & ESPH B Factor Prediction	81
7.5.1	Least Squares Fitting	81
7.5.2	Machine Learning	82
Chapter 8	Discussion	115
8.1	Element Specific Heat Maps	115
8.2	Hinge Detection	116
8.3	Fitting Models	117
8.3.1	MWCG	117
8.3.2	ASPH & ESPH	118
8.4	Machine Learning Models	118
8.4.1	MWCG	118
8.4.2	ASPH & ESPH	119
Chapter 9	Conclusions and Future Directions	120
9.1	Conclusions	120
9.2	Future Directions	127
9.2.1	Software Development	127

9.2.2	Inclusion of other datasets	128
9.2.3	Specific applications in drug design and docking pose	128
9.2.4	Other general approaches	129
BIBLIOGRAPHY		130

LIST OF TABLES

Table 2.1:	Notable molecular mechanic techniques and the year of introduction. . . .	11
Table 3.1:	Element pair combinations used in weighted colored graph.	14
Table 3.2:	Parameters used for correlation kernels in a parameter-free MWCG. Parameter optimization results originally published in Bramer <i>et al</i> [1]. . . .	20
Table 4.1:	Topological invariants displayed as Betti numbers. Betti-0 represents the number of connected components, Betti-1 the number of tunnels or circles, and Betti-2 the number of cavities or voids. Two auxiliary rings are added to the torus to illustrate that Betti-1=2.	23
Table 5.1:	The packing density distance parameters ($d \text{ \AA}$) used for generating short medium, and long packing density machine learning features.	42
Table 5.2:	Correlation kernel parameters used to generate parameter-free MWCG machine learning features. Parameters based on previous results.[1]	43
Table 5.3:	Parameters used for topological feature generation. All features used a cut-off of 11Å. Both lorentz (Lor) and exponential (exp) kernels and Bottleneck (B) and Wasserstein (W) distance metrics were used.	46
Table 5.4:	Parameters used for the element specific persistent homology features with a cutoff of 11 Å.	48
Table 5.5:	Boosted gradient tree parameters used for testing MWCG based B factor prediction. These parameters were determined using a grid search. Any hyper parameters not listed below were taken to be the default values provided by the python scikit-learn package. MWCG based GBT machine learning prediction results originally published in Bramer <i>et al</i> [2].	49
Table 5.6:	MWCG based deep Convolutional Neural Network (CNN) hyper-parameters used for testing. These hyper-parameters were determined using a grid search. Any hyper parameters not listed below were taken to be the default values provided by python with the Keras package. MWCG machine learning prediction results originally published in Bramer <i>et al</i> [2].	51
Table 5.7:	Boosted gradient tree parameters used for persistent homology based prediction testing. Parameters were determined using a grid search. Any hyper parameters not listed below were taken to be the default values provided by the python scikit-learn package.	52

Table 5.8: Convolutional Neural Network (CNN) parameters used for testing persistent homology based features. Parameters were determined using a grid search. Any hyper-parameters not listed below were taken to be the default values provided by python with the Keras package.	54
Table 7.4: Correlation coefficients for B factor prediction obtained by MWCG, optimal FRI (opFRI), parameter free FRI (pfFRI), and Gaussian normal mode (GNM) for a set of 364 proteins. GNM scores reported here are the result of tests with a processed set of PDB files as described in Chapter 2.2. MWCG results originally published in Bramer <i>et al</i> [1].	67
Table 7.12: Pearson correlation coefficients for cross protein heavy atom blind B factor prediction obtained by random forest (RF), boosted gradient (GBT), and convolutional neural network (CNN) for the Superset. Results reported use heavy atoms in both training and prediction. MWCG machine learning results originally published in Bramer <i>et al</i> [2].	75
Table 7.21: Pearson correlation coefficients of persistent homology based least squares fitting C_α B factor prediction of all proteins using 11Å cutoff. Two Bottleneck (B) and Wasserstein (W) metrics using various kernel choices are included.	83
Table 7.22: Persistent homology based Pearson correlation coefficients for cross protein C_α atom blind B factor prediction obtained by boosted gradient (GBT), convolutional neural network (CNN), and consensus method (CON) for the Superset.	92
Table 7.1: Correlation coefficients for B factor prediction obtained by optimal FRI (opFRI), parameter free FRI (pfFRI), and Gaussian normal mode (GNM) for small-size structures. Results for opFRI, pfFRI are taken from <i>Opron et al</i> [3]. GNM and NMA values are taken from the coarse-grained (C_α) results reported in <i>Park et al</i> [4]. MWCG results are parameter free and use all C, N, and O to predict C_α . MWCG results originally published in Bramer <i>et al</i> [1].	98
Table 7.2: Correlation coefficients for B factor prediction obtained by optimal FRI (opFRI), parameter free FRI (pfFRI) and Gaussian normal mode (GNM) for medium-size structures. Results for opFRI, pfFRI are taken from <i>Opron et al</i> [3]. GNM and NMA values are taken from the coarse-grained (C_α) results reported in <i>Park et al</i> [4]. MWCG results are parameter free and use all C, N, and O to predict C_α . MWCG results originally published in Bramer <i>et al</i> [1].	99

Table 7.3: Correlation coefficients for B factor prediction obtained by optimal FRI (opFRI), parameter free FRI (pfFRI), and Gaussian normal mode (GNM) for large-size structures. Results for opFRI, pfFRI are taken from Opron <i>et al</i> [3]. GNM and NMA values are taken from the coarse-grained (C_α) results reported in Park <i>et al</i> [4]. MWCG results are parameter free and use all C, N, and O to predict C_α . MWCG results originally published in Bramer <i>et al</i> [1].	100
Table 7.5: Average pearson correlation coefficients for C_α B factor prediction with FRI, GNM and NMA for three structure sets from Park et al. [4] and a superset of 364 structures. Results for opFRI, pfFRI are taken from Opron <i>et al</i> [3]. GNM and NMA values are taken from the coarse-grained (C_α) results reported in Park <i>et al</i> . [4] MWCG results are parameter free and use all C, N, and O to predict C_α . MWCG Results originally published in Bramer <i>et al</i> [1].	101
Table 7.6: Pearson Correlation coefficients for C_α , non C_α carbon, nitrogen, oxygen, and sulfur using parameter free MWCG. Only 215 of the 364 proteins contain sulfur atoms. MWCG results originally published in Bramer <i>et al</i> [1].	101
Table 7.7: CPU execution times, in seconds, from efficiency comparison between GNM [3], RF, GBT, and CNN. Results originally reported in Bramer <i>et al</i> [2] .	102
Table 7.8: Average Pearson correlation coefficients (PCC) both of all heavy atom and C_α only B factor predictions for small-, medium-, and large-sized protein sets along with the entire superset of the 364 protein dataset. Predictions of random forest (RF), gradient boosted tree (GBT), and convolutional neural network (CNN) are obtained by leave-one-protein-out (blind), while predictions of parameter-free flexibility-rigidity index (pfFRI), Gaussian network model (GNM) and normal mode analysis (NMA) were obtained via the least squares fitting of individual proteins. All machine learning models use all heavy atom information for training. MWCG machine learning B factor prediction results originally reported in Bramer <i>et al</i> [2].	103
Table 7.9: Pearson correlation coefficients for cross protein heavy atom blind MWCG B factor prediction obtained by random forest (RF), boosted gradient (GBT), and convolutional neural network (CNN) for the small-sized protein set. Results reported use heavy atoms in both training and prediction. Originally published in Bramer <i>et al</i> [2].	104

Table 7.10: Pearson correlation coefficients for cross protein heavy atom blind MWCG B factor prediction obtained by random forest (RF), boosted gradient (GBT), and convolutional neural network (CNN) for the medium-sized protein set. Results reported use heavy atoms in both training and prediction. Originally published in Bramer <i>et al</i> [2].	105
Table 7.11: Pearson correlation coefficients for cross protein heavy atom blind MWCG B factor prediction obtained by random forest (RF), boosted gradient (GBT), and convolutional neural network (CNN) for the large-sized protein set. Results reported use heavy atoms in both training and prediction. Originally published in Bramer <i>et al</i> [2].	106
Table 7.13: ASPH and ESPH average Pearson correlation coefficients C_α B factor predictions for small-, medium-, and large-sized protein sets along with the entire superset of the 364 protein dataset. Gradient boosted tree (GBT), convolutional neural network, and consensus(CON) results are obtained by leave-one-protein-out (blind). Predictions of parameter-free flexibility-rigidity index (pfFRI), Gaussian network model (GNM) and normal mode analysis (NMA) were obtained via the least squares fitting of individual proteins.	107
Table 7.14: ASPH and ESPH Pearson correlation coefficients for cross protein C_α atom blind B factor prediction obtained by boosted gradient (GBT), convolutional neural network (CNN), and consensus (CON) for the small-sized protein set.	108
Table 7.15: ASPH and ESPH Pearson correlation coefficients for cross protein C_α atom blind B factor prediction obtained by boosted gradient (GBT), convolutional neural network (CNN), and consensus (CON) for the medium-sized protein set.	109
Table 7.16: ASPH and ESPH Pearson correlation coefficients for cross protein C_α atom blind B factor prediction obtained boosted gradient (GBT), convolutional neural network (CNN), and consensus (CON) for the large-sized protein set.	110
Table 7.17: ASPH and ESPH Pearson correlation coefficients of least squares fitting C_α B factor prediction of small proteins using 11Å cutoff. Two Bottleneck (B) and Wasserstein (W) metrics using various kernel choices are included.	111
Table 7.18: ASPH and ESPH Pearson correlation coefficients of least squares fitting C_α B factor prediction of medium proteins using 11Å cutoff. Two Bottleneck (B) and Wasserstein (W) metrics using various kernel choices are included.	112

Table 7.19: ASPH and ESPH Pearson correlation coefficients of least squares fitting C_α B factor prediction of large proteins using 11Å cutoff. Two Bottleneck (B) and Wasserstein (W) metrics using various kernel choices are included. 113

Table 7.20: ASPH and ESPH average Pearson correlation coefficients of least squares fitting C_α B factor prediction of small, medium, large, and superset using 11Å cutoff. Two Bottleneck (B) and Wasserstein (W) metrics using various kernel choices are included. Results for pFRI are taken from Opron et al[3]. GNM and NMA value are taken from the course grained C_α results reported in Park et al[4]. 114

LIST OF FIGURES

Figure 3.1:	The average Pearson correlation coefficient (PCC) as found by optimizing individual kernels in the range of $\eta^n = 1, \dots, 40$. Parameter optimization results originally published in Bramer <i>et al</i> [1].	20
Figure 4.1:	From left to right an example of a 0-simplex, 1-simplex, 2-simplex, and 3-simplex.	24
Figure 4.2:	(a) An example of 5 points in \mathbb{R}^2 and (b) the corresponding topological barcode. The length of each barcode corresponds to the persistence of each topological object ($\beta_0, \beta_1, \beta_2, \text{etc.}$) over the filtration.	28
Figure 4.3:	Illustration of Atom-specific persistent homology point clouds. Top: the original point cloud. The atom of interest is at the center of the circle. Second row: a pair of conjugated sets of point clouds for atom-specific persistent homology. The rest: Four pairs of conjugated point clouds for atom-specific and element-specific persistent homology.	30
Figure 4.4:	Illustration of residue 338 C_α atom-specific persistent homology in the CC element-specific point cloud of protein PDBID 1AIE. For this example residues 332-339 are used and are shown on the left. The C_α location used to generate the barcodes (right) is highlighted in red in the left chart. Conjugated persistence barcodes are generated with and without the selected C_α	33
Figure 5.1:	An example of a perceptron, the basic functional unit of a neural network.	38
Figure 5.2:	An illustration of a fully connected deep neural network. Circles represent neurons and connections between neurons are indicated by arrows. Each connection has an associated weight. A neural network is considered “deep” when it uses several hidden layers.	38
Figure 5.3:	Frequency of the number of heavy elements from the 364 protein dataset. Figure originally published in Bramer <i>et al</i> [2].	41
Figure 5.4:	Illustration of modified persistence diagrams used in distance calculations. (a) Unchanged. (b) Rotated 30° . (c) rotated 60° . Black dots are Betti-0 events and triangles are Betti-1 events.	45
Figure 5.5:	Average Pearson correlation coefficient over the entire protein dataset fitting all 24 persistent homology features using various cutoff distances. . .	47
Figure 5.6:	The MWCG based deep convolutional neural network architecture used for B factor prediction. The plus symbol represents the concatenation of data sets. Figure originally published in Bramer <i>et al</i> [2].	50

Figure 5.7:	The deep learning architecture using a convolutional neural network combined with a deep neural network to predict B factor using PH based features. The plus symbol represents the concatenation of features. . . .	53
Figure 6.1:	Workflow for procedure in MWCG feature construction.	56
Figure 6.2:	Workflow for procedure in ASPH and ESPH feature construction.	57
Figure 6.3:	Workflow for procedure MWCG, ASPH, and ESPH based machine learning B factor prediction.	58
Figure 7.1:	(a) VMD representation of PBD ID 1AIE. (b) Correlation maps for nitrogen-nitrogen (NN) and (c) oxygen-oxygen (OO) interactions for protein 1AIE. The thicker band along the main diagonal of (b) and (c) corresponds to the alpha helix secondary structure in 1AIE. Figure originally published in Bramer <i>et al</i> [1].	61
Figure 7.2:	(a) VMD representation of PBD ID 1KGM. (b) Correlation maps for nitrogen-nitrogen (NN) and (c) oxygen-oxygen (OO) interactions for protein 1KGM. The bands perpendicular to the main diagonal of (b) and (c) correspond to the anti parallel beta sheet present in 1KGM. Figure originally published in Bramer <i>et al</i> [1].	62
Figure 7.3:	(a) VMD representation of PBD ID 5IIV. (b) Correlation maps for nitrogen-nitrogen (NN) and (c) oxygen-oxygen (OO) interactions for protein 5IIV. The presence of the two distinct thick bands along the main diagonal of (b) and (c) corresponds to the two alpha helices present in 5IIV. The off diagonal bands correspond to the bonding interaction between alpha helices. Figure originally published in Bramer <i>et al</i> [1].	63
Figure 7.4:	(a) A visual comparison of experimental B factors , (b) WCG predicted B factors, (c) and GNM predicted B factors for the ribosomal protein L14 (PDB ID:1WHI). (d) The experimental and predicted B factor values plotted per residue. GNM represents predicted B factors using GNM with a cutoff distance of 7 Å. WCG is parametrized using CC, CN, CO kernels of the exponential type with fixed parameters $\kappa = 1$, and $\eta = 3$ Å. Figure originally published in Bramer <i>et al</i> [1].	64
Figure 7.5:	(a) The structure of calmodulin (PDB ID: 1CLL) visualized in Visual Molecular Dynamics (VMD)18 and colored by experimental B factors, (b) MWCG predicted B factors, (c) WCG predicted B factors, (d) and GNM predicted B factors with red representing the most flexible regions. Figure originally published in Bramer <i>et al</i> [1].	65

Figure 7.5:	(Continued) (e) The experimental (Exp) and predicted B factor values plotted per residue for PDB ID 1CLL. The GNM is for the GNM method with a cutoff distance of 7 Å. We see that GNM clearly misses the flexible hinge region. WCG is parametrized using CC, CN, CO kernels of the exponential type with fixed parameters $\kappa = 1$, and $\eta = 3$ Å. MWCG represents B factor predictions determined from the MWCG method using the fixed parameters listed in Table 3.2. Figure originally published in Bramer <i>et al</i> [1].	66
Figure 7.6:	A visual comparison of experimental B factors (a), WCG predicted B factors (b), and GNM predicted B factors (c) for the engineered cyan fluorescent protein, mTFP1 (PDB ID:2HQK). (d) The experimental (Exp) and predicted B factor values plotted per residue for PDB ID 2HQK. The GNM is for the GNM method with a cutoff distance of 7 Å. WCG is parametrized using CC, CN, CO kernels of the exponential type with fixed parameters $\kappa = 1$, and $\eta = 3$ Å. Figure originally published in Bramer <i>et al</i> [1].	97
Figure 7.7:	CPU Efficiency comparison between GNM [3], RF, GBT, and CNN algorithms for MWCG B factor prediction. Execution times in seconds (s) versus number of residues. A set of 34 proteins, listed in Table 7.7, were used to evaluate the computational complexity. Result originally published in Bramer <i>et al</i> [2].	101
Figure 7.8:	Individual feature importance for the MWCG random forest model averaged over the data set. Reported feature selection includes the use heavy atoms in the model. Figure originally published in Bramer <i>et al</i> [2].	103
Figure 7.9:	Average feature importance for the MWCG random forest model with the angle, secondary, MWCG, atom type, protein size, amino acid, and packing density features aggregated. Reported feature selection includes the use heavy atoms in the model. Figure originally published in Bramer <i>et al</i> [2].	107

KEY TO ABBREVIATIONS

aFRI	Anisotropic Flexibility Rigidity Index
ANM	Anisotropic Network Model
ASPH	Atom Specific Persistent Homology
CNN	Convolutional Neural Network
DNN	Deep Neural Network
ESPH	Element Specific Persistent Homology
fFRI	Fast Flexibility Rigidity Index
FRI	Flexibility Rigidity Index
GBT	Gradient Boosting Trees
GNM	Gaussian Network Model
mFRI	Multiscale Flexibility Rigidity Index
MWCG	Multiscale Weighted Colored Graph
NMA	Normal Mode Analysis
NMR	Nuclear Magnetic Resonance
MD	Molecular Dynamics
PDB	Protein Data Bank
PH	Persistent Homology
RF	Random Forest
WCG	Weighted Colored Graph

Chapter 1

Overview

X-ray crystallography is an impressive experimental tool that provides three dimensional (3D) spatial coordinates and thermal fluctuation data of atoms within a crystallized molecule in the form of a PDB data file. Using data contained in a protein PDB file, one can validate mathematical models to understand protein dynamics and flexibility. The protein data bank is massive, containing over 140,000 structures as of March 2019, with more structures submitted annually.

Even with the solution of many protein structures, there is still an important need for robust and accurate mathematical models. Many important classes of proteins are difficult to crystallize and some may even prove to be impossible. Protein crystallization difficulty increases proportionally with the size of a protein. Highly flexible proteins represent another class of proteins that are difficult to crystallize due to their resistance in forming a crystal lattice structure. Other examples of proteins which are difficult to crystallize include small heat shock proteins, transmembrane and membrane proteins, and intrinsically disordered proteins. Heat shock proteins are an important class of proteins related to cardiovascular function, immunity, and cancer. Transmembrane and Membrane proteins are targets for the majority of modern drugs. Intrinsically disordered proteins are also vitally important to understand as they have been implicated in a number of diseases such as Bovine Spongiform Encephalopathy (mad cow disease), Creutzfeldt-Jakob disease, Alzheimer's disease, and

Parkinson’s disease.

In this work new and efficient methods for protein analysis are introduced that improve upon existing methods in several ways. These methods are the first protein B factor prediction methods to incorporate additional protein information from non- C_α atoms in the form of element specific interaction pairs. Moreover, this work introduces methods that are entirely new to B factor prediction. These methods are capable of successful cross protein B factor prediction using only information from other proteins. The methods presented use advanced graph theory based techniques, machine learning algorithms, and the first known topological data analysis based persistent homology method, to successfully analyze protein flexibility and dynamics. Lastly, the methods provide the best predictive results, to date, for both protein B factor prediction within a protein and cross protein B factor prediction. The results are validated through extensive testing on a large and diverse set of proteins. Using these methods many protein analysis tools can be constructed. In addition to the protein B factor prediction, several applications of these methods are provided in this work. Examples include hinge detection, element specific protein correlation maps, and protein model relative feature importance ranking.

This work first introduces an efficient and accurate advanced graph theory based multiscale weighted colored graph (MWCG) method for analyzing protein flexibility and dynamics. The weighted colored graph (WCG) theory is based on the hypothesis that the most fundamental properties of proteins are determined by the geometric structure of the protein. The WCG method does not require costly matrix diagonalization like other commonly used methods such as Normal Mode Analysis (NMA) and Gaussian Network Model (GNM). Given a protein of N atoms, the computational complexity of the WCG method is approximately $\mathcal{O}(N^2)$ whereas methods like Normal Mode Analysis and Gaussian Network

Model are $\mathcal{O}(N^3)$ due to the fact that they require diagonalization of a large matrix.

Next a multiscale formulation of the WCG method is introduced to incorporate the multiscale interactions that occur within a protein into the model. Protein interactions take place over a variety of different scales, so any reliable model should take this property into account. To reduce computational complexity, most elastic network models include a predefined cutoff distance. However, the computational cost saved by using a cutoff in ENM incurs a cost in the overall accuracy of such models. By prescribing a distance based cutoff these models fail to capture protein interactions that take place across multiple characteristic length scales. The MWCG model was developed to capture the multiscale behavior of protein interactions. To capture various interaction scales within a protein the MWCGs used in this work were parameterized using three correlation kernels parameterized at different length scales. However the method is general and adaptable, so the number of correlation kernels can be adjusted to fit the users performance needs.

To test the efficacy of the WCG approach, the method is tested on a set of over 300 protein structures taken from X-ray crystallography data provided by the Protein Data Bank. The accuracy of B factor prediction between MWCG is compared to the most commonly used approaches, parameter-free FRI (pfFRI), NMA, and GNM. Averaged over a large and diverse set of over 300 proteins the results demonstrate a significant improvement. Averaged over the entire protein test set, the MWCG method is over 28% more accurate than the best previous method opFRI and 42% more accurate than GNM. To further demonstrate the utility of the WCG method, applications such as element specific protein heat maps and hinge detection visualizations are included.

Accurate identification of hinge regions and hinge motion is an important topic that has been highly studied[5, 6, 7, 8, 9]. Hinge residue detection is integral for molecules that are

too large for MD simulation over meaningful time scales. In the past, methods such as GNN and NMA have been used to detect hinges for proteins where MD is intractable. This work compares the ability of GNM, WCG, and MWCG methods to identify the hinge regions of several proteins. The work demonstrates several instances where WCG and MWCG accurately identify hinge regions where GNM at the same time fails to do so. This highlights the overall efficacy of this method and the multiscale behavior captured by MWCG.

Element specific correlation maps provide a new way to visualize secondary and tertiary protein structure using a two dimensional (2D) image where flexibility is represented by the color of each pixel of the image. These correlation maps have been introduced in the past for C_α atoms [3]. In this work we introduce more general element specific correlation maps. Examples of nitrogen-nitrogen and oxygen-oxygen element interaction correlation maps are provided for several proteins. This demonstrates the adaptability of the WCG and MWCG methods presented here. The provided examples clearly demonstrate important secondary structures such as alpha helix and beta sheets as well as their primary and secondary interactions.

Previous protein B factor prediction methods are not capable of accurate prediction of B factor across proteins. The MWCG method, along with other engineered features are used to create machine learning based B factor prediction models. The model captures various interaction scales within an individual protein. To capture distinctions between proteins other global features such as protein resolution are included as feature inputs. The machine learning algorithms used in this work are trained using nine MWCG kernels with various parameterizations. Other local and global features are also included to improve the robustness of the feature set. The algorithms were trained using leave-one-protein-out cross validation, where the algorithm trains on all protein data except the protein of interest,

then the test set is taken to be the protein of interest. Extensive numerical testing indicates that the MWCG cross B factor predictions obtained are more accurate than any B factor prediction using existing traditional methods. The approach introduced here is particularly notable because it accurately predicts cross-protein B factors.

In recent years topological data analysis (TDA) has been successfully applied to protein analysis in a variety of areas. The basic idea of TDA is to use tools from topology to analyze high dimensional datasets that may be noisy or incomplete. Techniques from TDA reduce the dimensionality of the dataset, and allow the user their choice of metric. These techniques are a good fit for protein analysis, where one wants to infer high dimensional structure from low dimensional representations, capture multiple scales, and assemble discrete point data into a global structure. The point cloud of 3D spatial coordinates provided for proteins in the protein databank PDB files can be converted into a family of simplicial complexes. These simplicial complexes are indexed by a proximity parameter. Then, converting the dataset into global topological objects, tools from algebraic topology can be applied for protein analysis.

Persistent homology theory allows the persistent homology of a filtered simplicial complex to be uniquely represented with a barcode. In this work protein data is encoded into a barcode by taking a filtration over simplicial complexes that have been constructed from element specific protein spatial data. The protein barcodes provide global invariant topological features of the protein. By comparing two related barcodes for each atom of interest, this technique can be used to predict local atomic flexibility. The two barcodes are constructed such that one barcode is constructed using a point cloud that includes the atom of interest, and another is constructed using the same point cloud but without the atom of interest. The similarity or difference between barcodes is compared using bottleneck or wasserstein met-

rics. This provides atomic specific persistent homology protein flexibility analysis. Including various element interaction pairs one may also generate element specific persistent homology (ESPH) to capture element specific interactions. To the author's knowledge no previous protein flexibility models have used persistent homology to predict B factors of atoms in a protein in this way.

In this work ASPH and ESPH features are generated for each C_α of a given protein. However, this method is a general framework that can be applied to any element in a protein, including hydrogen. The method allows for several parameterizations that can be tuned by the user. To validate this approach several PH features are generated and used to fit B factor prediction models using linear least squares fitting. Later, the features are used with machine learning techniques. Both cases are validated using a large and diverse data set of proteins from the protein data bank. These results provide good predictions that are comparable to the aforementioned MWCG results.

Chapter 2

Background

Currently Nuclear Magnetic Resonance (NMR) Spectroscopy and X-ray crystallography are the two major experimental techniques used for protein dynamics and flexibility analysis. Techniques for NMR were previously very challenging but are now becoming more routine. At a basic level, NMR works by mapping the magnitude or intensity of magnetic resonance signals as a magnetic field is applied to a protein sample. X-ray crystallography determines protein structure by measuring the diffraction patterns of an intense beam of X-rays of a crystallized protein. The crystal is rotated many times and with each rotation a new set of diffraction patterns is collected. After tens of thousands of rotations, the data is combined and computationally processed into a final atomic arrangement known as the protein crystal structure.

At the time of this dissertation, over 90% of the protein data bank (PDB) files have been solved using X-ray crystallography while less than 10% have been solved using NMR. Unlike X-ray crystallography, NMR results do not provide atomic flexibility information. In contrast, X-ray crystallography data includes flexibility information in the form of atomic B factor (temperature factor, B value, or Debye-Waller factor), which is a measurement of the X-ray scattering of atoms or groups of atoms in a protein. Atomic B factor has been observed to correlate with atomic flexibility from Molecular dynamics (MD) and Normal mode analysis (NMA) experiments thus it provides a good experimental gold standard to

compare theoretical methods.

2.1 Computing Protein Flexibility and Dynamics

Many methods exist for studying protein structure and function; however, there is room for substantial improvement. Algorithms which require X-ray crystallography are limited by the availability of previously crystallized proteins. Surely the protein databank will continue to grow as scientists crystallize proteins with ever increasing efficiency. However, for many types of proteins, crystallization is very difficult or impossible. This calls for new approaches to theoretical protein analysis.

MD simulation is one method for protein analysis that has made a serious contribution to our understanding of the conformational landscapes of proteins. It has been particularly helpful in understanding proteins that are difficult to study experimentally such as amyloid fibrils, intrinsically disordered proteins, and partially disordered proteins. Even so, the dynamics of large proteins generally takes place over long time scales that are inaccessible to modern MD simulations. MD simulations are computationally intractable for larger macromolecules and in systems of multiple molecules as the time scales required are unreasonable for current technology. As such MD continues to be limited to systems of low complexity due to the methods high degree of freedom.

To address the limitations of time-dependant MD approaches several time independent approaches to protein dynamics and flexibility analysis have been developed. NMA was one of the first successful time-independent methods used for protein analysis[10, 11, 12, 13, 14]. NMA achieves time-independence by adopting an interaction Hamiltonian based on protein molecular mechanics. In this approach bond lengths and angles are fixed, and NMA

is computed by the diagonalization of a Hamiltonian on an energy minimized structure. Normal modes are the orthogonal resonant patterns of the molecular mechanic system. A superposition of the normal modes provides the collective motion of the protein. Low frequency modes correspond to cooperative motions and are meaningful in applications like hinge detection and MD where slow, collective motion is relevant. The transition pathways of macromolecules are also highly correlated with the low-frequency modes of NMA[14]. NMA provides good coarse grained deformation motion of supramolecular complexes. The success of NMA has resulted in several related methods that improve the computational cost and quality of the generated results.

The elastic network model (ENM) was proposed in 1996 as a simplified NMA approach[15]. The ENM is based on a statistical mechanics approach where a molecule is treated as a system of N nodes with each node corresponding to an atom or residue within the molecular network[16]. This approach provides good prediction of global motions but does not reliably predict local motion and requires costly diagonalization of the large corresponding Hessian matrix. The Anisotropic network model (ANM) was model introduced using the ENM framework to account for 3D directionality. The ANM uses a spring network with a simple spring potential between C_α atoms[17]. Given N atoms, ANM requires a $3N \times 3N$ matrix Diagonalization of the resulting Hessian. This provides the modes of the system that correspond to cooperative motions. Lower eigenvalue and eigenvectors can be used to estimate protein flexibility. In ANM all springs use the same force constant. The ANM provides good insight into the protein dynamics at a lower computational cost than other normal mode analysis based methods.

The Gaussian network model (GNM) is a related ENM developed around the same time as ANM that provides a good course grained, isotropic, low cost approach[18, 19]. In GNM

the Hessian is replaced by a Kirchoff matrix. The diagonalization of the Kirchoff matrix gives rise to eigenmodes and eigenvalues for describing protein fluctuations that correspond to B factors. GNM is both accurate and efficient compared to other previous approaches[20].

To bypass costly large matrix diagonalization the flexibility-rigidity index (FRI) was more recently introduced[21, 3, 22]. FRI is a mathematical method based on geometric graphs, that makes use of protein graph connectivity and node centrality to analyze protein flexibility. The method is based on the hypothesis that protein interactions and protein structure are inextricably linked in a given environment. That is, protein flexibility and function are determined by protein structure and environment. Since the FRI approach is not based on molecular mechanics it does not require a protein interaction Hamiltonian like those used in spectral graph theory, to analyze protein flexibility. The FRI approach works well as long as the accurate structure of the protein and its environment is known. As such FRI is restricted to proteins with solved 3D X-ray crystal structures. The FRI method provided a significant improvement in computational speed compared to previous protein analysis methods. The first FRI method [21] is of computational complexity $\mathcal{O}(N^2)$ [21]. Later fast FRI (fFRI) [3] was introduced to reduce computational cost further. The fFRI method is of computational complexity $\mathcal{O}(N)$. Anisotropic FRI (aFRI) [3] and generalized FRI (gFRI) [23] have also since been developed. To capture the multiscale interactions seen in macromolecules the multiscale FRI (mFRI) method was introduced[24]. Compared to GNM, the mFRI algorithm was shown to be approximately 20%, more accurate averaged over a large and diverse set proteins [24]. The fFRI algorithm was shown to be significantly faster than GNM[3]. Generalized GNM (gGNM), generalized ANM (gANM), multiscale GNM (mGNM), and multiscale ANM (mANM) methods have been recently constructed using FRI matrices [25]. These generalized algorithms provide major improvements to the

accuracy of original algorithms for protein flexibility analysis. A summary of when the different approaches to protein flexibility and dynamics were first introduced is provided in Table 2.1.

Table 2.1: Notable molecular mechanic techniques and the year of introduction.

Molecular Mechanics Technique	Year of Introduction
Molecular Dynamics (MD)	1977[26]
Normal Mode Analysis (NMA)	1982[11]
Elastic Network Model (ENM)	1996[15]
Gaussian Network Model (GNM)	1996[18]
Anisotropic Network Model (ANM)	2001[17]
Flexibility Rigidity Index (FRI)	2014[3]

While the previous methods provide good results, there is still room for significant improvement. The average pearson correlation coefficient of the B factor predictions of the aforementioned methods is generally below 0.7. Knowing the importance of protein flexibility analysis, it is crucial to improve these results. Moreover the above methods do not provide satisfactory results when predicting cross protein B factor. Given the the many classes of proteins with no X-ray crystal structure this is an important problem with no existing reliable solutions.

2.2 Data

Two data sets are used for testing and validation in this work: one from Refs. [3, 24] and the other from Park, Jernigan, and Wu [4]. The first data set contains 364 proteins [3, 24], and the second contains 3 subsets of small, medium, and large sized proteins [4]. All protein PDB structures have a resolution of 3 Å or higher and an average resolution of 1.3 Å. The PDB data sets include proteins that range in size from 4 to 3912 residues [4]. This work excludes protein 1AGN due to known data issues. Proteins 1NKO, 2OCT, and 3FVA are

also excluded as these proteins have PDB files with residues whose B factors are reported as zero which is nonphysical. For all machine learning results provided in this work, the STRIDE software is unable to provide the required secondary features for proteins 1OB4, 1OB7, 2OLX, and 3MD5 so these also excluded.

Chapter 3

Multiscale Weighted Colored Graphs

3.1 Weighted colored graphs

For this approach, each protein is considered to be a network in the form of a mathematical graph. That is, a protein a network where atoms represent nodes or vertices of the graph and edges are weighted connections between nodes that are determined by a distance based radial function. Colored graphs are constructed based on heavy element (carbon, nitrogen, oxygen, sulfur) interaction pairs. Provided it is available, one may even include hydrogen atoms. Hydrogen atoms have a high degree of uncertainty, and cannot be accurately measured by X-ray crystallography so we exclude them from this work. A graph is denoted as $G(V, E)$ where V represents a set of nodes called vertices and E the set of edges of the graph that relate vertices pairwise. This work defines a protein network to be a graph whose nodes and edges have specific attributes corresponding to the protein. In particular, individual atoms correspond to graph nodes, and the edges to a distance based correlation metric. This approach makes sense from a biophysical point of view since interaction strength is inversely proportional to distance. Further, many existing B factor prediction methods use three-dimensional (3D) networks of spatial atomic coordinate data from the protein databank.

The most basic component of this method is a weighted colored graph. A WCG converts 3D geometric protein spatial information, provided as atomic coordinates by a PDB data

file, into a protein connectivity network. All existing previous methods only take C_α atoms into consideration when constructing graph theoretic approaches. However, in this work all N atoms in a protein are considered. Given the colored graph $G(V, E)$, the i th atom is labeled by its element type α_j and position \mathbf{r}_j and thus

$$V = \{(\mathbf{r}_j, \alpha_j) | \mathbf{r}_j \in \mathbb{R}^3; \alpha_j \in \mathcal{C}; j = 1, 2, \dots, N\},$$

where $\mathcal{C} = \{C, N, O, S\}$ is the set containing the chosen element types of interest in a protein. The set of edges, \mathcal{P} , in a colored protein graph is defined to be the set of all element specific pairs of \mathcal{C} . This choice of \mathcal{C} results in 16 element directed interaction pairs. Table 3.1 illustrates the 16 possible element interaction pairs. For this work \mathcal{P} is defined to be

Table 3.1: Element pair combinations used in weighted colored graph.

	C	N	O	S
C	CC	CN	CO	CS
N	NC	NN	NO	NS
O	OC	ON	OO	OS
S	SC	SN	SO	SS

$$\mathcal{P} = \{CC, CN, CO, CS, NC, NN, NO, NS, OC, ON, OO, OS, SC, SN, SO, SS\}.$$

For example, the subset $\mathcal{P}_3 = \{CO\}$ contains all directed CO pairs in the protein such that the first atom is a carbon and the second one is a oxygen. Mathematically, E is the set of weighted directed edges describing the potential interaction pairs of atoms given by

$$E = \{\Phi^k(\|\mathbf{r}_i - \mathbf{r}_j\|; \eta_{ij}) | (\alpha_i \alpha_j) \in \mathcal{P}_k; k = 1, 2, \dots, 16; i, j = 1, 2, \dots, N\}, \quad (3.1)$$

where $\|\mathbf{r}_i - \mathbf{r}_j\|$ is defined to be the Euclidean distance between the i^{th} and j^{th} atoms, η_{ij} a characteristic distance between the atoms, and $(\alpha_i\alpha_j)$ a directed pair of element types. In this work Φ^k is a correlation function with the following properties [3]

$$\Phi^k(\|\mathbf{r}_i - \mathbf{r}_j\|; \eta_{ij}) = 1, \text{ as } \|\mathbf{r}_i - \mathbf{r}_j\| \rightarrow 0 \quad (\alpha_i\alpha_j) \in \mathcal{P}_k, \quad (3.2)$$

$$\Phi^k(\|\mathbf{r}_i - \mathbf{r}_j\|; \eta_{ij}) = 0 \text{ as } \|\mathbf{r}_i - \mathbf{r}_j\| \rightarrow \infty, \quad (\alpha_i\alpha_j) \in \mathcal{P}_k. \quad (3.3)$$

Previous work by Opron et al[3] has shown that generalized exponential functions of the form,

$$\Phi^k(\|\mathbf{r}_i - \mathbf{r}_j\|; \eta_{ij}) = e^{-(\|\mathbf{r}_i - \mathbf{r}_j\|/\eta_{ij})^\kappa}, \quad (\alpha_i\alpha_j) \in \mathcal{P}_k; \quad \kappa > 0, \quad (3.4)$$

and generalized Lorentz functions of the form,

$$\Phi^k(\|\mathbf{r}_i - \mathbf{r}_j\|; \eta_{ij}) = \frac{1}{1 + (\|\mathbf{r}_i - \mathbf{r}_j\|/\eta_{ij})^\nu}, \quad (\alpha_i\alpha_j) \in \mathcal{P}_k; \quad \nu > 0, \quad (3.5)$$

are good choices for correlation functions that satisfy the above properties.

3.2 WCG Centrality

Given a graph, centrality provides a measure of the importance of a node. Centrality is an important concept in graph theory that has a wide variety of applications including social network analysis, identification of critical genes, traffic flows, and epidemics[27, 28, 29]. There are several types of centrality measures. For example, the normalized closeness centrality [30] of node \mathbf{r}_i is defined as

$$\frac{1}{\sum_j \|\mathbf{r}_i - \mathbf{r}_j\|}$$

and the Harmonic centrality [31] of node \mathbf{r}_i in a connected graph is defined as

$$\sum_j \frac{1}{\|\mathbf{r}_i - \mathbf{r}_j\|}.$$

In this work the notion of Harmonic centrality is extended to subgraphs with weighted edges defined by generalized correlation functions. The generalized centrality metric used in this work is defined as

$$\mu_i^k = \sum_{j=1}^N w_{ij} \Phi^k(\|\mathbf{r}_i - \mathbf{r}_j\|; \eta_{ij}), \quad (\alpha_i \alpha_j) \in \mathcal{P}_k, \quad \forall i = 1, 2, \dots, N, \quad (3.6)$$

where w_{ij} is a weight function related to the element type. The WCG centrality in Equation (3.6) provides the atom specific rigidity index of the i^{th} atom. This is a measure of the stiffness of the i^{th} atom that corresponds to the k th set of contact atoms.

3.3 Weighted Colored Graph Flexibility Analysis

Given a rigidity index, its reciprocal function provides a corresponding measure of flexibility, or flexibility index. Thus the general flexibility index on subgraphs is given by

$$f_i^k = \frac{1}{\mu_i^k}, \quad (\alpha_i \alpha_j) \in \mathcal{P}_k, \quad \forall i = 1, 2, \dots, N. \quad (3.7)$$

Previous work by Ngyuen et al shows that other flexibility index forms work equally as well [23]. At each atom, the flexibility index corresponds to temperature fluctuation. Thus we

can model the B factor of the i th atom as

$$B_i^t = \sum_k c_k f_i^k + b, \quad \forall i = 1, 2, \dots, N, \quad (3.8)$$

where B_i^t represents the theoretically predicted B factor of the i^{th} atom. The coefficients c_k and b are determined by minimizing the linear system given by

$$\min_{c_k, b} \left\{ \sum_{i=1}^N \left| B_i^t - B_i^e \right|^2 \right\}, \quad (3.9)$$

where B_i^e is the experimentally measured B factor of the i^{th} atom.

3.4 Multiscale Weighted Colored Graph Flexibility Analysis

Macromolecular interactions consist of a complex interplay of short, medium, and long range interactions. Covalent bonds dominate short-range type interactions. Medium-range interactions consist mainly of hydrogen bonds, electrostatics and van der Waals interactions. Lastly, hydrophobicity is the main contributor to long-range molecular interactions. As such, a protein's flexibility is inherently connected to multiple characteristic length scales. This work proposes multiscale weighted colored graphs to characterize the multiscale interactions that exist within a protein. The flexibility of i^{th} atom at n^{th} scale corresponding to the k^{th} set of interaction atoms is given by

$$f_i^{k,n} = \frac{1}{\sum_{j=1}^N w_{ij}^n \Phi^k(\|\mathbf{r}_i - \mathbf{r}_j\|; \eta_{ij}^n)}, \quad (\alpha_i \alpha_j) \in \mathcal{P}_k, \quad (3.10)$$

where w_{ij}^n is an atomic type dependent parameter, $\Phi^k(\|\mathbf{r}_i - \mathbf{r}_j\|; \eta_{ij}^n)$ a correlation kernel, and η_{ij}^n a scale parameter. Minimization takes the form

$$\min_{c_k^n, b} \left\{ \sum_i \left| \sum_{k,n} c_k^n f_i^{k,n} + b - B_i^e \right|^2 \right\}, \quad (3.11)$$

where B_i^e are experimental B factors. In this work we construct three correlation kernels using two generalized Lorentz kernels and a generalized exponential kernel to capture multiple length scales. The method provided here is made parameter free by choosing appropriate values for η , ν , and κ .

Sulfur atoms play an important role in proteins but they are also very sparse in proteins. As such, this work provides some results using sulfur atoms but for most of the testing provided sulfur atoms are excluded as they have a negligible overall effect on the model. Thus, unless otherwise noted this work considers the following subset of \mathcal{P} for the lion's share of computations.

$$\hat{\mathcal{P}} = \{ \text{CC, CN, CO, NC, NN, NO, OC, ON, OO} \}. \quad (3.12)$$

This work chooses to focus on C, N, and O due to their high occurrence in proteins and important biological relevance. However, it should be noted that the general method presented here can be adapted to include any element the user chooses. For WCG calculations of B factor predictions all possible element pairs, SC, SN, SO, and SS are considered.

This method is unique compared to other B factor prediction methods. The WCG method considers not only C_α interactions but the effects of interactions between nitrogen, oxygen, and other non- C_α carbon atoms. For this work, three element specific correlation kernels

are constructed for all carbon-carbon (CC), carbon-nitrogen (CN), and carbon-oxygen (CO) interactions within a protein. To capture multiscale interactions this work also includes three different scale parameterizations for each kernel. In total this generates 9 correlation kernels to characterize element specific multiscale protein interactions in terms of their corresponding graph centralities and atomic flexibility. The result of this method can be used directly, fitted using linear least squares, or as a machine learning feature. Previously existing methods such as mFRI, GNM, and NMA fail to take into account the element specific interactions that the WCG method presented here captures. Since this method provides a general framework for any element, in addition to carbon, WCG can also be used to predict the B factor of any heavy element.

3.5 Parameterization

In this work a total of 9 unique correlation kernels are used based on the CC, CN, and CO element specific correlation kernels described in Eq. (3.10). For simplification purposes, all B factor prediction computed in this work through fitting and machine learning uses $w_{ij} = w_{ij}^n = 1$ and $\eta_{ij}^n = \eta^n$.

A basic grid search over the 364 dataset determined the near optimal parameters for MWCG based C_α B factor predictions. Three kernels are used with $\nu = \{1, 3\}$ for Lorentz kernels and $\kappa = 1$ for the Exponential kernel, respectively. To improve the efficiency, a radial cutoff distance may be used. However, the WCG fitting and MWCG based machine learning results presented in this work do not use a cutoff.

The first kernel considered is a Lorentz function, and its near optimal η^1 was found to be $\eta^1 = 16$ as shown in Fig. 3.1. Then, fixing $\eta^1 = 16$, a parameter grid search is used

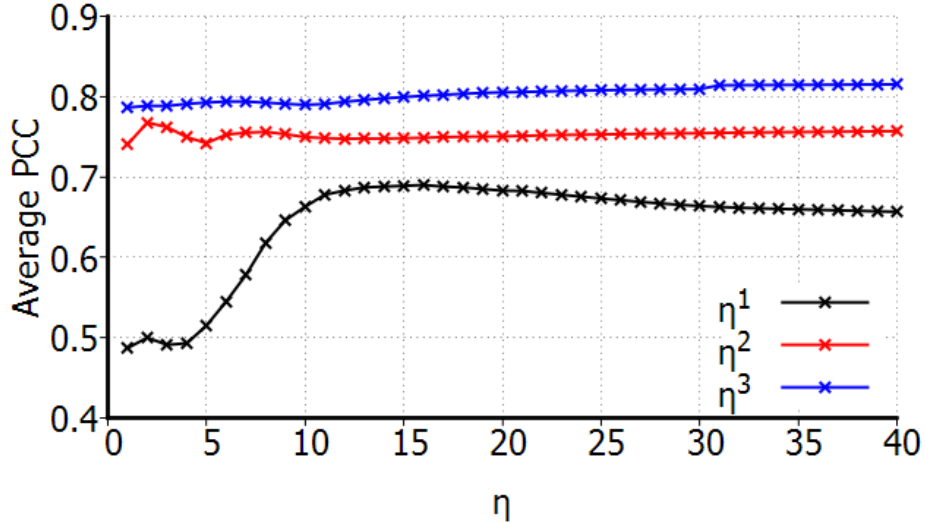


Figure 3.1: The average Pearson correlation coefficient (PCC) as found by optimizing individual kernels in the range of $\eta^n = 1, \dots, 40$. Parameter optimization results originally published in Bramer *et al* [1].

Table 3.2: Parameters used for correlation kernels in a parameter-free MWCG. Parameter optimization results originally published in Bramer *et al* [1].

Kernel Type	κ	η^n	ν
Lorentz ($n = 1$)	-	16	3
Lorentz ($n = 2$)	-	2	1
Exponential ($n = 3$)	1	31	-

to determine optimal η^2 for a second Lorentz kernel. The second Lorentz kernel was found to provide optimal predictions for $\eta^2 = 2$ as shown in Fig. 3.1. Lastly, fixing $\eta^1 = 16$ and $\eta^2 = 2$, a parameter search is used to determine optimal values for η^3 used in an exponential kernel. Given the fixed parameters of the Lorentz kernel the average Pearson correlation coefficient (PCC) does not decay even for very large values of η^3 as indicated in Fig. 3.1. Given the multiscale nature of these three parameters this behavior is reasonable. With only a single kernel, the strongest interactions, which provide good approximations, can be obtained for $12 \leq \eta \leq 17$. To capture close range interactions, the second η provides the best results for small values. The large values seen in the third η appear to capture

large scale interactions. This result corresponds to the dominance of these length dependent interaction types. Because it decays so quickly, the exponential kernel is used to capture large scale interaction effect. Of course large η values are very costly due to the structure of the kernel. So for η^3 a value of 31 is used in the testing published in Bramer *et al* [1, 2] for the parameter-free MWCG method as listed in Table 3.2.


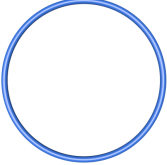
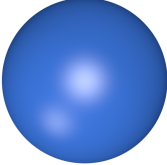
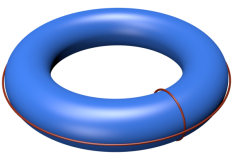
Chapter 4

Atom Specific Persistent Homology

4.1 Overview

Most existing protein analysis methods are structure or geometry based models. Many of these models struggle with the high dimensional space of protein data. Put another way, any model that is too fine grained will inherently fail in the high dimensional protein data space due to the associated computational complexity. The study of topology provides the connectivity of components, and characterizes independent entities, rings, and high dimensional topological faces within a space. Applied to proteins, topology provides a powerful tool for analysis of several important biological processes. Examples include hot spot detection, assembly/disassembly of virus capsids, ligand binding state, ion channel state, and protein folding[32, 33, 34, 35, 36, 37, 38, 39]. Topology provides a high level of abstraction and in its purely mathematical form is free of metrics of coordinates which can be problematic for the study of biological macro-molecules. Topological data analysis allows the extraction of invariant features that are embedded in the high dimensional data space of biomolecules. Persistent homology is one component of TDA that provides useful bridge between the high dimensional protein data space and the abstract low dimensional topological analysis of the protein data space. PH embeds multiscale geometric information into topological invariants, this works well for the aforementioned examples but oversimplifies the atomic properties of

Table 4.1: Topological invariants displayed as Betti numbers. Betti-0 represents the number of connected components, Betti-1 the number of tunnels or circles, and Betti-2 the number of cavities or voids. Two auxiliary rings are added to the torus to illustrate that Betti-1=2.

				
Example	Point	Circle	Sphere	Torus
Betti-0	1	1	1	1
Betti-1	0	1	0	2
Betti-2	0	0	1	1

macro-molecules making it challenging to use directly for atomic level analysis. In this work we provide a new approach that uses techniques from topological data analysis to provide element specific protein analysis at atomic resolution.

To apply TDA techniques, data must first be described as a simplicial complex or a graph network. Specifically, simplicial homology is concerned with the identification of topological invariants from a set of discrete nodes such as the atomic coordinates of a protein. Given a point cloud, Betti numbers describe the topological variants of connected components, rings, and cavities. Table 4.1 provides examples of the Betti-0, Betti-1, and Betti-2 numbers of a point, circle, sphere, and torus.

To determine topological invariants, a simplicial complex, such as Vietoris-Rips (VR) complex, Čech complex, or an alpha complex is constructed using a fixed filtration parameter. The simplicial complex is made up of vertices, edges, triangles, and tetrahedrons, denoted 0-simplex, 1-simplex, 2-simplex, and 3-simplex respectively. Basic examples are provided in Figure 4.1. By varying the filtration parameter over an interval a persistence diagram can be generated from a simplicial complex. A persistence diagram, or barcode, provides the birth and death (appearance and cessation) of Betti features for each node. The difference between

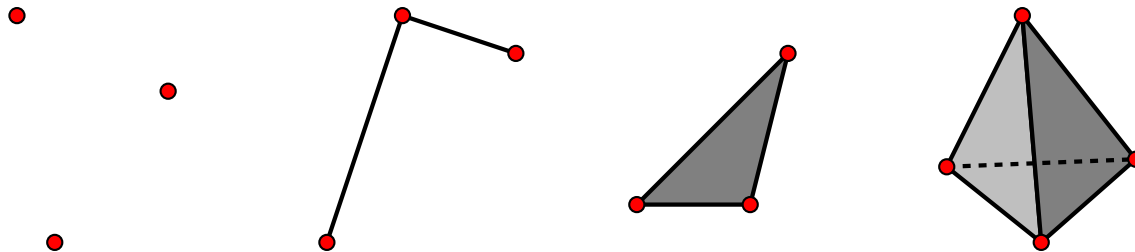


Figure 4.1: From left to right an example of a 0-simplex, 1-simplex, 2-simplex, and 3-simplex.

two persistence diagrams can be compared using Bottleneck and Wasserstein distances.

The main idea of atom-specific persistent homology and element-specific persistent homology is to extract atomic molecular information using global persistent homology techniques. To generate an atom-specific description using a global topological description we construct a pair of conjugated point clouds for each atom of interest. One point cloud is centered about the original atom of interest and all nearby atoms within a prescribed radial cutoff. The conjugate point cloud consists of the same point cloud minus the atom of interest. Then for each atom of interest, Bottleneck and Wasserstein distances are computed between the corresponding conjugate pairs which provides the desired topological information of each atom.

4.2 Simplex & Simplicial Complex

A simplex is a generalization of a triangle or tetrahedron to arbitrary dimensions. A k -simplex is a convex hull of $k + 1$ vertices represented by a set of affinely independent points

$$\sigma = \{\lambda_0 u_0 + \lambda_1 u_1 + \dots + \lambda_k u_k \mid \sum \lambda_i = 1, \lambda_i \geq 0, i = 0, 1, \dots, k\}, \quad (4.1)$$

where $\{u_0, u_1, \dots, u_k\} \subset \mathbb{R}^k$ is the set of points, σ is the k -simplex, and constraints on λ_i 's ensure the formation of a convex hull. A convex combination of points can have at most

$k + 1$ points in \mathbb{R}^k . For example a 1-simplex is a line segment, a 2-simplex a triangle, and a 3-simplex a tetrahedron. A subset of the $k + 1$ vertices of a k simplex with $m + 1$ vertices forms a convex hull in a lower dimension and is called an m -face of the k -simplex. An m -face is proper for $m < k$. The boundary of a k -simplex σ , is defined as the formal sum of its $(k - 1)$ faces. Given as

$$\partial_k \sigma = \sum_{i=0}^k [u_0, \dots, \hat{u}_i, \dots, u_k]^k (-1)^i [u_0, \dots, \hat{u}_i, \dots, u_k]^k, \quad (4.2)$$

where $[u_0, \dots, \hat{u}_i, \dots, u_k]$ denotes the convex hull formed by vertices of σ with the vertex u_i excluded and ∂_k is called the boundary operator. A collection of finitely many simplices forms a simplicial complex denoted by \mathcal{K} . All simplicial complexes satisfy the following conditions.

1. Faces of any simplex in \mathcal{K} are also simplices in \mathcal{K} .
2. The intersection of any two simplices $\sigma_1, \sigma_2 \in \mathcal{K}$ is a face of both σ_1 and σ_2 .

4.3 Homology

Given a simplicial complex \mathcal{K} , a k -chain c_k of \mathcal{K} is a formal sum of the k -simplices in \mathcal{K} with k no greater than dimension of \mathcal{K} and is defined as $c_k = \sum a_i \sigma_i$ where σ_i are the k -simplices and a_i 's coefficients. Generally, a_i can be in any field such as \mathbb{R} , \mathbb{Q} , or \mathbb{Z} . Here we choose a_i to be in \mathbb{Z}_2 for simplicity. Let the group of k -chains in \mathcal{K} be denoted by C_k . Then (C_k, \mathbb{Z}_2) forms an Abelian group under addition in modulo two. This allows us to extend the definition of the boundary operator introduced in Equation 4.2 to chains.

The boundary operator applied to a k -chain c_k is defined as

$$\partial_k c_k = \sum a_i \partial_k \sigma_i, \quad (4.3)$$

where σ_i 's are k -simplices. The boundary operator is a map from \mathcal{C}_k to \mathcal{C}_{k-1} , which is also known as a boundary map for chains. Note that operator ∂_k satisfies the property that $\partial_k \circ \partial_{k+1} \sigma = 0$ for any $(k+1)$ -simplex σ following the fact that any $(k-1)$ -face of σ is contained in exactly two k -faces of σ . The chain complex is defined as a sequence of chains connected by boundary maps with decreasing dimension and is denoted

$$\dots \rightarrow \mathcal{C}_n(\mathcal{K}) \xrightarrow{\partial_n} \mathcal{C}_{n-1}(\mathcal{K}) \xrightarrow{\partial_{n-1}} \dots \xrightarrow{\partial_1} \mathcal{C}_0(\mathcal{K}) \xrightarrow{\partial_0} 0. \quad (4.4)$$

The k -cycle group and k -boundary group are then defined as kernel and image of ∂_k and ∂_{k+1} respectively, and

$$\mathcal{Z}_k = \text{Ker} \partial_k = \{c \in \mathcal{C}_k \mid \partial_k c = 0\}, \quad (4.5)$$

$$\mathcal{B}_k = \text{Im} \partial_k = \{\partial_k c \mid c \in \mathcal{C}_k\}, \quad (4.6)$$

where \mathcal{Z}_k is the k -cycle group and \mathcal{B}_k is the k -boundary group. Since $\partial_k \circ \partial_{k+1} = 0$, we have $\mathcal{B}_k \subset \mathcal{Z}_k \subset \mathcal{C}_k$. Then the k -homology group is defined to be the quotient group of the k -cycle group modulo the k -boundary group,

$$\mathcal{H}_k = \mathcal{Z}_k / \mathcal{B}_k \quad (4.7)$$

where \mathcal{H}_k is the k -homology group. The k th Betti number is defined to be rank of the

k -homology group as $\beta_k = \text{rank}(\mathcal{H}_k)$.

4.4 Filtration & Persistence

For a simplicial complex \mathcal{K} , we define a filtration of \mathcal{K} as a nested sequence of sub-complexes of \mathcal{K} ,

$$\emptyset \subseteq \mathcal{K}_0 \subseteq \mathcal{K}_1 \dots \subseteq \mathcal{K}_n = \mathcal{K} \quad (4.8)$$

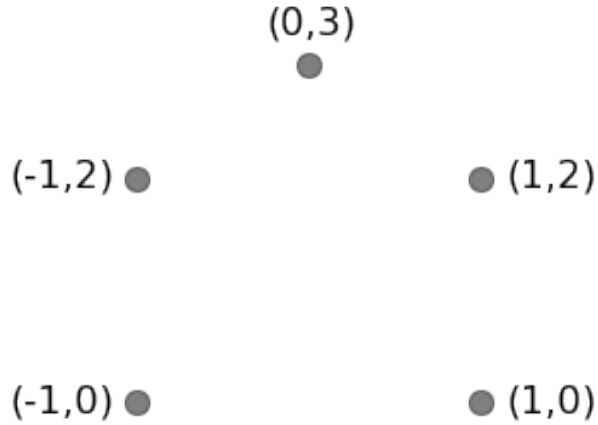
In persistent homology, the nested sequence of sub-complexes usually depends on a filtration parameter. The persistence of a topological feature is denoted graphically by its life span with respect to filtration parameter. Sub-complexes corresponding to various filtration parameters offer the topological fingerprints over multiple scales. The k^{th} persistent Betti numbers $\mathcal{B}_k^{i,j}$ represent the ranks of the k^{th} homology groups of \mathcal{K}_i that are alive and are defined as

$$\mathcal{B}_k^{i,j} = \text{rank}(\mathcal{H}_k^{i,j}) = \text{rank}(\mathcal{Z}_k(\mathcal{K}_i) / (\mathcal{B}_k(\mathcal{K}_j) \cap \mathcal{Z}_k(\mathcal{K}_i))). \quad (4.9)$$

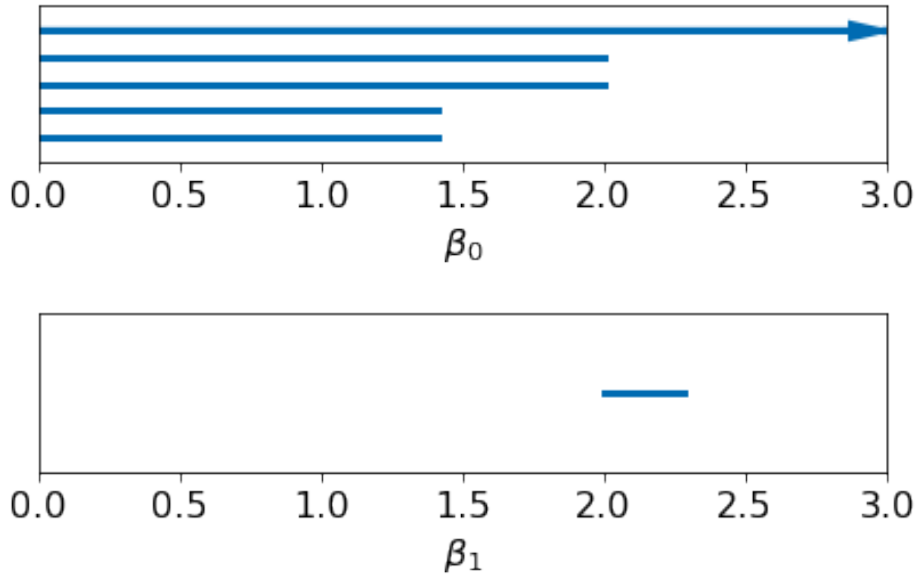
respectively where X and Y are persistence barcodes and $B_{ij}(X, Y)$ the collection of all bijections from X to Y . An example of a barcode is provided in Figure 4.2.

4.5 Similarity and distance

In this work, both Bottleneck and Wasserstein distances are used to compare conjugate persistence diagrams. This provides the models with atom-specific topological information and facilitates atom-specific persistent homology. Let X and Y be multisets of data points,



(a) Example Points



(b) Barcode

Figure 4.2: (a) An example of 5 points in \mathbb{R}^2 and (b) the corresponding topological barcode. The length of each barcode corresponds to the persistence of each topological object ($\beta_0, \beta_1, \beta_2, \text{etc.}$) over the filtration.

the Bottleneck and Wasserstein distances of X and Y are given by [40]

$$d_B(X, Y) = \inf_{\gamma \in B(X, Y)} \sup_{x \in X} \|x - \gamma(x)\|_\infty, \quad (4.10)$$

and [41]

$$d_W^p(X, Y) = \left(\inf_{\gamma \in B(X, Y)} \sum_{x \in X} \|x - \gamma(x)\|_\infty^p \right)^{1/p}, \quad (4.11)$$

respectively. Here $B(X, Y)$ is the collection of all bijections from X to Y . In this work topological invariants of different dimensions are compared separately.

4.6 Vietoris-Rips Complex

Given a metric space M and a cutoff distance d , a simplex is formed if all points have pairwise distances no greater than d . All such simplices form the Vietoris-Rips complex.

The abstract nature of the VR complex allows the construction of simplicial complexes for correlation function based metric spaces, which models pairwise interaction of atoms using correlation functions versus more standard spatial metrics.

4.7 Atom Specific Persistent Homology & Element Specific Persistent Homology

To embed the chemical biological protein information into topological invariants, element-specific persistent homology was introduced by Cang *et al*[42, 43]. The basic idea of ESPH is to use subset of atoms of various element types within a protein to construct topological representations. The corresponding persistence diagrams then represent different interactions that occur within a protein. For example selecting all carbon atoms would result in barcodes that coded the network and strength of the hydrophobicity in the protein.

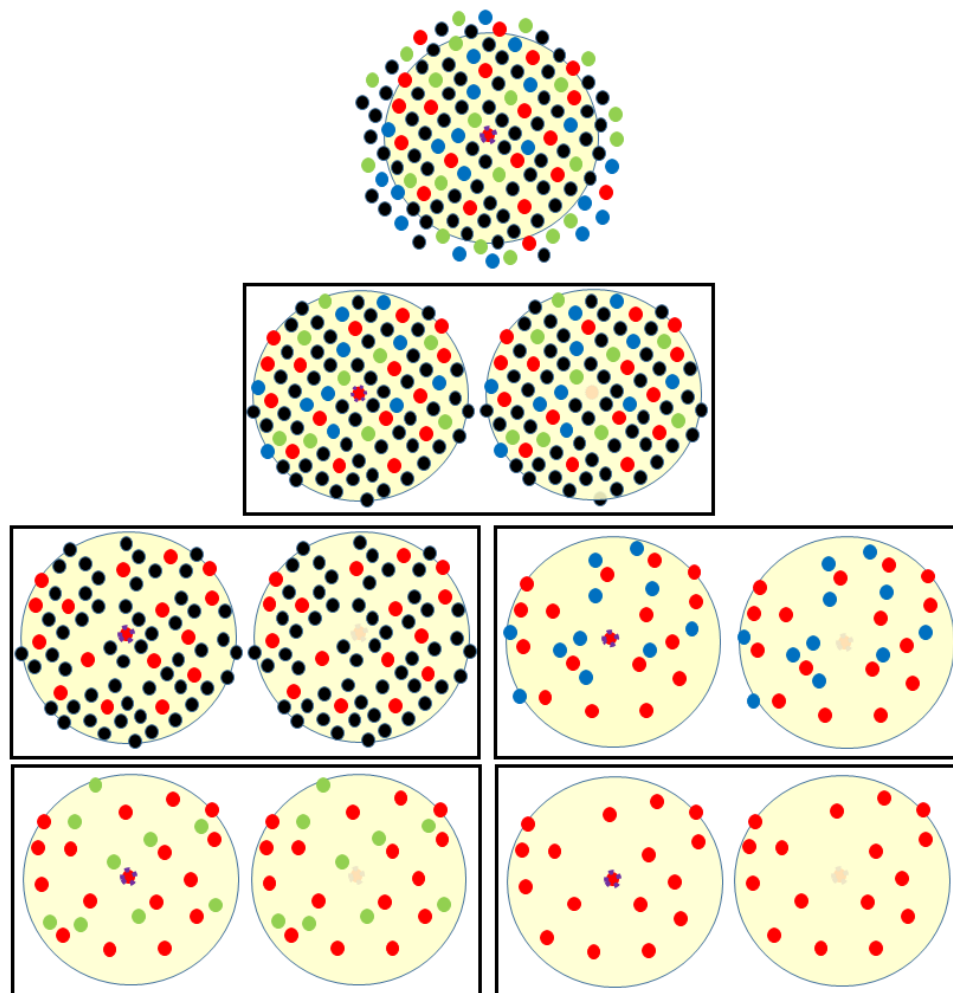


Figure 4.3: Illustration of Atom-specific persistent homology point clouds. Top: the original point cloud. The atom of interest is at the center of the circle. Second row: a pair of conjugated sets of point clouds for atom-specific persistent homology. The rest: Four pairs of conjugated point clouds for atom-specific and element-specific persistent homology.

To represent the topological importance of a given *atom*, atom-specific persistent homology is introduced. This works by constructing two conjugated point clouds centered about a given atom of interest within a biomolecule. The point clouds consists of one that includes the atom of interest and all nearby atoms within a prescribed cutoff, and another identical point cloud minus the atom of interest. Then, conjugated simplicial complexes, conjugated homology groups and conjugated topological invariants are generated for each

conjugate pair of points clouds. Wasserstein and Bottleneck distances can then be used to measure the difference between conjugated topological invariants which provides a topological representation of the atom of interest. Figure 4.3 provides an example of atom-specific and element-specific conjugated point clouds can be constructed for a given toy dataset.

This work generates only C_α B factor predictions however the method is general and can be used to predict the B factor of any atom. To create a diverse topological representation for each C_α element specific persistent homology is used. Atom-specific persistent homology is also used to contribute a precise topological representation at each C_α atom. Using the conjugate pair subsets, Vietoris-Rips complexes are constructed by contact maps or matrix filtration [44].

To capture element-specific interactions three subsets of carbon-carbon, carbon-nitrogen, and carbon-oxygen point clouds are used. This gives the following element specific pairs,

$$\mathcal{P} = \{\text{CC}, \text{CN}, \text{CO}\}. \quad (4.12)$$

For a given Protein Data Bank (PDB) file, persistence barcodes are calculated as follows. Given a specific C_α of interest, $\mathbf{r}_i^k \in \mathcal{P}_k$ in an element specific set \mathcal{P}_k ($\mathcal{P}_1 = \text{CC}$, $\mathcal{P}_2 = \text{CN}$, and $\mathcal{P}_3 = \text{CO}$), a point cloud consisting of all atoms within a pre-defined cutoff radius r_c is defined as

$$\mathcal{R}_i^k = \{\mathbf{r}_j^k \mid \|\mathbf{r}_i^k - \mathbf{r}_j^k\| < r_c, \quad \mathbf{r}_i^k, \mathbf{r}_j^k \in \mathcal{P}_k, \forall j \in 1, 2, \dots, N\}, \quad (4.13)$$

where N is the number of atoms in the k th element pair \mathcal{P}_k . A conjugated set of point cloud, $\hat{\mathcal{R}}_i^k$, includes the same set of atoms, except for \mathbf{r}_i^k . For a given pair of conjugated point clouds \mathcal{R}_i^k and $\hat{\mathcal{R}}_i^k$, conjugated simplicial complexes, conjugated homology groups, and conjugated persistence barcodes are computed. Euclidean distance based filtration is

computed using the Vietoris-Rips complex. Given set of atoms selected according to atom-specific and element specific constructions, a family of multi-resolution persistence barcodes is generated by a resolution controlled filtration matrix given by [44]

$$M_{nm}(\vartheta) = 1 - \Phi(\|\mathbf{r}_n - \mathbf{r}_m\|; \vartheta), \quad (4.14)$$

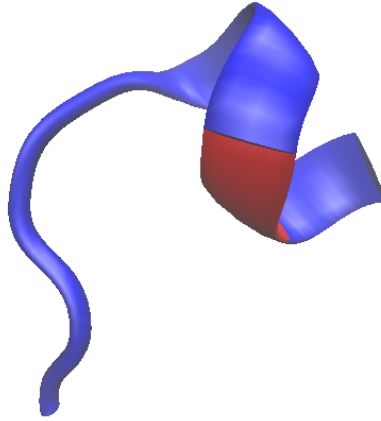
where ϑ denotes a set of kernel parameters. We have used both exponential kernels

$$\Phi(\|\mathbf{r}_n - \mathbf{r}_m\|; \eta, \kappa) = e^{-(\|\mathbf{r}_n - \mathbf{r}_m\|/\eta)^\kappa}, \quad \kappa > 0 \quad (4.15)$$

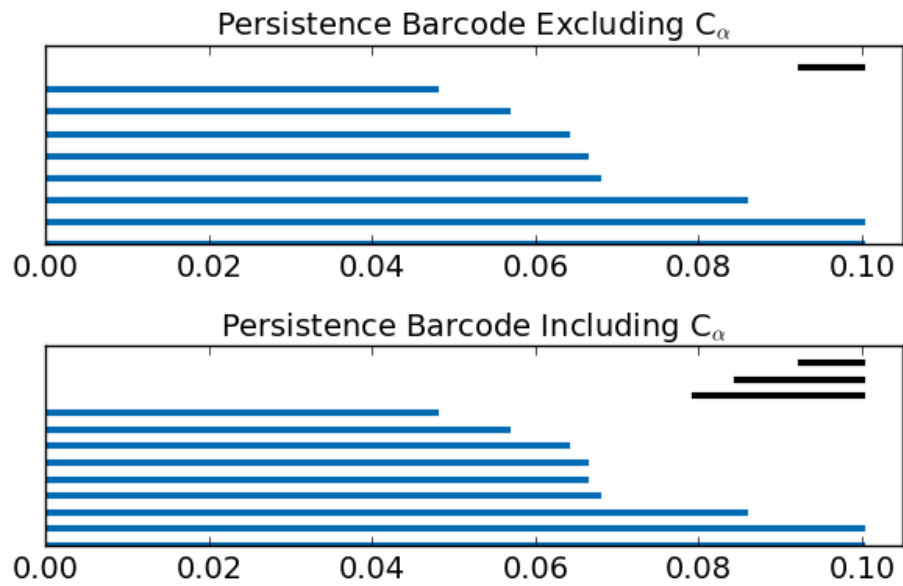
and Lorentz kernels

$$\Phi(\|\mathbf{r}_n - \mathbf{r}_m\|; \eta, \nu) = \frac{1}{1 + (\|\mathbf{r}_n - \mathbf{r}_m\|/\eta)^\nu}, \quad \nu > 0 \quad (4.16)$$

where η , κ , and ν are pre-defined constants. This filtration matrix is used in association with the Vietoris-Rips complex to generate persistence barcodes or persistence diagrams. These topological invariants are then compared using both Bottleneck and Wasserstein distances. An example of the conjugated persistence barcode pair generated for a C_α atom is illustrated in Figure 4.4.



(a) 1AIE Subunit



(b) Barcode

Figure 4.4: Illustration of residue 338 C_α atom-specific persistent homology in the CC element-specific point cloud of protein PDBID 1AIE. For this example residues 332-339 are used and are shown on the left. The C_α location used to generate the barcodes (right) is highlighted in red in the left chart. Conjugated persistence barcodes are generated with and without the selected C_α.

Chapter 5

Machine Learning

Machine learning is a subset of artificial intelligence based on statistical and probabilistic methods to “learn” patterns in data given a training set. This means that unlike other mathematical models, the structure of the algorithm is not known *a priori*. Broadly speaking, machine learning tasks are classified into supervised, semi-supervised, or unsupervised learning. Supervised learning involves training on data that contains both input data and some desired output data, semi-supervised training on data where some of the outputs are unknown, and unsupervised training on data without known output. Supervised and semi-supervised algorithms can then be trained for regression or classification tasks depending on the desired output. Since they have no target output, unsupervised algorithms can only find structure in data such as in the clustering or grouping data.

Machine learning algorithms differ by their internal representation. These algorithms are first classified as parametric or non-parametric depending on whether they have fixed number of parameters regardless of sample size, or whether the number of parameters is allowed to grow with sample size respectively. In practice parametric machine learning algorithms are computationally fast, require less data, and easy to implement compared to non-parametric machine learning algorithms. However, parametric machine learning algorithms can suffer from poor fitting due to overly strong assumptions about the underlying mapping function. In contrast non-parametric machine learning models are able to fit a

larger variety of functional forms and can thus produce more robust models.

The work by Wolpert et al suggests that learning algorithms cannot be universally good[45]. That is, a machine learning algorithm that provides a good model for one problem may not work for a different problem. As such, it is standard practice when using machine learning, to test several different machine learning algorithms to determine which of the algorithms are best suited to the problem.

The task of B factor prediction is a supervised regression task. It is supervised because B factors are known from experimental data and the prediction task is regression because B factor takes continuous values. Taking the aforementioned considerations into mind, this work considers several non-parametric machine learning algorithms. In particular, random forests, gradient boosted trees, convolutional neural networks, and deep neural networks are all considered in this work. All machine learning results are reported in Chapter 7. The following sections provide a detailed description of the algorithms, feature inputs, parameterizations, and datasets used for testing.

5.1 Machine Learning Algorithms

The following subsections provide a brief overview of each of type of machine learning algorithm used in this work.

5.1.1 Ensemble Methods

Ensemble methods are a class of machine learning algorithms that generate a strong predictive model based on a large number of simple weak learning models. The basic idea is that taken together, a large number of weak learners, those who do only slightly better

than chance, can generate a robust predictive model. Two of the most popular ensemble algorithms, which are used in this work, are random forests of trees and gradient boosting trees[46, 47, 48, 49].

5.1.1.1 Random forest

Random forests are an ensemble machine learning method used for classification or regression tasks. For regression tasks random forests train many decision trees then output the mean prediction of the individual trees. Compared to other machine learning algorithms, random forests are advantageous because they have few hyper-parameters, are generally robust against overfitting, and invariant to scaling.

Machine learning approaches are commonly criticized as “black box” approaches. That is, while the input and output of a machine learning algorithm are well known the internal model the algorithm is using is generally hidden to the user. Ensemble methods like random forests address this issue in part by providing variable importance of the trained model. Variable importance is one important way that users can understand which features give the model the most predictive power. Random forests are invariant to scaling, so they do not require the feature data to be pre-processed.

Random forests require minimal hyperparameter tuning. The only hyper parameter required is the number of n decision trees. While random forests are generally robust to overfitting if n is chosen to be too large it is possible for a random forests to overfit a dataset. Thus too few trees and the model will have poor predictive power and too many trees may lead to overfitting and be computationally costly. The user must take special care to determine the right amount of decision trees. For this work, the choice of decision trees was determined by testing various values of n to strike a balance between performance and

cost.

5.1.1.2 Gradient boosted trees

Like random forests, gradient boosting trees (GBTs) are an ensemble method. GBTs incorporate boosting to reduce bias and variance and utilize a number of “weak learners” to iteratively construct a predictive model. The algorithm is optimized using gradient descent, minimizing the residual of a predefined loss function. At each step, GBTs incorporate decision trees to improve their predictive power. Gradient boosting trees and other related ensemble methods are useful because they have strong predictive power, do not require normalization of the dataset, and are typically robust to outliers and overfitting.

5.1.2 Neural Networks

Recent advances in GPU computing have allowed neural networks to be computationally tractable machine learning models. Modeled after neurons in the brain, neural networks apply layers of activation functions, called perceptrons, to inputs. Weights of the neural network are trained to minimize a loss function over many passes of a training dataset. Many neural networks utilize back-propagation, which allows the error to propagate to the previous layer, to adjust neuron weights and improve output error until it is below a preset threshold. In short, neural networks begin with an initial random guess at an output then repeatedly adjust the neuronal weights until the output error is satisfactorily reduced. Neural networks with several “hidden” layers of perceptrons are known as deep neural networks (DNNs). Figures 5.1 and 5.2 provide examples of the basic perceptron and deep neural network framework.

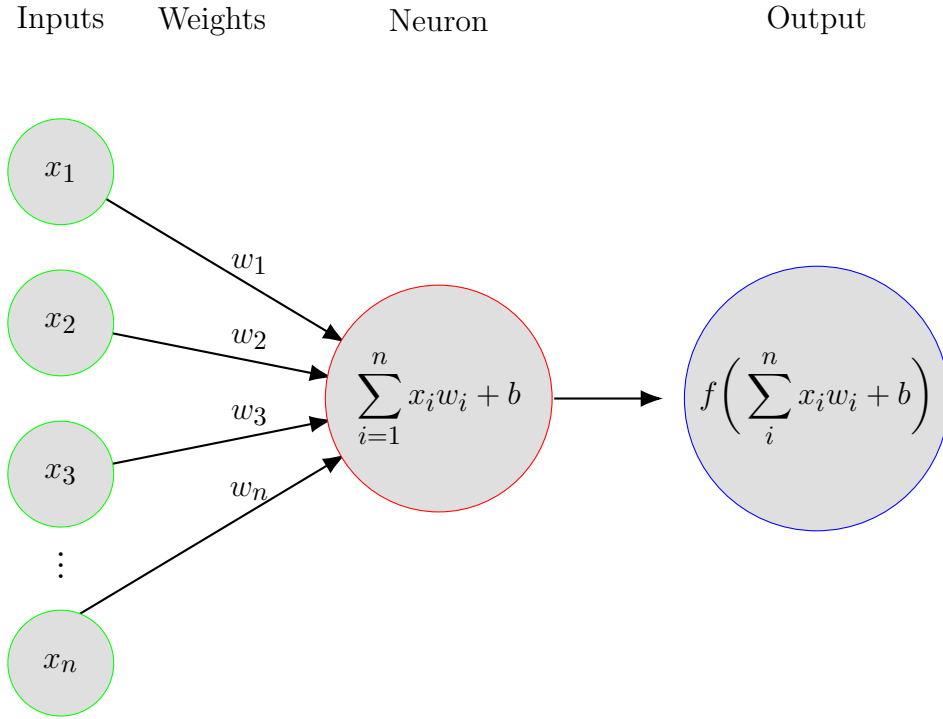


Figure 5.1: An example of a perceptron, the basic functional unit of a neural network.

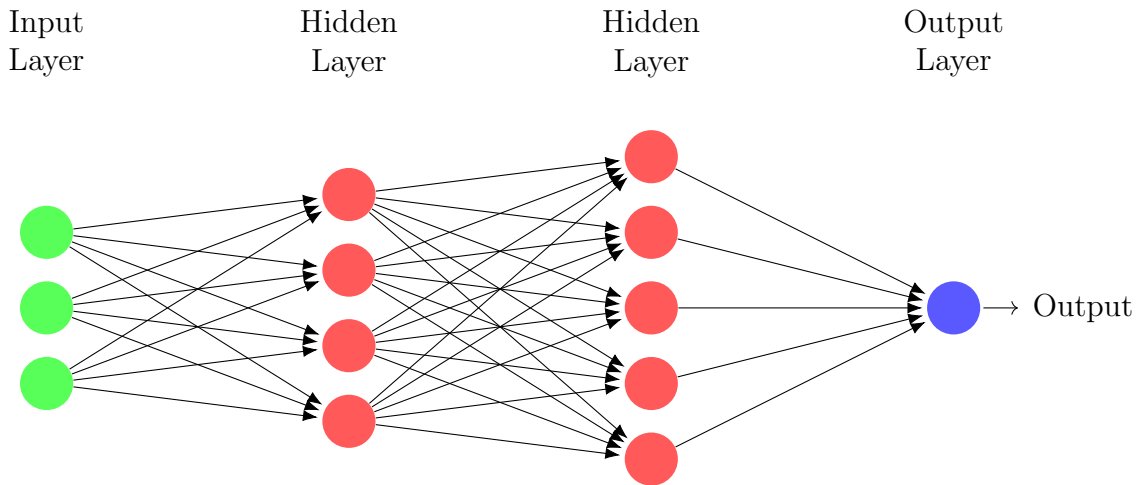


Figure 5.2: An illustration of a fully connected deep neural network. Circles represent neurons and connections between neurons are indicated by arrows. Each connection has an associated weight. A neural network is considered “deep” when it uses several hidden layers.

5.1.2.1 Convolutional Neural Network

Convolutional neural networks (CNNs) are a type of neural network that have recently had great success in the field of image classification. CNNs work by applying convolutional

filters over several layers, and by doing so extract successively higher-level features from input images. For image data CNNs are more advantageous than fully connected neural networks because they can often outperform fully connected neural networks with a fraction of training parameters.

5.1.3 Consensus methods

It is often the case that one machine learning model may outperform others in certain areas but do worse in others. As such a consensus model can provide a useful tool that may improve overall results. As such, for PH based C_α only B factor prediction, this work also includes B factor prediction results using a consensus model. The consensus model prediction used here is generated by combining the B factor predictions of the two PH based machine learning models. In particular, the consensus prediction for each C_α is the average of C_α B factor values predicted from the PH based GBT and deep CNN B factor prediction.

5.2 General Machine Learning Features

3D spatial atomic coordinates of each atom in a protein are provided by Protein Databank (PDB) .pdb files. The PDB files also provide additional experimental data that can be used as local and global input features for machine learning algorithms. All machine learning algorithms used in this work make use of both global and local protein features described in the sections 5.2.1 and 5.2.2. To study the impact of the MWCG, ESPH, and ASPH methods these features are tested separately in different machine learning algorithms. The parameters used to generate these machine learning features in this work are described in detail in sections 5.3 and 5.4 and below.

5.2.1 Global features

The global protein features described in this section were used in all the machine learning models in this work. The global features that were used in this work are R-value, resolution, and total number of heavy atoms. These features are obtained via the experimental data recorded in PDB file of each protein. Both R-value and resolution provide measures of the quality of the atomic model obtained from the X-ray crystallography. Also included as a global feature is the total protein size which is determined as the sum of heavy elements (carbon, nitrogen, oxygen, and sulfur) present in the protein. To code the protein size data, it is organized into one of 10 discrete size classes using one hot encoding. The size ranges are given based on the distribution of total number of heavy elements of each protein. For this work we use the following size classes. A frequency distribution of the size categories is provided in Figure 5.3.

[500, 750, 1000, 1500, 2000, 2500, 3000, 4000, 5000, 30000]

Using one-hot coding, a protein element feature size will take on 1 if the number of heavy atoms (carbon, nitrogen, or oxygen) of the protein is less than or equal to the corresponding size and zero for the other sizes. For example, a protein with 600 heavy elements would have the feature size vector for all of its atoms given by

[0, 1, 0, 0, 0, 0, 0, 0, 0, 0].

The maximum size bin is 30,000 since all proteins in the dataset have less than 30,000 heavy elements.

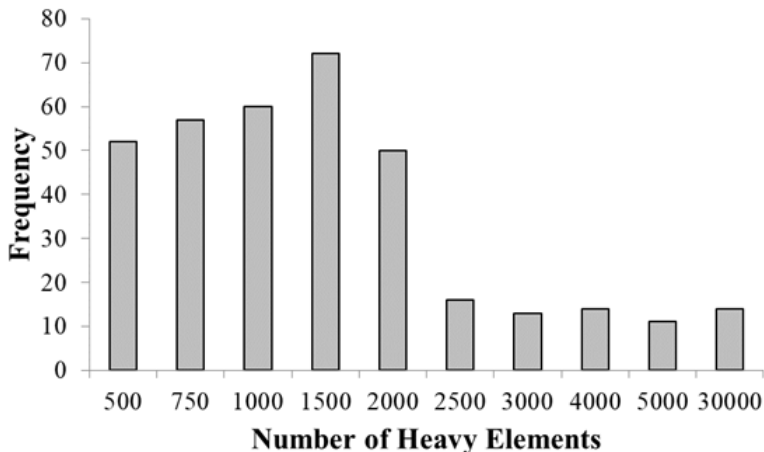


Figure 5.3: Frequency of the number of heavy elements from the 364 protein dataset. Figure originally published in Bramer *et al* [2].

5.2.2 Local features

In addition to the features discussed above, PDB files contain the amino acid corresponding to each heavy element. Like the protein size feature, amino acid information is included by using one hot encoding for each heavy element which results in twenty amino acid features. More locally, each of the the four different heavy element types carbon, nitrogen, oxygen, and sulfur for each element are one hot coded which results in another four features.

To explicitly take the density of nearby atoms into account, this work includes packing density as an additional model feature. Short, medium and long packing density features for each heavy atom are generated and included in all the machine learning models used in this work. Mathematically, the packing density of the i^{th} atom is defined as

$$p_i^d = \frac{N_d}{N},$$

where d is the given cutoff in angstroms, N_d is the number of atoms within the Euclidean distance of the cutoff to the i^{th} atom, and N the total number of heavy atoms of the protein.

Table 5.1 provides the packing density cutoffs used in this work.

Table 5.1: The packing density distance parameters (d Å) used for generating short medium, and long packing density machine learning features.

Short	Medium	Long
$d < 3$	$3 \leq d < 5$	$5 \leq d$

Secondary structures also play an important role in protein interactions. This work includes several secondary structural machine learning features for all the machine learning models used. Several software packages exist for the prediction of secondary protein structures. All secondary protein machine learning features used in this work were generated using the STRIDE software. This software returns secondary structure results that are in maximal agreement with X-ray crystallography data through the use of an optimized knowledge based algorithm. STRIDE takes 3D atomic coordinates in the form of protein PDB files as input and assigns each atom to a corresponding secondary structural group. STRIDE assigns each atom as belonging to a alpha helix, 3-10 helix, PI-helix, extended conformation, isolated bridge, turn, or a coil. Solvent accessible surface area, ϕ and ψ angle information are also generated by the software. This provides a total of 12 secondary structure features that are used in all the machine learning models in this work.

5.3 MWCG Features

The MWCG flexibility index described in Chapter 3 is used to create feature vectors for carbon, nitrogen, and oxygen interactions with each heavy element. To capture multiscale interactions 3 different kernel parameterizations are used for each interaction type. This provides a total of nine MWCG machine learning features for each heavy element. The kernel parameters used in this work are based off previous results. Specific parameters for

the kernels used here were originally published in Bramer *et al* and are provided in Table 5.2.[1]

Table 5.2: Correlation kernel parameters used to generate parameter-free MWCG machine learning features. Parameters based on previous results.[1]

Kernel Type	κ	η^n	ν
Lorentz ($n = 1$)	-	16	3
Lorentz ($n = 2$)	-	2	1
Exponential ($n = 3$)	1	31	-

5.3.1 Image-like MWCG Features

Convolutional neural networks make use of the large amount of data provided in images by applying a convolution operation. Due to the massive amount of trainable parameters, fully connected feed forward neural networks are computationally prohibitive for images. Convolutional operations greatly reduce the number of free parameters, thereby striking good balance between deep predictive power and computational cost. For this work MWCG images are generated for every heavy atom in the data set then used in a deep CNN model. Multiscale images are generated using both Lorentz and exponential radial basis functions for all heavy atoms in the data set. The generated images capture multiscale interactions by using a number of different parameterizations of κ , ν , and η in their kernels. To capture a large range of protein atomic interaction scales this work uses the following values are used for κ , ν , and η .

$$\eta = \{1, 2, 3, 4, 5, 10, 15, 20\}$$

$$\kappa, \nu = \{2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, 7, 8, 9, 10, 11\}.$$

Taken together as a matrix, each generated “image” results in three 2D MWCG images

of dimension (8, 30) for each heavy atom in the data set. For this work MWCG images are generated for all carbon, nitrogen, and oxygen interactions for each heavy atom. This results in a total of three channels for each image and a final image dimension of (8, 30, 3) for each atom used the MWCG deep CNN testing.

The image matrix is given by F_i^k in equation 5.1, where each atom $f_i^k(l, m, n)$ represents the flexibility index of the i^{th} atom, and k^{th} atom interaction (C, N, or O), $l = \eta$, $m = \{\kappa, \nu\}$, and n the type of radial basis function. Values of $n = 1$ and $n = 2$ correspond to exponential and Lorentz radial basis functions respectively.

$$F_i^k = \left[\begin{array}{cccccccc} f_i^k(1, 2, 1) & f_i^k(1, 2.5, 1) & \dots & f_i^k(1, 11, 1) & f_i^k(1, 2, 2) & f_i^k(1, 2.5, 2) & \dots & f_i^k(1, 11, 2) \\ f_i^k(2, 2, 1) & f_i^k(2, 2.5, 1) & \dots & f_i^k(2, 11, 1) & f_i^k(2, 2, 2) & f_i^k(2, 2.5, 2) & \dots & f_i^k(2, 11, 2) \\ \vdots & & & & & & & \vdots \\ f_i^k(15, 2, 1) & f_i^k(15, 2.5, 1) & \dots & f_i^k(15, 11, 1) & f_i^k(15, 2, 2) & f_i^k(15, 2.5, 2) & \dots & f_i^k(15, 11, 2) \\ \underbrace{f_i^k(20, 2, 1) \quad f_i^k(20, 2.5, 1) \quad \dots \quad f_i^k(20, 11, 1)}_{\kappa} & \underbrace{f_i^k(20, 2, 2) \quad f_i^k(20, 2.5, 2) \quad \dots \quad f_i^k(20, 11, 2)}_{\nu} \end{array} \right] \eta \quad (5.1)$$

5.4 ASPH & ESPH Features

A variety of element-specific and atom-specific persistent homology features, as described in Chapter 4, are generated as local machine learning features. The ASPH and ESPH features are generated in several ways by varying kernels (Lorentz and exponential), element-specific pairs (CC, CN, CO), and distance metrics (Wasserstein-0 and Wasserstein-1, Bottleneck-0 and Bottleneck-1). For this work, all persistent homology features were generated with a radial cutoff of 11Å.

The distances determined by Wasserstein and Bottleneck metrics are dependent on the boundary of the corresponding persistence diagrams. In other words any events from one

diagram that do not match an event on the other diagram can contribute to the final Wasserstein or Bottleneck distance by their distances from the boundary. Considering these effects, this work includes two additional persistence diagrams. The additional diagrams are constructed by rotating the y -axis is rotated clockwise by 30° or 60° , respectively. Figure 5.4 provides an example of these modifications. By introducing this modification, the Bottleneck and Wasserstein distances correspondingly allow the model to recognize elements that have a short persistence, or lifespan. As a final consideration, a feature is generated by reflecting the original persistence diagram about the diagonal axis. An example of this modification is provided in Figure 5.4. A list of kernels, kernel parameters, y -axis change, distance metric, and element-specific pairs used to generate features in machine learning models is provided in Table 5.3.

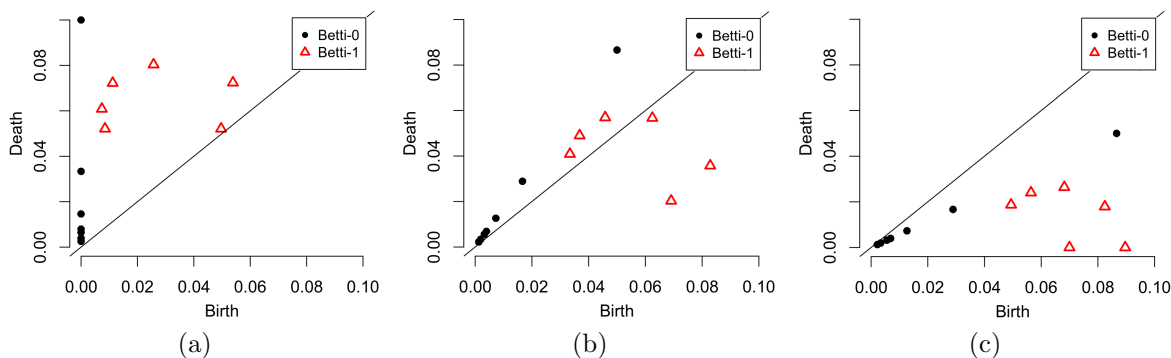


Figure 5.4: Illustration of modified persistence diagrams used in distance calculations. (a) Unchanged. (b) Rotated 30° . (c) rotated 60° . Black dots are Betti-0 events and triangles are Betti-1 events.

5.4.1 Image-like ASPH & ESPH Features

2D image-like persistent homology (PH) features for each C_α of the proteins are generated using the process described in Section 4.7. The images-like features are generated by taking various values of η and κ using the kernel function. An exponential kernel is used with a

Table 5.3: Parameters used for topological feature generation. All features used a cutoff of 11Å. Both lorentz (Lor) and exponential (exp) kernels and Bottleneck (B) and Wasserstein (W) distance metrics were used.

No. features	Kernel	Kernel parameter	Diagram	Distance metric	Element pair
12	Lor	$\eta = 21, \nu = 5$	Unchanged	B, W	CC, CN, CO
12	Exp	$\eta = 10, \kappa = 1$	Unchanged	B, W	CC, CN, CO
12	Exp	$\eta = 2, \kappa = 1$	Diagonal reflection	B, W	CC, CN, CO
12	Exp	$\eta = 2, \kappa = 1$	Rotated 30°	B, W	CC, CN, CO
12	Exp	$\eta = 2, \kappa = 1$	Rotated 60°	B, W	CC, CN, CO

radial cutoff of 11Å. Different values of η and κ are used to capture multiple interaction scales. The values used in this work are

$$\eta = \{1, 2, 3, 4, 5, 10, 15, 20\},$$

and

$$\kappa = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}.$$

This results in an image-like matrix given by PH_i^k in Eq. (5.2). Each atom $\text{PH}_i^k(l, m)$ represents the PH feature of the i^{th} C_α atom, and k^{th} atom interaction (C, N, or O), $l = \eta$, and $m = \kappa$.

$$\text{PH}_i^k = \underbrace{\left[\begin{array}{ccccc} f_i^k(1, 1) & f_i^k(1, 2) & \dots & f_i^k(1, 9) & f_i^k(1, 10) \\ f_i^k(2, 1) & f_i^k(2, 2) & \dots & f_i^k(2, 9) & f_i^k(2, 10) \\ \vdots & & \vdots & & \\ f_i^k(15, 1) & f_i^k(15, 2) & \dots & f_i^k(15, 9) & f_i^k(15, 10) \\ f_i^k(20, 1) & f_i^k(20, 2) & \dots & f_i^k(20, 9) & f_i^k(20, 10) \end{array} \right]}_{\kappa} \Bigg\} \eta \quad (5.2)$$

This generates 2D PH image-like features of dimension (8,10). Compared to MWCG images,

the PH images have lower resolution than the MWCG images due to the cost of calculating PH features. Images are generated for carbon, nitrogen, and oxygen element-specific interactions with each C_α atom. As a result, the final image feature input has a dimension of (8,10,3) for each C_α atom.

5.4.2 Cutoff Distance

For this work a cutoff of 11Å is used to generate all persistent homology machine learning features. The cutoff was determined using a basic grid search over various cutoff distances. Figure 5.5 displays the average Pearson correlation coefficient, obtained via fitting with experimental B factors, over the entire dataset using all persistent homology metrics with various point cloud distance cutoffs. The parameters listed in Table 5.4 are used to generate

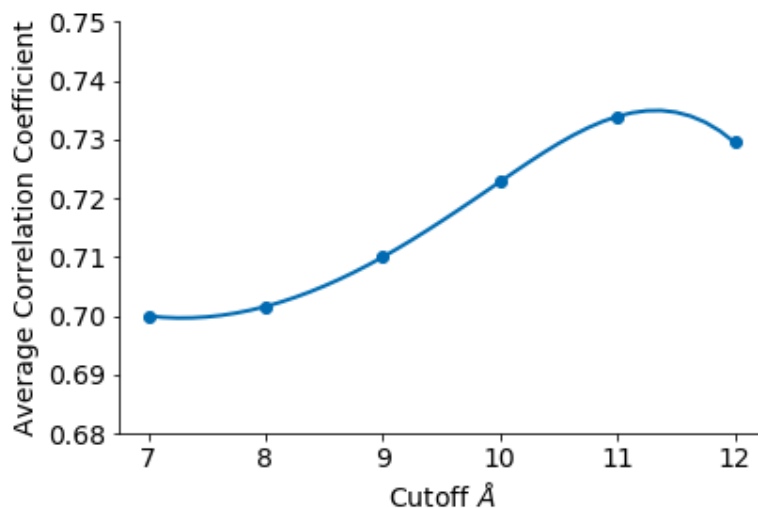


Figure 5.5: Average Pearson correlation coefficient over the entire protein dataset fitting all 24 persistent homology features using various cutoff distances.

PH features for each protein. These parameters were determined using a grid search over various ν , η , and κ .

Table 5.4: Parameters used for the element specific persistent homology features with a cutoff of 11 Å.

Kernel Type	ν	η^n	κ
Lorentz ($n = 1$)	5	21	-
Exponential ($n = 2$)	-	10	1

5.5 Machine Learning Model Parameters

For this work several machine learning models were generated. All machine learning models used in this study include the global and local features described sections 5.2.1 and 5.2.2. Two classes of machine models are generated for this work. The first includes random forest, gradient boosted tree, and deep convolutional neural networks that use MWCG input features in addition the general global and local features mentioned above. The second class of machine learning models use the ASPH and ESPH input features in addition to the general and local features. Each model has specific parameters than can be tuned. The following sections outline the parameters used in this work.

5.5.1 MWCG

5.5.1.1 Random Forest

Random forests only require the user to determine the amount of n trees. The predictive power of random forests generally increases with the number of trees used and these models are robust to over fitting. However increasing the number of trees comes at a computational cost. To balance performance with computational cost, this work uses $n = 500$ trees for all MWCG based random forest B factor prediction.

5.5.1.2 Gradient Boosted Trees

Several hyperparameters within the gradient boosted tree method can be tuned. The MWCG based GBT hyperparameters used in this work are determined using the standard practice of a grid search. Testing parameters are provided in Table 5.5. Any hyper parameters not listed below were taken to be the default values provided by the python scikit-learn package.

Table 5.5: Boosted gradient tree parameters used for testing MWCG based B factor prediction. These parameters were determined using a grid search. Any hyper parameters not listed below were taken to be the default values provided by the python scikit-learn package. MWCG based GBT machine learning prediction results originally published in Bramer *et al* [2].

Parameter	Setting
Loss Function	Quantile
Alpha	0.95
Estimators	1000
Learning Rate	0.001
Max Depth	4
Min Samples Leaf	9
Min Samples Split	9

5.5.1.3 Deep Convolutional Neural Network

This work uses 3 channel (8,30) MWCG based image-like correlation maps, as described in Section 5.3, as CNN input data for each image. The CNN output is flattened and concatenated with global and local protein features, as described in Sections 5.2.1 and 5.2.2, then input into a deep neural network to predict atomic B factor. A diagram of the MWCG based deep CNN architecture is provided in Figure 5.6.

The CNN input image used for MWCG based B factor in this work is a three-channel MWCG image of dimension (8,30,3). The deep CNN applies two convolutional layers with 2x2 filters, a dropout layer of 0.5, a dense layer, then flattens the resulting output. The

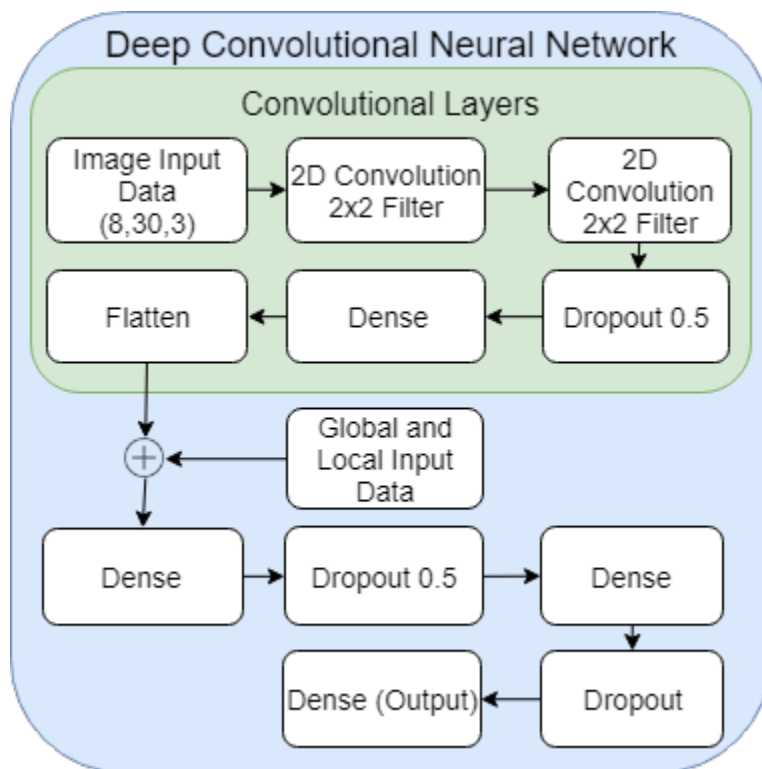


Figure 5.6: The MWCG based deep convolutional neural network architecture used for B factor prediction. The plus symbol represents the concatenation of data sets. Figure originally published in Bramer *et al* [2].

flattened output from the CNN is concatenated with the other global and local features into a dense layer of 59 neurons followed by a dropout layer of 0.5, another dense layer of 100 neurons, a dropout layer of 0.25, a dense layer of 10 neurons, and finishes with a dense output layer. This results in a total of 21,584 trainable parameters for the deep CNN used in MWCG based B factor prediction. A diagram of the deep CNN architecture is illustrated in Figure 5.6.

Convolutional neural networks have several hyper-parameters. The hyper parameters for the MWCG based deep CNN used in this work are optimized using a grid search. Table 5.6 provides a list of the hyper-parameter values used for testing. Any hyper parameters not listed below were taken to be the default values provided by the python Keras package.

Table 5.6: MWCG based deep Convolutional Neural Network (CNN) hyper-parameters used for testing. These hyper-parameters were determined using a grid search. Any hyper parameters not listed below were taken to be the default values provided by python with the Keras package. MWCG machine learning prediction results originally published in Bramer *et al* [2].

Parameter	Setting
Learning Rate	0.001
Epoch	100
Batch Size	100
Loss	Mean Absolute Error
Optimizer	Adam

5.5.2 ASPH & ESPH

The generated ASPH & ESPH features described in section 4.7 are used for prediction of protein B factor using both least squares fitting and machine learning as described in the following sections.

5.5.2.1 Gradient Boosted Trees

The persistent homology based GBT hyper-parameters used in this work are optimized using a grid search. The parameters used for testing are provided in 5.7. Any hyper-parameters not listed in the table were taken to be the default values provided by the python scikit-learn package.

Table 5.7: Boosted gradient tree parameters used for persistent homology based prediction testing. Parameters were determined using a grid search. Any hyper parameters not listed below were taken to be the default values provided by the python scikit-learn package.

Parameter	Setting
Loss Function	Quantile
Alpha	0.975
Estimators	500
Learning Rate	0.25
Max Depth	4
Min Samples Leaf	9
Min Samples Split	9

5.5.2.2 Deep Convolutional Neural Network

The deep CNN used in this work uses input images generated from an image-like correlation map. These images are generated by using a range of kernel parameters for atom-specific and element-specific persistent homology as described in Section 5.4.1. The CNN output is flattened and then input into a DNN along with global and local protein features. This allows the deep CNN to use the same feature set as the boosted gradient method to be used as well as the generated PH image-like data. Figure 5.7 provides a diagram of the CNN architecture used for the PH based B factor prediction in this work.

The CNN is passed a three-channel persistent homology image of dimension (8,10,3) for each C_α of the training set. The model used in this work takes the input image data and

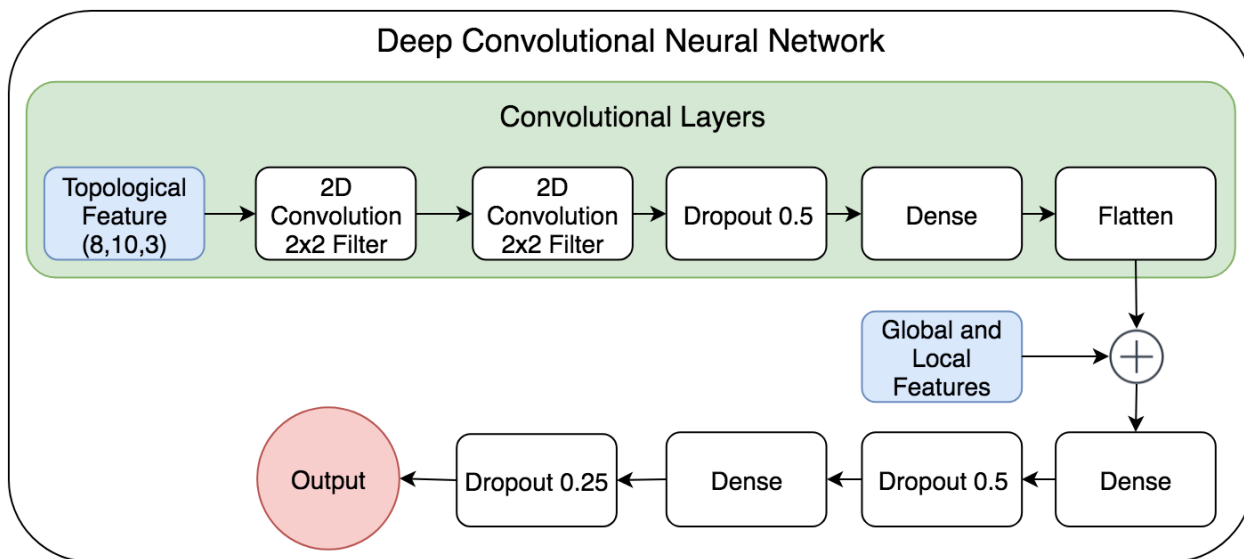


Figure 5.7: The deep learning architecture using a convolutional neural network combined with a deep neural network to predict B factor using PH based features. The plus symbol represents the concatenation of features.

applies two convolutional layers with 2x2 filters, followed by a dropout of 0.5. The image data is then passed through a dense layer, flattened, then joined with the other global and local features to form a dense layer of 218 neurons. This is followed by a dropout layer of 0.5, another dense layer of 100 neurons, a dropout layer of 0.25, a dense layer of 10 neurons, and finishes with a dense layer of the B factor prediction output. Figure 5.7 provides an illustration of the deep CNN used in this work.

Several hyper-parameters of the deep convolutional neural network can be tuned. The deep convolutional neural network hyper-parameters are optimized using a basic grid search. Table 5.8 provides the parameters used for testing. Any hyper-parameters not listed in the provided table were taken to be the default values provided by the python Keras package.

Table 5.8: Convolutional Neural Network (CNN) parameters used for testing persistent homology based features. Parameters were determined using a grid search. Any hyper-parameters not listed below were taken to be the default values provided by python with the Keras package.

Parameter	Setting
Learning Rate	0.001
Epoch	1000
Batch Size	1000
Loss	Mean Squared Error
Optimizer	Adam

5.6 Machine Learning Datasets

The image like features used in all convolutional neural networks were standardized with mean 0 and variance of 1. Because the STRIDE software is unable to provide features for these proteins, 1OB4, 1OB7, 2OLX, and 3MD5 are excluded from the data set. Protein 1AGN is also excluded due to known problems with this protein data. Lastly, proteins 1NKO, 2OCT, and 3FVA are excluded because they have residues with B factors reported as zero, which is unphysical.

5.6.1 Training set and test set

The PH and MWCG based machine learning algorithms used in this work are all trained and tested using a leave-one-protein-out approach. For each protein a machine learning model is built using the entire dataset but excluding data from the protein whose B factors are to be predicted. The dataset contains over 620,000 atoms in total which provides a training set of roughly 600,000 data points (i.e., atoms) for each protein. Each heavy atom in the training set has an associated set of input features, as described in Sections 5.3 and 4.7, and a B factor output. The feature inputs and the outputs in the training set are used to train each machine learning model. Since the predictions are leave-one-protein-out, data from

each protein is taken as a test set when its B factors are to be blindly predicted.

All random forest models and boosted gradient models are implemented using the scikit-learn python package. All deep CNN models are implemented using the python package Keras with tensorflow as a backend.

Chapter 6

Workflow

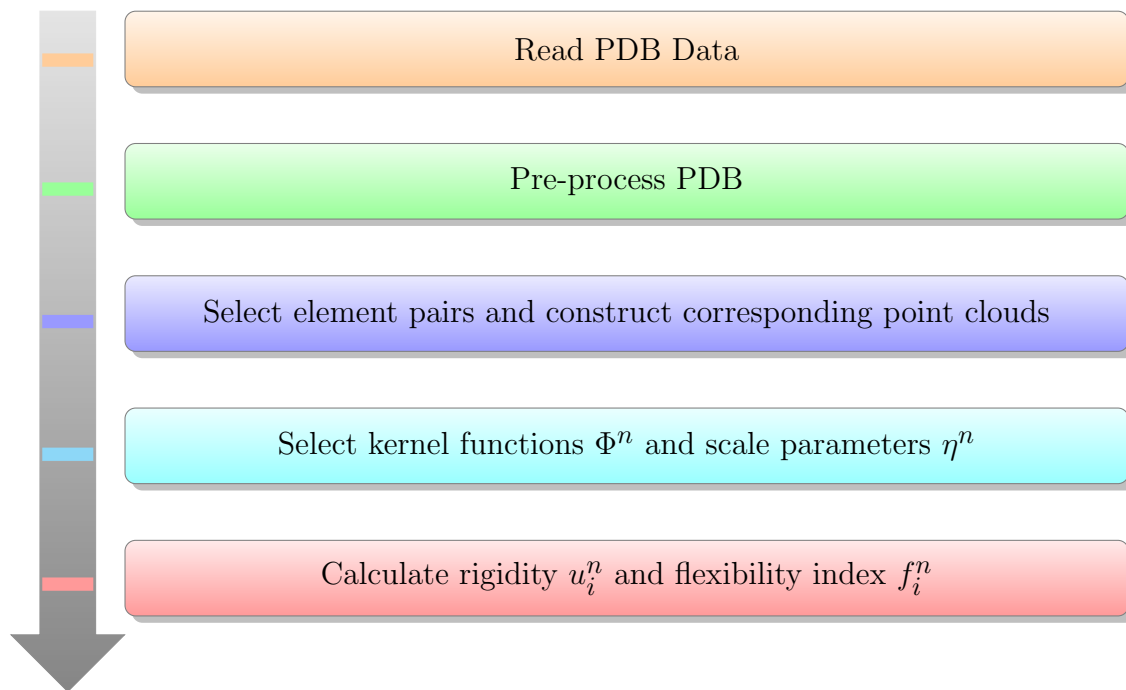


Figure 6.1: Workflow for procedure in MWCG feature construction.

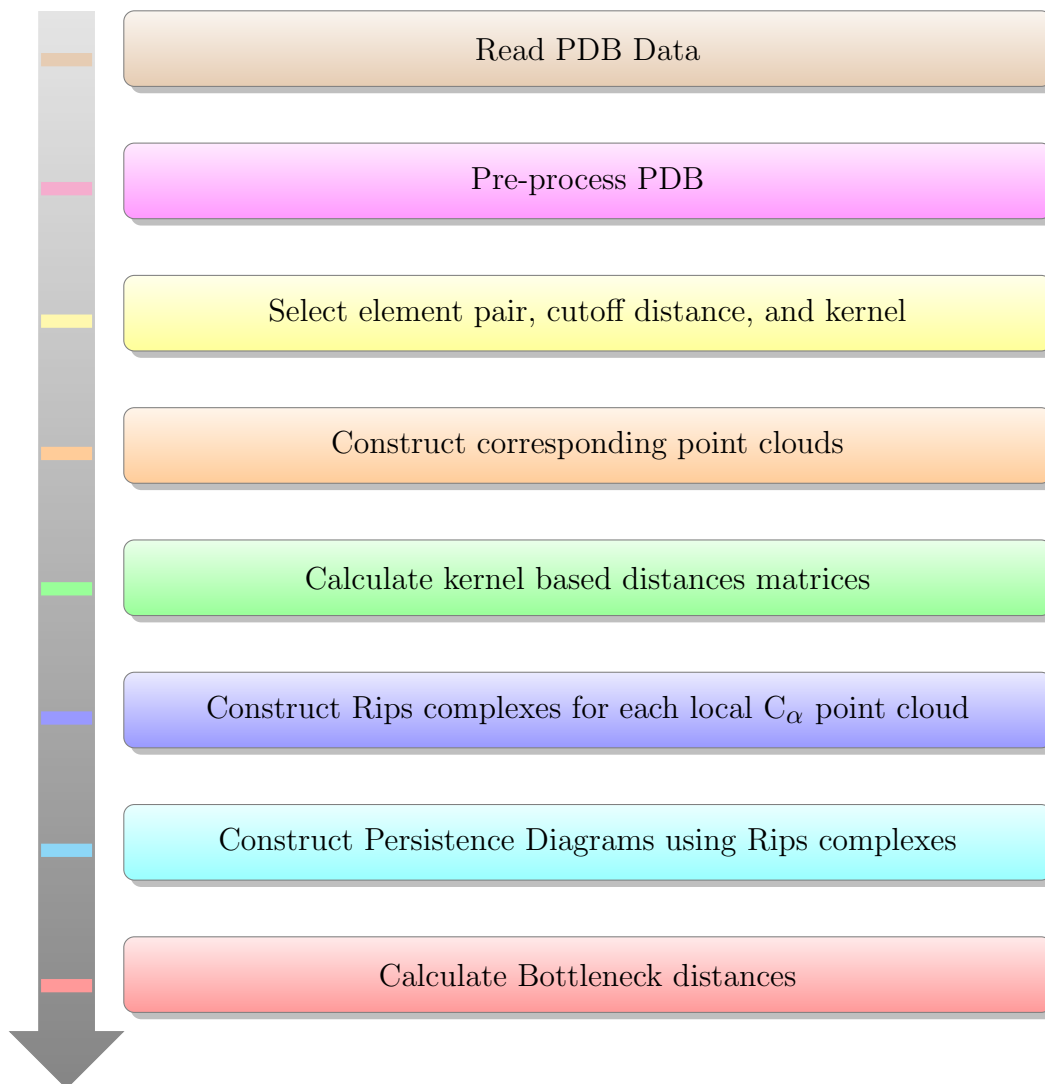


Figure 6.2: Workflow for procedure in ASPH and ESPH feature construction.

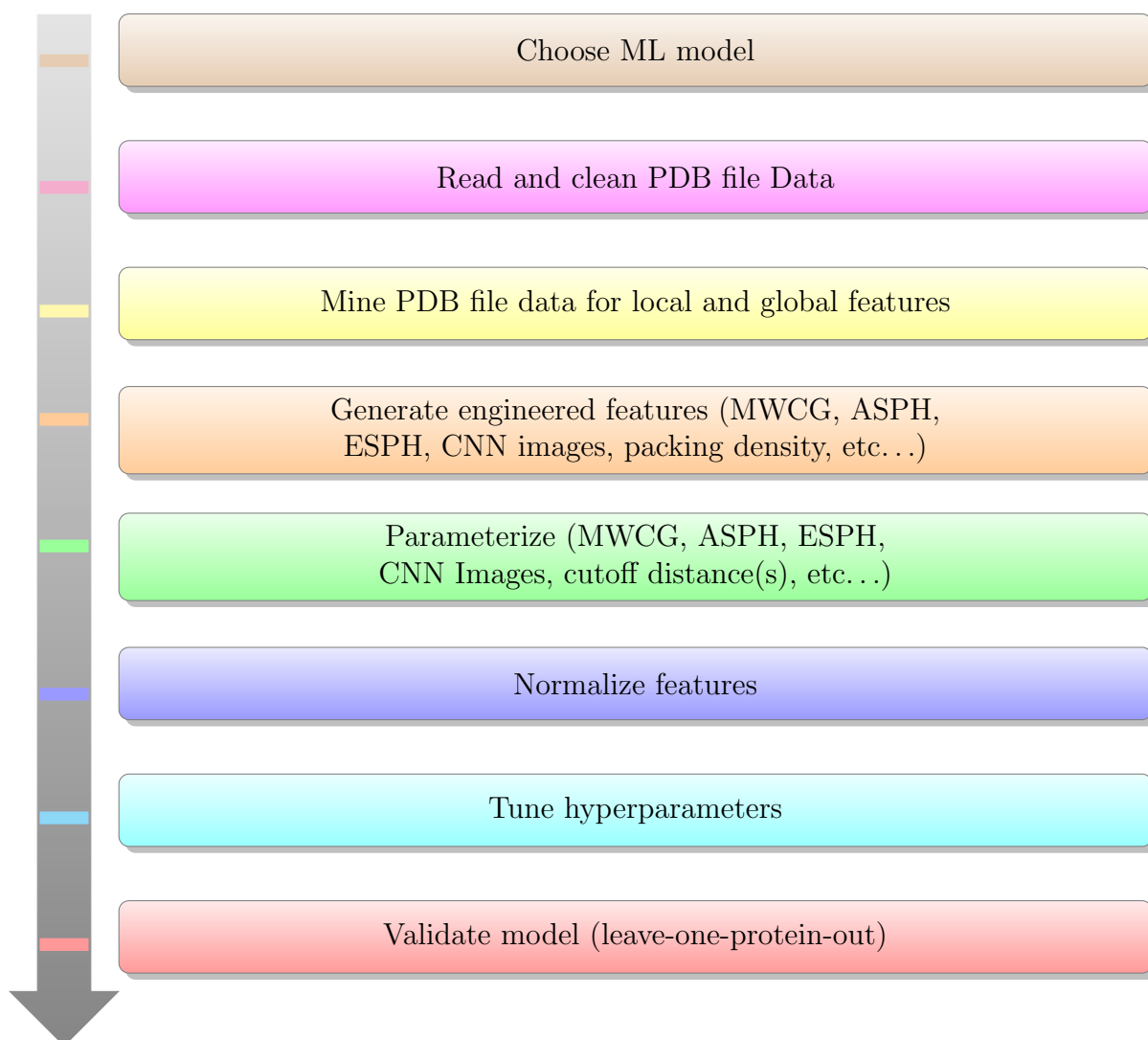


Figure 6.3: Workflow for procedure MWCG, ASPH, and ESPH based machine learning B factor prediction.

Chapter 7

Results

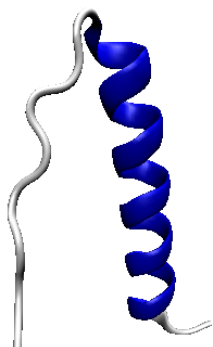
7.1 Visualization of Element Specific Correlation Maps

In this result the radial basis functions are used in the MWCG method to construct various element specific correlation heat maps of a given protein. For this study we consider carbon, nitrogen, and oxygen interactions and create correlation heat maps using both nitrogen-nitrogen and carbon-carbon interaction pairs. Only one spatial scale is used to illustrate the element specific feature of the MWCG method. This is abbreviated as WCG in the related tables. Given an element pair, each map was calculated used the average of the three kernels described in Chapter 3. Axes of each correlation map correspond to individual atoms of each carbon, nitrogen, or oxygen atom in the given protein. In this work correlation heat maps are generated using the three proteins with PDB ID 3TYS, 1AIE, and 3PSM. Nitrogen-nitrogen and oxygen-oxygen correlation heat maps are provided in Figures 7.1, 7.2, and 7.3. Each figure also includes a 3D representation, generated using Visual Molecular Dynamics (VMD) software, of each protein for reference.

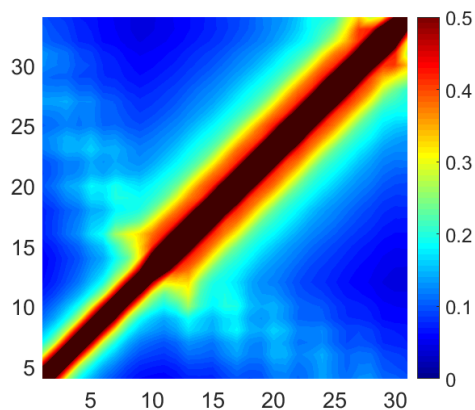
7.2 Hinge Detection

Accurate and robust identification of hinge regions is an ongoing problem. An important application of hinge region detection is domain identification. Hinge regions of proteins also

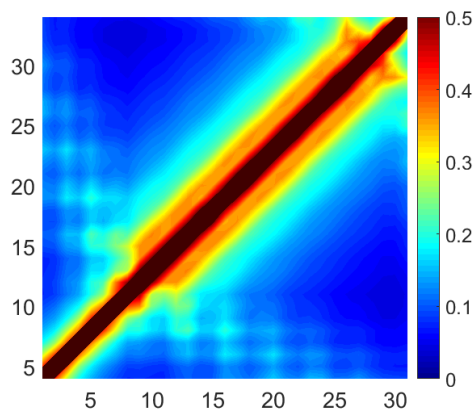
play an essential role in enzymatic catalysis due to their ability to allow conformational changes to the protein. Binding by ligands can be accommodated by a flexible active site as seen in hinge regions. With these considerations in mind, hinge prediction cannot be overlooked when developing methods for protein flexibility and dynamics analysis. The MWCG presented here can be used as a hinge detection tool. In this work we consider three interesting examples. Calmodulin provides an example of a protein hinge that effects both the structure and function of the protein. For this result experimental protein B factors of C_α atoms are compared with predictions from the WCG method and GNM for calmodulin (PDB ID 1CLL), ribosomal protein (PDB ID 1WHI), and engineered fluorescent cyan protein (PDB ID 2HQK). To highlight the value of the element specific feature of the MWCG only one scale is used so that the method is simply WCG. For comparison protein PDB ID 1CLL includes MWCG and WCG predictions to compare and contrast the element specific and multiscale nature of the MWCG method. Results are generated with carbon-carbon, carbon-nitrogen, and carbon-oxygen interaction pairs. Exponential type kernels are used with fixed parameters $\kappa = 1$, and $\eta = 3 \text{ \AA}$. The results are displayed in Figures 7.5, 7.4, and 7.6.



(a) 1AIE

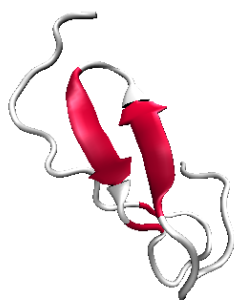


(b) Amine Nitrogens

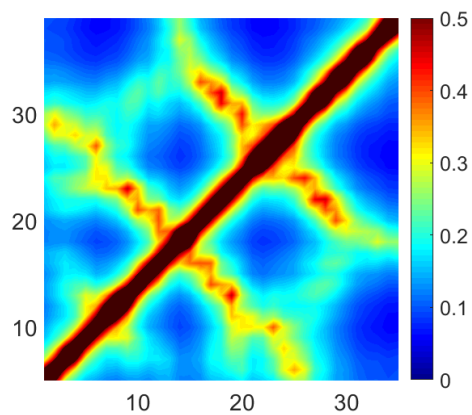


(c) Double Bonded Carboxyl Oxygens

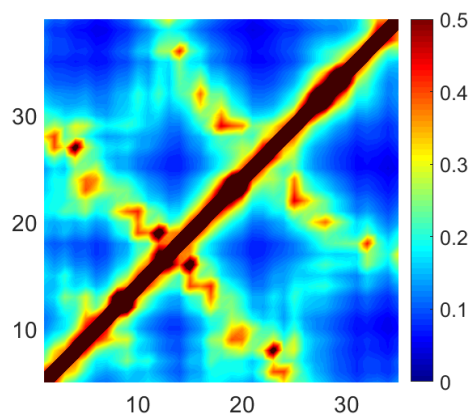
Figure 7.1: (a) VMD representation of PBD ID 1AIE. (b) Correlation maps for nitrogen-nitrogen (NN) and (c) oxygen-oxygen (OO) interactions for protein 1AIE. The thicker band along the main diagonal of (b) and (c) corresponds to the alpha helix secondary structure in 1AIE. Figure originally published in Bramer *et al* [1].



(a) 1KGM

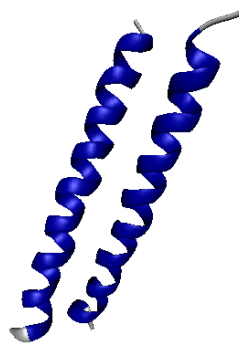


(b) Amine Nitrogens

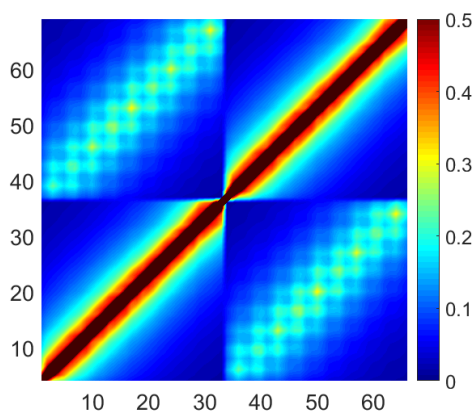


(c) Double Bonded Carboxyl Oxygens

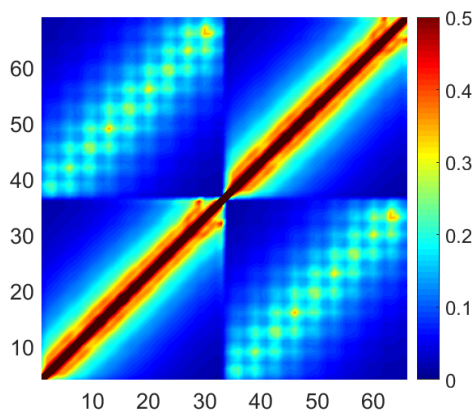
Figure 7.2: (a) VMD representation of PBD ID 1KGM. (b) Correlation maps for nitrogen-nitrogen (NN) and (c) oxygen-oxygen (OO) interactions for protein 1KGM. The bands perpendicular to the main diagonal of (b) and (c) correspond to the anti parallel beta sheet present in 1KGM. Figure originally published in Bramer *et al* [1].



(a) 5IIV

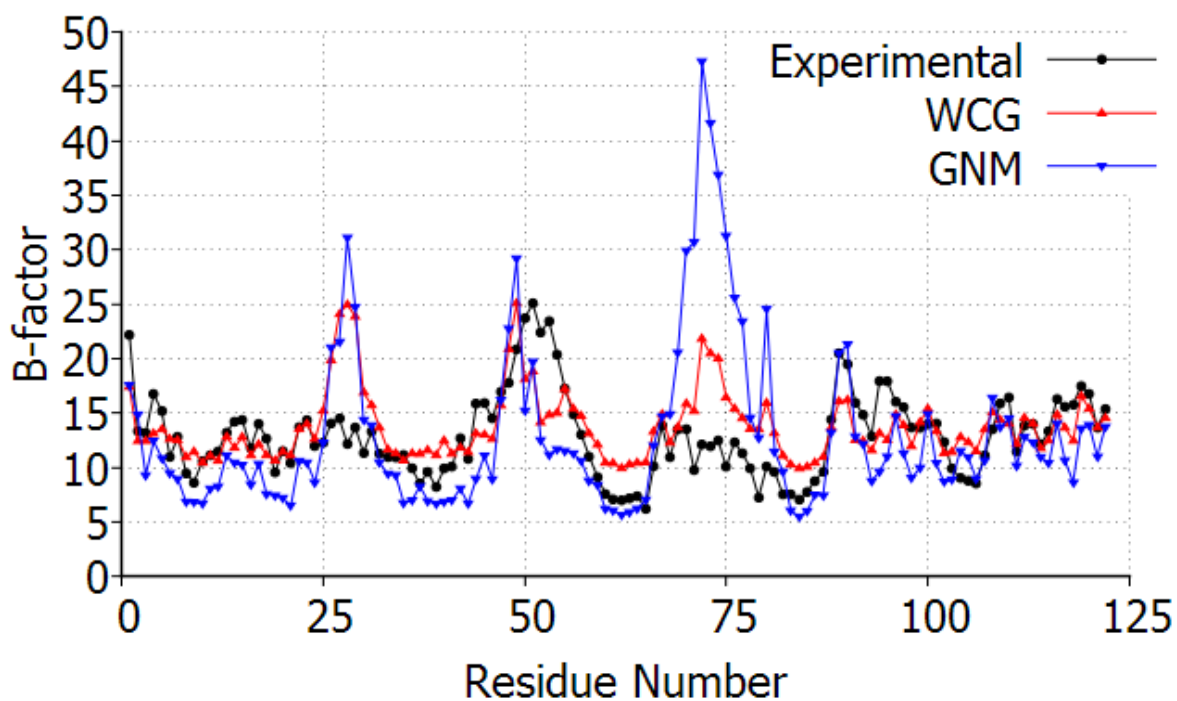
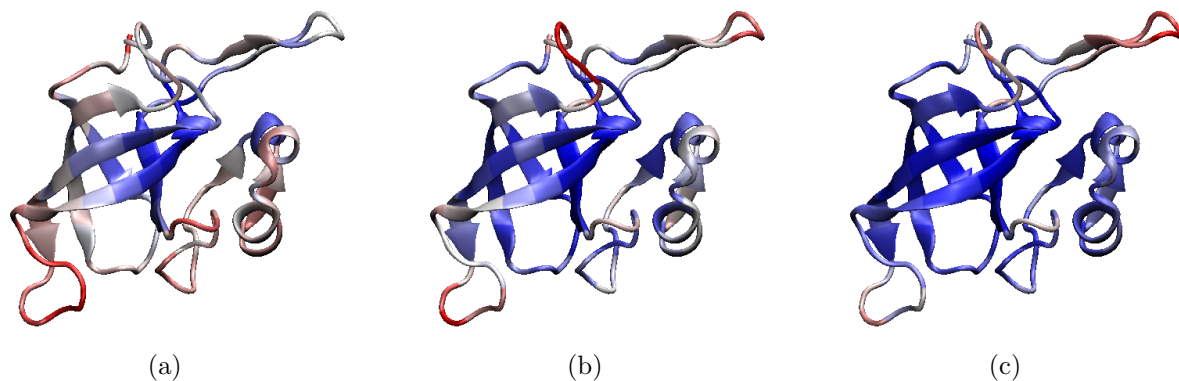


(b) Amine Nitrogen



(c) Double Bonded Carboxyl Oxygens

Figure 7.3: (a) VMD representation of PBD ID 5IIV. (b) Correlation maps for nitrogen-nitrogen (NN) and (c) oxygen-oxygen (OO) interactions for protein 5IIV. The presence of the two distinct thick bands along the main diagonal of (b) and (c) corresponds to the two alpha helices present in 5IIV. The off diagonal bands correspond to the bonding interaction between alpha helices. Figure originally published in Bramer *et al* [1].



(d)

Figure 7.4: (a) A visual comparison of experimental B factors , (b) WCG predicted B factors, (c) and GNM predicted B factors for the ribosomal protein L14 (PDB ID:1WHI). (d) The experimental and predicted B factor values plotted per residue. GNM represents predicted B factors using GNM with a cutoff distance of 7 Å. WCG is parametrized using CC, CN, CO kernels of the exponential type with fixed parameters $\kappa = 1$, and $\eta = 3$ Å. Figure originally published in Bramer *et al* [1].

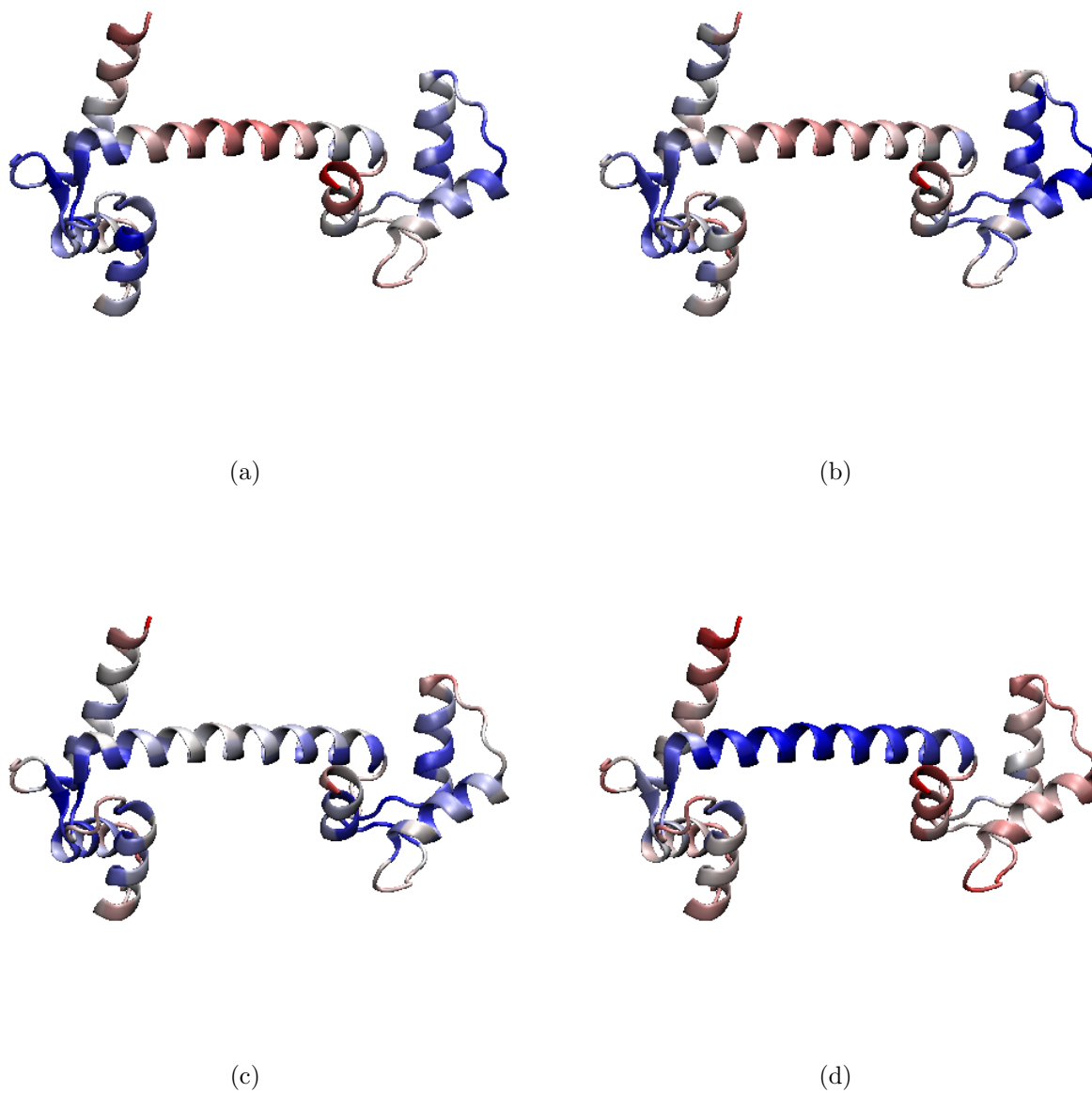
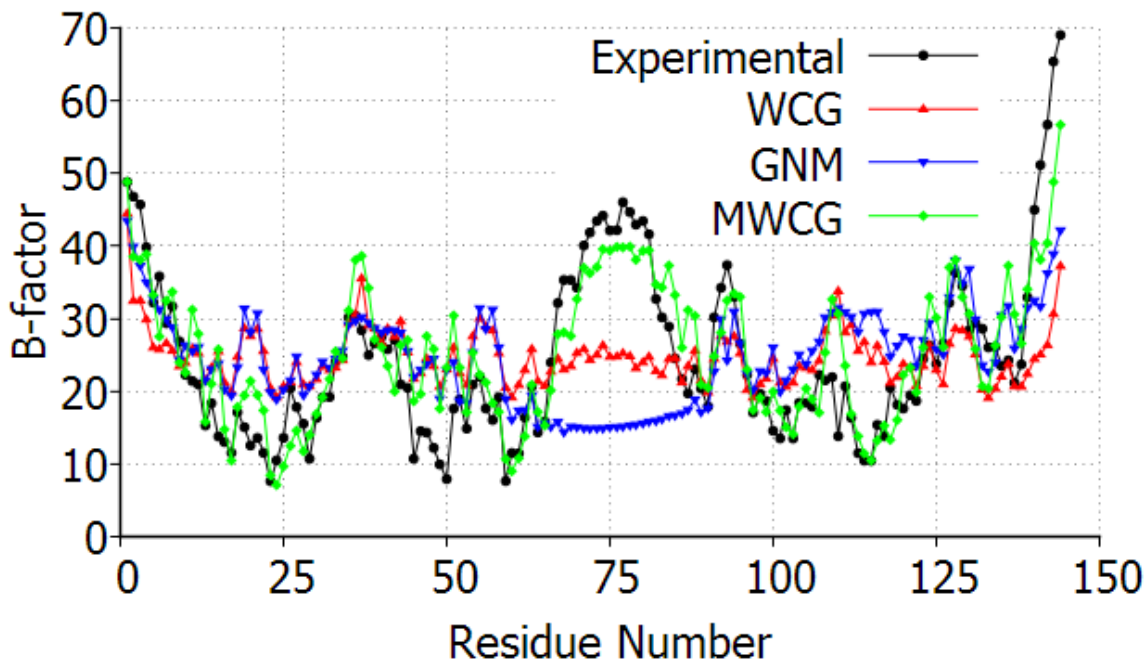


Figure 7.5: (a) The structure of calmodulin (PDB ID: 1CLL) visualized in Visual Molecular Dynamics (VMD)18 and colored by experimental B factors, (b) MWCG predicted B factors, (c) WCG predicted B factors, (d) and GNM predicted B factors with red representing the most flexible regions. Figure originally published in Bramer *et al* [1].



(e)

Figure 7.5: (Continued) (e) The experimental (Exp) and predicted B factor values plotted per residue for PDB ID 1CLL. The GNM is for the GNM method with a cutoff distance of 7 Å. We see that GNM clearly misses the flexible hinge region. WCG is parametrized using CC, CN, CO kernels of the exponential type with fixed parameters $\kappa = 1$, and $\eta = 3$ Å. MWCG represents B factor predictions determined from the MWCG method using the fixed parameters listed in Table 3.2. Figure originally published in Bramer *et al* [1].

7.3 MWCG

7.3.1 Validation

The Pearson correlation coefficient is used to quantitatively assess the prediction results.

The Pearson correlation coefficient for B factor prediction used in this work is given by

$$\text{PCC} = \frac{\sum_{i=1}^N (B_i^e - \bar{B}^e)(B_i^t - \bar{B}^t)}{\left[\sum_{i=1}^N (B_i^e - \bar{B}^e)^2 \sum_{i=1}^N (B_i^t - \bar{B}^t)^2 \right]^{1/2}}, \quad (7.1)$$

where $B_i^t, i = 1, 2, \dots, N$ are predicted B factors using the proposed method and $B_i^e, i = 1, 2, \dots, N$ are experimental B factors from the PDB file. The terms B_i^t and B_i^e represent the i^{th} theoretical and experimental B factors respectively. Here \bar{B}^e and \bar{B}^t are averaged B factors.

7.3.2 Fitting Results

Tables 7.1-7.6, and 7.4 provide the average Pearson correlation coefficient obtained using the MWCG method as outlined in Chapter 3. The MWCG method is compared to other commonly used protein B factor prediction methods. The MWCG B factor Pearson correlation coefficient results for all 364 proteins in the dataset are provided in table 7.4. The proposed MWCG method, optimal FRI (opFRI), parameter free FRI (pfFRI), and GNM methods are all compared. The same comparison for proteins of relatively, small, medium, and large sizes are provided in tables 7.1, 7.2, and 7.3.

Table 7.4: Correlation coefficients for B factor prediction obtained by MWCG, optimal FRI (opFRI), parameter free FRI (pfFRI), and Gaussian normal mode (GNM) for a set of 364 proteins. GNM scores reported here are the result of tests with a processed set of PDB files as described in Chapter 2.2. MWCG results originally published in Bramer *et al* [1].

PDB ID	N	MWCG	opFRI	pfFRI	GNM	PDB ID	N	MWCG	opFRI	pfFRI	GNM
1ABA	87	0.855	0.727	0.698	0.613	1PEF	18	0.989	0.888	0.826	0.808
1AHO	64	0.768	0.698	0.625	0.562	1PEN	16	0.957	0.516	0.465	0.27
1AIE	31	0.969	0.588	0.416	0.155	1PMY	123	0.701	0.671	0.654	0.685
1AKG	16	0.945	0.373	0.35	0.185	1PZ4	114	0.921	0.828	0.781	0.843
1ATG	231	0.843	0.613	0.578	0.497	1Q9B	43	0.957	0.746	0.726	0.656
1BGF	124	0.834	0.603	0.539	0.543	1QAU	112	0.786	0.678	0.672	0.62

Table 7.4 (cont'd)

PDB ID	N	MWCG	opFRI	pfFRI	GNM	PDB ID	N	MWCG	opFRI	pfFRI	GNM
1BX7	51	0.896	0.726	0.623	0.706	1QKI	3912	0.508	0.809	0.751	0.645
1BYI	224	0.600	0.543	0.491	0.552	1QTO	122	0.809	0.543	0.52	0.334
1CCR	111	0.741	0.58	0.512	0.351	1R29	122	0.787	0.65	0.631	0.556
1CYO	88	0.860	0.751	0.702	0.741	1R7J	90	0.859	0.789	0.621	0.368
1DF4	57	0.941	0.912	0.889	0.832	1RJU	36	0.805	0.517	0.447	0.431
1E5K	188	0.848	0.746	0.732	0.859	1RRO	112	0.748	0.435	0.372	0.529
1ES5	260	0.700	0.653	0.638	0.677	1SAU	114	0.819	0.742	0.671	0.596
1ETL	12	0.932	0.71	0.609	0.628	1TGR	104	0.810	0.72	0.711	0.714
1ETM	12	0.941	0.544	0.393	0.432	1TZV	141	0.869	0.837	0.82	0.841
1ETN	12	0.949	0.089	0.023	-0.274	1U06	55	0.774	0.474	0.429	0.434
1EW4	106	0.804	0.65	0.644	0.547	1U7I	267	0.885	0.778	0.762	0.691
1F8R	1932	0.504	0.878	0.859	0.738	1U9C	221	0.764	0.6	0.577	0.522
1FF4	65	0.933	0.718	0.613	0.674	1UHA	83	0.838	0.726	0.665	0.638
1FK5	93	0.648	0.59	0.568	0.485	1UKU	102	0.765	0.665	0.661	0.742
1GCO	1044	0.839	0.766	0.693	0.646	1ULR	87	0.718	0.639	0.594	0.495
1GK7	39	0.984	0.845	0.773	0.821	1UOY	64	0.769	0.713	0.653	0.671
1GVD	52	0.849	0.781	0.732	0.591	1USE	40	0.960	0.438	0.146	-0.142
1GXU	88	0.901	0.748	0.634	0.421	1USM	77	0.819	0.832	0.809	0.798
1H6V	2927	0.133	0.488	0.429	0.306	1UTG	70	0.745	0.691	0.61	0.538
1HJE	13	0.931	0.811	0.686	0.616	1V05	96	0.841	0.629	0.599	0.632
1I71	83	0.798	0.549	0.516	0.549	1V70	105	0.854	0.622	0.492	0.162
1IDP	441	0.827	0.735	0.715	0.69	1VRZ	21	0.995	0.792	0.695	0.677
1IFR	113	0.875	0.697	0.689	0.637	1W2L	97	0.747	0.691	0.564	0.397
1K8U	89	0.856	0.553	0.531	0.378	1WBE	204	0.767	0.591	0.577	0.549
1KMM	1499	0.740	0.749	0.744	0.558	1WHI	122	0.804	0.601	0.539	0.27
1KNG	144	0.810	0.547	0.536	0.512	1WLY	322	0.728	0.695	0.679	0.666
1KR4	110	0.892	0.635	0.612	0.466	1WPA	107	0.797	0.634	0.577	0.417
1KYC	15	0.971	0.796	0.763	0.754	1X3O	80	0.787	0.6	0.559	0.654
1LR7	73	0.929	0.679	0.657	0.62	1XY1	18	0.933	0.832	0.645	0.447
1MF7	194	0.757	0.687	0.681	0.7	1XY2	8	1.000	0.619	0.57	0.562
1N7E	95	0.812	0.651	0.609	0.497	1Y6X	87	0.838	0.596	0.524	0.366
1NKD	59	0.911	0.75	0.703	0.631	1YJO	6	1.000	0.375	0.333	0.434
1NKO	122	0.831	0.619	0.535	0.368	1YZM	46	0.970	0.842	0.834	0.901
1NLS	238	0.799	0.669	0.53	0.523	1Z21	96	0.725	0.662	0.638	0.433
1NNX	93	0.834	0.795	0.789	0.631	1ZCE	146	0.898	0.808	0.757	0.77
1NOA	113	0.808	0.622	0.604	0.615	1ZVA	75	0.911	0.756	0.579	0.69
1NOT	13	0.937	0.746	0.622	0.523	2A50	457	0.704	0.564	0.524	0.281
1O06	20	0.988	0.91	0.874	0.844	2AGK	233	0.821	0.705	0.694	0.512
1O08	221	0.516	0.562	0.333	0.309	2AH1	939	0.462	0.684	0.593	0.521
1OB4	16	1.000	0.776	0.763	0.75	2B0A	186	0.805	0.639	0.603	0.467
1OB7	16	1.000	0.737	0.545	0.652	2BCM	413	0.695	0.555	0.551	0.477
1OPD	85	0.607	0.555	0.409	0.398	2BF9	36	0.714	0.606	0.554	0.68

Table 7.4 (cont'd)

PDB ID	N	MWCG	opFRI	pfFRI	GNM	PDB ID	N	MWCG	opFRI	pfFRI	GNM
1P9I	29	0.841	0.754	0.742	0.625	2BRF	100	0.873	0.795	0.764	0.71
2CE0	99	0.824	0.706	0.598	0.529	2C71	205	0.773	0.658	0.649	0.56
2CG7	90	0.738	0.551	0.539	0.379	2OLX	4	1.000	0.917	0.888	0.885
2COV	534	0.895	0.846	0.823	0.812	2PKT	93	0.762	0.162	0.003	0.193
2CWS	227	0.756	0.647	0.64	0.696	2PLT	99	0.635	0.508	0.484	0.509
2D5W	1214	0.448	0.689	0.682	0.681	2PMR	76	0.799	0.693	0.682	0.619
2DKO	253	0.873	0.816	0.812	0.69	2POF	440	0.743	0.682	0.651	0.589
2DPL	565	0.721	0.596	0.538	0.658	2PPN	107	0.673	0.677	0.638	0.668
2DSX	52	0.704	0.337	0.333	0.127	2PSF	608	0.641	0.526	0.5	0.565
2E10	439	0.808	0.798	0.796	0.692	2PTH	193	0.901	0.822	0.784	0.767
2E3H	81	0.794	0.692	0.682	0.605	2Q4N	153	0.846	0.711	0.667	0.74
2EAQ	89	0.817	0.753	0.69	0.695	2Q52	412	0.510	0.756	0.748	0.621
2EHP	248	0.832	0.804	0.804	0.773	2QJL	99	0.611	0.594	0.584	0.594
2EHS	75	0.805	0.72	0.713	0.747	2R16	176	0.640	0.582	0.495	0.618
2ERW	53	0.513	0.461	0.253	0.199	2R6Q	138	0.915	0.603	0.54	0.529
2ETX	389	0.854	0.58	0.556	0.632	2RB8	93	0.840	0.727	0.614	0.517
2FB6	116	0.850	0.791	0.786	0.74	2RE2	238	0.711	0.652	0.613	0.673
2FG1	157	0.719	0.62	0.617	0.584	2RFR	154	0.826	0.693	0.671	0.753
2FN9	560	0.704	0.607	0.595	0.611	2V9V	135	0.697	0.555	0.548	0.528
2FQ3	85	0.844	0.719	0.692	0.348	2VE8	515	0.698	0.744	0.643	0.616
2G69	99	0.850	0.622	0.59	0.436	2VH7	94	0.851	0.775	0.726	0.596
2G7O	68	0.888	0.785	0.784	0.66	2VIM	104	0.859	0.413	0.393	0.212
2G7S	190	0.756	0.67	0.644	0.649	2VPA	204	0.757	0.763	0.755	0.576
2GKG	122	0.748	0.688	0.646	0.711	2VQ4	106	0.776	0.68	0.679	0.555
2GOM	121	0.874	0.586	0.584	0.491	2VY8	149	0.759	0.77	0.724	0.533
2GXG	140	0.901	0.847	0.78	0.52	2VYO	210	0.777	0.675	0.648	0.729
2GZQ	191	0.462	0.505	0.382	0.369	2W1V	548	0.761	0.68	0.68	0.571
2HQB	213	0.897	0.824	0.809	0.365	2W2A	350	0.819	0.706	0.638	0.589
2HYK	238	0.728	0.585	0.575	0.51	2W6A	117	0.804	0.823	0.748	0.647
2I24	113	0.672	0.593	0.498	0.494	2WJ5	96	0.821	0.484	0.44	0.357
2I49	398	0.766	0.714	0.683	0.601	2WUJ	100	0.919	0.739	0.598	0.598
2IBL	108	0.919	0.629	0.625	0.352	2WW7	150	0.629	0.499	0.471	0.356
2IGD	61	0.865	0.585	0.481	0.386	2WWE	111	0.903	0.692	0.582	0.628
2IMF	203	0.798	0.652	0.625	0.514	2X1Q	240	0.505	0.534	0.478	0.443
2IP6	87	0.841	0.654	0.578	0.572	2X25	168	0.710	0.632	0.598	0.403
2IVY	88	0.837	0.544	0.483	0.271	2X3M	166	0.875	0.744	0.717	0.655
2J32	244	0.878	0.863	0.848	0.855	2X5Y	171	0.799	0.718	0.705	0.694
2J9W	200	0.741	0.716	0.705	0.662	2X9Z	262	0.726	0.583	0.578	0.574
2JKU	35	0.926	0.805	0.695	0.656	2XHF	310	0.830	0.606	0.591	0.569
2JLI	100	0.937	0.779	0.613	0.622	2Y0T	101	0.834	0.778	0.774	0.798
2JLJ	115	0.811	0.741	0.72	0.527	2Y72	170	0.926	0.78	0.754	0.766
2MCM	113	0.867	0.789	0.713	0.639	2Y7L	319	0.939	0.928	0.797	0.747

Table 7.4 (cont'd)

PDB ID	N	MWCG	opFRI	pfFRI	GNM	PDB ID	N	MWCG	opFRI	pfFRI	GNM
2NLS	36	0.937	0.605	0.559	0.53	2Y9F	149	0.769	0.771	0.762	0.664
2NR7	194	0.885	0.803	0.785	0.727	2YLB	400	0.820	0.807	0.807	0.675
2NUH	104	0.922	0.835	0.691	0.771	2YNY	315	0.836	0.813	0.804	0.706
2O6X	306	0.825	0.814	0.799	0.651	2ZCM	357	0.723	0.458	0.422	0.42
2OA2	132	0.703	0.571	0.456	0.458	2ZU1	360	0.753	0.689	0.672	0.653
2OCT	192	0.673	0.567	0.55	0.54	3A0M	148	0.916	0.807	0.712	0.392
2OHW	256	0.743	0.614	0.539	0.475	3A7L	128	0.806	0.713	0.663	0.756
2OKT	342	0.779	0.433	0.411	0.336	3AMC	614	0.758	0.675	0.669	0.581
2OL9	6	1.000	0.909	0.904	0.689	3AUB	116	0.650	0.614	0.608	0.637
3BA1	312	0.827	0.661	0.624	0.621	3B5O	230	0.729	0.644	0.629	0.601
3BED	261	0.874	0.845	0.82	0.684	3MD4	12	0.999	0.86	0.781	0.914
3BQX	139	0.900	0.634	0.481	0.297	3MD5	12	0.998	0.649	0.413	-0.218
3BZQ	99	0.848	0.532	0.516	0.466	3MEA	166	0.872	0.669	0.669	0.6
3BZZ	100	0.783	0.485	0.45	0.6	3MGN	348	0.742	0.205	0.119	0.193
3DRF	547	0.781	0.559	0.549	0.488	3MRE	383	0.675	0.661	0.641	0.567
3DWV	325	0.754	0.707	0.661	0.547	3N11	325	0.736	0.614	0.583	0.517
3E5T	228	0.731	0.502	0.489	0.296	3NE0	208	0.859	0.706	0.645	0.659
3E7R	40	0.769	0.706	0.687	0.642	3NGG	94	0.867	0.696	0.689	0.719
3EUR	140	0.874	0.431	0.427	0.577	3NPV	495	0.855	0.702	0.653	0.677
3F2Z	149	0.877	0.824	0.792	0.74	3NVG	6	1.000	0.721	0.617	0.597
3F7E	254	0.847	0.812	0.803	0.811	3NZL	73	0.713	0.627	0.583	0.506
3FCN	158	0.741	0.64	0.606	0.632	3O0P	194	0.898	0.727	0.706	0.734
3FE7	91	0.914	0.583	0.533	0.276	3O5P	128	0.787	0.734	0.698	0.63
3FKE	250	0.755	0.525	0.476	0.435	3OBQ	150	0.877	0.649	0.645	0.655
3FMY	66	0.857	0.701	0.655	0.556	3OQY	234	0.807	0.698	0.686	0.637
3FOD	48	0.725	0.532	0.44	-0.126	3P6J	125	0.689	0.774	0.767	0.81
3FSO	221	0.906	0.831	0.817	0.793	3PD7	188	0.848	0.77	0.723	0.589
3FTD	240	0.818	0.722	0.713	0.634	3PES	165	0.861	0.697	0.642	0.683
3FVA	6	1.000	0.835	0.825	0.789	3PID	387	0.677	0.537	0.531	0.642
3G1S	418	0.879	0.771	0.7	0.63	3PIW	154	0.772	0.758	0.744	0.717
3GBW	161	0.864	0.82	0.747	0.51	3PKV	221	0.731	0.625	0.597	0.568
3GHJ	116	0.864	0.732	0.511	0.196	3PSM	94	0.914	0.876	0.79	0.745
3HFO	197	0.825	0.691	0.67	0.518	3PTL	289	0.611	0.543	0.541	0.468
3HHP	1234	0.830	0.72	0.716	0.683	3PVE	347	0.785	0.718	0.667	0.568
3HNY	156	0.885	0.793	0.723	0.758	3PZ9	357	0.758	0.709	0.709	0.678
3HP4	183	0.690	0.534	0.5	0.573	3PZZ	12	0.998	0.945	0.922	0.95
3HWU	144	0.905	0.754	0.748	0.841	3Q2X	6	1.000	0.922	0.904	0.866
3HYD	7	1.000	0.966	0.95	0.867	3Q6L	131	0.723	0.622	0.577	0.605
3HZ8	192	0.857	0.617	0.502	0.475	3QDS	284	0.782	0.78	0.745	0.568
3I2V	124	0.879	0.486	0.441	0.301	3QPA	197	0.616	0.587	0.442	0.503
3I2Z	138	0.732	0.613	0.599	0.317	3R6D	221	0.854	0.688	0.669	0.495
3I4O	135	0.767	0.735	0.714	0.738	3R87	132	0.861	0.452	0.419	0.286

Table 7.4 (cont'd)

PDB ID	N	MWCG	opFRI	pfFRI	GNM	PDB ID	N	MWCG	opFRI	pfFRI	GNM
3I7M	134	0.604	0.667	0.635	0.695	3RQ9	162	0.711	0.51	0.403	0.242
3IHS	169	0.807	0.586	0.565	0.409	3RY0	128	0.790	0.616	0.606	0.47
3IVV	149	0.866	0.817	0.797	0.693	3RZY	139	0.867	0.8	0.784	0.849
3K6Y	227	0.817	0.586	0.535	0.301	3S0A	119	0.713	0.562	0.524	0.526
3KBE	140	0.743	0.705	0.704	0.611	3SD2	86	0.842	0.523	0.421	0.237
3K GK	190	0.798	0.784	0.775	0.68	3SEB	238	0.879	0.801	0.712	0.826
3KZD	85	0.789	0.647	0.611	0.475	3SED	124	0.870	0.709	0.658	0.712
3L41	220	0.776	0.718	0.716	0.669	3SO6	150	0.747	0.675	0.666	0.63
3LAA	169	0.880	0.827	0.647	0.659	3SR3	637	0.633	0.619	0.611	0.624
3LAX	106	0.924	0.734	0.73	0.584	3SUK	248	0.721	0.644	0.633	0.567
3LG3	833	0.701	0.658	0.614	0.589	3SZH	697	0.860	0.817	0.815	0.697
3LJI	272	0.720	0.612	0.608	0.551	3T0H	208	0.897	0.808	0.775	0.694
3M3P	249	0.697	0.584	0.554	0.338	3T3K	122	0.803	0.796	0.748	0.735
3M8J	178	0.813	0.73	0.728	0.628	3T47	141	0.759	0.592	0.527	0.447
3M9J	210	0.867	0.639	0.574	0.296	3TDN	357	0.668	0.458	0.419	0.24
3M9Q	176	0.851	0.591	0.51	0.471	3TOW	152	0.722	0.578	0.556	0.571
3MAB	173	0.770	0.664	0.591	0.451	3TUA	210	0.696	0.665	0.658	0.588
3U6G	248	0.808	0.635	0.632	0.526	3TYS	75	0.918	0.853	0.8	0.791
3U97	77	0.819	0.753	0.736	0.712	4DT4	160	0.784	0.776	0.738	0.716
3UCI	72	0.689	0.589	0.526	0.495	4EK3	287	0.830	0.68	0.68	0.674
3UR8	637	0.832	0.666	0.652	0.597	4ERY	318	0.801	0.74	0.701	0.688
3US6	148	0.668	0.698	0.586	0.553	4ES1	95	0.820	0.648	0.625	0.551
3V1A	48	0.811	0.531	0.487	0.583	4EUG	225	0.592	0.57	0.529	0.405
3V75	285	0.674	0.604	0.596	0.491	4F01	448	0.883	0.633	0.372	0.688
3VN0	193	0.889	0.84	0.837	0.812	4F3J	143	0.879	0.617	0.598	0.551
3VOR	182	0.686	0.602	0.557	0.484	4FR9	141	0.806	0.671	0.655	0.501
3VUB	101	0.852	0.625	0.61	0.607	4G14	15	1.000	0.467	0.323	0.356
3VVV	108	0.951	0.833	0.741	0.753	4G2E	151	0.835	0.76	0.755	0.758
3VZ9	163	0.887	0.785	0.749	0.695	4G5X	550	0.822	0.786	0.754	0.743
3W4Q	773	0.798	0.737	0.725	0.649	4G6C	658	0.834	0.591	0.59	0.528
3ZBD	213	0.891	0.651	0.516	0.632	4G7X	194	0.840	0.688	0.587	0.624
3ZIT	152	0.641	0.43	0.404	0.392	4GA2	144	0.782	0.528	0.485	0.406
3ZRX	221	0.639	0.59	0.562	0.391	4GMQ	92	0.794	0.678	0.628	0.55
3ZSL	138	0.903	0.691	0.687	0.526	4GS3	90	0.698	0.544	0.522	0.547
3ZZP	74	0.692	0.524	0.46	0.448	4H4J	236	0.866	0.81	0.806	0.689
3ZZY	226	0.804	0.746	0.709	0.728	4H89	168	0.624	0.682	0.588	0.596
4A02	166	0.730	0.618	0.516	0.303	4HDE	168	0.783	0.745	0.728	0.615
4ACJ	167	0.827	0.748	0.746	0.759	4HJP	281	0.730	0.703	0.649	0.51
4AE7	186	0.862	0.724	0.717	0.717	4HWM	117	0.807	0.638	0.622	0.499
4AM1	345	0.796	0.674	0.619	0.46	4IL7	85	0.719	0.446	0.404	0.316
4ANN	176	0.562	0.551	0.536	0.47	4J11	357	0.726	0.62	0.562	0.401
4AVR	188	0.759	0.68	0.605	0.65	4J5O	220	0.817	0.793	0.757	0.777

Table 7.4 (cont'd)

PDB ID	N	MWCG	opFRI	pfFRI	GNM	PDB ID	N	MWCG	opFRI	pfFRI	GNM
4AXY	54	0.973	0.7	0.623	0.72	4J5Q	146	0.851	0.742	0.742	0.689
4B6G	558	0.804	0.765	0.756	0.669	4J78	305	0.729	0.658	0.648	0.608
4B9G	292	0.855	0.844	0.816	0.763	4JG2	185	0.889	0.746	0.736	0.543
4DD5	387	0.850	0.615	0.596	0.351	4JVU	207	0.800	0.723	0.697	0.553
4DKN	423	0.786	0.781	0.761	0.539	4JYP	534	0.800	0.688	0.682	0.538
4DND	95	0.829	0.763	0.75	0.582	4KEF	133	0.704	0.58	0.53	0.324
4DPZ	109	0.837	0.73	0.726	0.651	5CYT	103	0.548	0.441	0.421	0.331
4DQ7	328	0.776	0.69	0.683	0.376	6RXN	45	0.583	0.614	0.574	0.594

As reported in Bramer et al [1], the Pearson correlation coefficients for small, medium and large proteins, as well as the average Pearson correlation coefficient of the protein superset, are provided in Table 7.5. In addition to MWCG, the average Pearson correlation coefficients for opFRI, pFRI, GNM, and NMA are also included for comparison. As determined by Park *et al*, GNM is more accurate than NMA, as analyzed by Park *et al* [4]. Moreover, opFRI and pfFRI are more accurate than GNM and the MWCG method presented in this work is on average approximately 28% more accurate than pfFRI and 42% more accurate than GNM.

Table 7.6 provides the average Pearson correlation coefficient obtained from MWCG linear least square fitting for C_α , non C_α carbon, nitrogen, oxygen, and sulfur atom based B factor prediction. It is notable that these predictions were not available to earlier GNM and FRI methods, thus no comparison can be provided for this result.

7.4 Machine Learning Results

7.4.1 MWCG

B factors of all carbon, nitrogen, and oxygen atoms present in a given protein were blindly predicted using a leave-one-(protein)-out approach. Results for predicted C_α B factors are

also included for comparison between other methods. These results are predicted in the same way as other heavy atoms. The machine learning B factor prediction models were trained using the generated input feature and B factor data from a training data set as described in Sections 5.6.1 and 5.6. After training the model is used to predict B factors for all heavy atoms in a given protein using only its feature input data.

7.4.1.1 Efficiency comparison

It is important for any algorithmic approach to consider the computational efficiency of the method. For B factor predictions this is a particularly important consideration for large proteins. The running times of the GNM, RF, GBT, and CNN models used for testing the MWCG method are provided in Table 7.7. Figure 7.7 provides a log-log comparison of these times. The protein set used to test the computational complexity were the same as those used by Opron *et al* [3]. Because GNM only provides C_α B factor predictions, only B factors of C_α atoms are predicted by the RF, GBT, and CNN models for this comparison. Because GNM is computational prohibitive for large proteins, several proteins were excluded from the test set for GNM predictions. All testing excludes the time it takes to load PDB files and feature data. The RF, GBT, and CNN times exclude the training of the model which can be used for the prediction of all proteins once they are trained. The results agree with the theoretical complexity $\mathcal{O}(N^3)$ for GNM. This is due to the matrix diagonalization required for GNM. In contrast the machine learning algorithms are close to $\mathcal{O}(N)$, where N is the number of atoms. The lines of best fit for CPU time (t) are $t \approx (4 \times 10^{-8}) * N^{3.09}$ for GNM, $t \approx (9 \times 10^{-6}) * N^{0.78}$ for RF, $t \approx (4 \times 10^{-6}) * N^{0.87}$ for GBT, and $t \approx (1.1 \times 10^{-3}) * N^{0.97}$ for CNN.

7.4.1.2 Machine learning performance

Table 7.8 provides the results for the blind prediction of all heavy atoms over the protein dataset. Overall the convolutional neural network method performs best with average Pearson correlation coefficient of 0.69. Both gradient boosted and random forest perform similarly with Pearson correlation coefficients of 0.63 and 0.59 respectively. Table 7.8 provides the results of the average Pearson correlation coefficients for C_α only B factor predictions, which are obtained in the same manner as other heavy atoms. This allows a comparison between previous methods. For comparison, the parameter-free flexibility-rigidity index (pfFRI), Gaussian network model (GNM) and normal mode analysis (NMA) are all included. The predictions of these previous methods are all obtained via the least squares fitting of each protein.

B factor prediction results are also included in Tables 7.9, 7.10, and 7.11 for the small-, medium-, and large-sized protein data subsets [4]. The results B factor predictions of all proteins in the protein Superset are provided in Table 7.12. The averages over the data subsets and superset is provided in Table 7.8. Over the different subsets all methods provided similar performance in terms of Pearson correlation coefficient. The deep convolutional neural network performed best on the protein Superset for both C_α only and all heavy atom B factor predictions.

The blind cross protein B factor prediction obtained in this work is particularly notable because it improves upon the best existing fitting methods. Previous work by Opron *et al* used the single protein linear least squares parameter-free FRI (pfFRI) method to obtain an average Pearson correlation coefficient of 0.63 averaged over the superset [3]. GNM performs worse with an overall Pearson correlation coefficient of 0.57 averaged over the superset [3].

Cross protein blind prediction is a much more difficult task than linear fitting. Table 7.12 shows that none of the machine learning methods outperform one another over the entire data set. Averaged over the superset, the Pearson correlation coefficient for the all heavy atom B factor prediction of the convolutional neural network outperformed the boosted gradient and random forest by 10% and 17% respectively.

Table 7.12: Pearson correlation coefficients for cross protein heavy atom blind B factor prediction obtained by random forest (RF), boosted gradient (GBT), and convolutional neural network (CNN) for the Superset. Results reported use heavy atoms in both training and prediction. MWCG machine learning results originally published in Bramer *et al* [2].

PDB ID	N	RF	GBT	CNN	PDB ID	N	RF	GBT	CNN
1ABA	728	0.74	0.77	0.73	2X5Y	1352	0.75	0.79	0.72
1AHO	482	0.62	0.71	0.76	2X9Z	1956	0.71	0.72	0.76
1AIE	235	0.62	0.53	0.60	2XHF	2432	0.65	0.71	0.70
1AKG	108	0.41	0.51	0.70	2Y0T	757	0.59	0.75	0.73
1ATG	1689	0.61	0.66	0.63	2Y72	1171	0.73	0.80	0.75
1BGF	1018	0.58	0.63	0.63	2Y7L	2398	0.81	0.82	0.62
1BX7	345	0.55	0.67	0.63	2Y9F	1212	0.72	0.77	0.64
1BYI	1540	0.59	0.63	0.59	2YLB	3065	0.60	0.69	0.63
1CCR	837	0.70	0.67	0.66	2YNY	2364	0.67	0.71	0.68
1CYO	697	0.66	0.68	0.76	2ZCM	2959	0.41	0.45	0.44
1DF4	463	0.79	0.75	0.64	2ZU1	2794	0.59	0.73	0.17
1E5K	1423	0.70	0.73	0.74	3A0M	823	0.65	0.47	0.74
1ES5	1912	0.63	0.68	0.66	3A7L	963	0.66	0.75	0.81
1ETL	76	0.27	0.03	0.48	3AMC	5174	0.72	0.75	0.62
1ETM	80	0.46	0.13	0.48	3AUB	782	0.63	0.62	0.74
1ETN	77	0.33	0.25	0.20	3B5O	1510	0.53	0.55	0.65
1EW4	863	0.70	0.71	0.61	3BA1	2391	0.65	0.64	0.44
1F8R	15291	0.64	0.64	0.83	3BED	1570	0.73	0.73	0.70
1FF4	477	0.55	0.59	0.76	3BQX	1028	0.52	0.59	0.85
1FK5	626	0.62	0.71	0.63	3BZQ	742	0.60	0.61	0.43
1GCO	7888	0.64	0.61	0.71	3BZZ	773	0.45	0.45	0.77
1GK7	321	0.53	0.73	0.72	3DRF	4101	0.67	0.66	0.81
1GVD	401	0.66	0.69	0.71	3DWV	2363	0.60	0.67	0.87
1GXU	694	0.65	0.67	0.66	3E5T	1543	0.71	0.72	0.75
1H6V	22514	0.39	0.40	0.58	3E7R	295	0.60	0.60	0.81
1HJE	73	-0.07	0.46	0.37	3EUR	1059	0.47	0.50	0.82
1I71	683	0.57	0.62	0.66	3F2Z	1160	0.78	0.78	0.88
1IDP	3661	0.69	0.74	0.83	3F7E	1912	0.61	0.67	0.69
1IFR	878	0.72	0.74	0.73	3FCN	1039	0.68	0.71	0.73

Table 7.12 (cont'd)

PDB ID	N	RF	GBT	CNN	PDB ID	N	RF	GBT	CNN
1K8U	686	0.65	0.68	0.74	3FE7	710	0.62	0.71	0.83
1KMM	11632	0.65	0.70	0.87	3FKE	1938	0.57	0.56	0.76
1KNG	1016	0.61	0.56	0.55	3FMY	470	0.73	0.75	0.84
1KR4	906	0.73	0.76	0.72	3FOD	328	0.30	0.45	0.78
1KYC	138	0.43	0.30	0.32	3FSO	197	0.71	0.73	0.85
1LR7	522	0.53	0.70	0.71	3FTD	1795	0.75	0.75	0.69
1MF7	1551	0.68	0.68	0.70	3G1S	3196	0.74	0.76	0.72
1N7E	700	0.62	0.65	0.71	3GBW	1275	0.75	0.76	0.68
1NKD	426	0.56	0.59	0.63	3GHJ	808	0.66	0.71	0.44
1NLS	1746	0.61	0.64	0.56	3HFO	1432	0.65	0.72	0.70
1NNX	674	0.69	0.73	0.53	3HHP	8495	0.71	0.74	0.62
1NOA	778	0.52	0.57	0.57	3HNY	1351	0.73	0.73	0.58
1NOT	96	-0.18	0.81	0.63	3HP4	1322	0.61	0.63	0.65
1O06	142	0.51	0.64	0.65	3HWU	934	0.51	0.69	0.51
1O08	1722	0.51	0.58	0.55	3HYD	52	-0.05	0.28	0.60
1OPD	642	0.55	0.60	0.62	3HZ8	1459	0.51	0.54	0.76
1P9I	203	0.73	0.77	0.77	3I2V	929	0.50	0.54	0.81
1PEF	153	0.60	0.64	0.76	3I2Z	1039	0.63	0.64	0.75
1PEN	109	0.34	0.24	0.21	3I4O	969	0.66	0.64	0.87
1PMY	937	0.64	0.65	0.67	3I7M	928	0.56	0.60	0.87
1PZ4	874	0.73	0.73	0.74	3IHS	1120	0.66	0.65	0.81
1Q9B	303	0.41	0.67	0.75	3IVV	1097	0.72	0.81	0.85
1QAU	812	0.57	0.58	0.57	3K6Y	1617	0.62	0.65	0.90
1QKI	31154	0.44	0.27	0.84	3KBE	829	0.75	0.76	0.86
1QTO	934	0.61	0.55	0.63	3K GK	1492	0.75	0.78	0.87
1R29	971	0.61	0.73	0.72	3KZD	605	0.64	0.70	0.74
1R7J	729	0.71	0.70	0.65	3L41	1735	0.73	0.76	0.88
1RJU	257	0.71	0.75	0.73	3LAA	1112	0.54	0.46	0.89
1RRO	846	0.56	0.52	0.54	3LAX	753	0.69	0.71	0.89
1SAU	830	0.62	0.68	0.60	3LG3	6061	0.57	0.59	0.91
1TGR	749	0.61	0.65	0.67	3LJI	1946	0.46	0.54	0.50
1TZV	1051	0.75	0.77	0.75	3M3P	1858	0.57	0.62	0.68
1U06	432	0.55	0.68	0.61	3M8J	1396	0.78	0.77	0.68
1U7I	1988	0.73	0.75	0.77	3M9J	1329	0.66	0.74	0.50
1U9C	1712	0.61	0.64	0.58	3M9Q	1359	0.52	0.53	0.48
1UHA	623	0.74	0.80	0.75	3MAB	1311	0.63	0.65	0.59
1UKU	873	0.74	0.75	0.70	3MD4	81	0.36	0.61	0.79
1ULR	677	0.69	0.71	0.68	3MEA	1236	0.58	0.64	0.93
1UOY	452	0.55	0.56	0.55	3MGN	2236	0.15	0.03	0.82
1USE	290	0.25	0.50	0.68	3MRE	2598	0.57	0.56	0.84
1USM	631	0.59	0.78	0.67	3N11	2501	0.52	0.57	0.85
1UTG	548	0.58	0.55	0.62	3NE0	1551	0.68	0.69	0.85

Table 7.12 (cont'd)

PDB ID	N	RF	GBT	CNN	PDB ID	N	RF	GBT	CNN
1V05	17	-0.20	0.02	0.60	3NGG	702	0.63	0.75	0.83
1V70	784	0.70	0.67	0.62	3NPV	3655	0.70	0.75	0.84
1VRZ	66	0.38	-0.17	0.09	3NVG	50	-0.08	0.08	0.88
1W2L	746	0.62	0.68	0.69	3NZL	567	0.59	0.65	0.63
1WBE	1542	0.59	0.61	0.63	3O0P	1452	0.55	0.65	0.63
1WHI	937	0.74	0.77	0.71	3O5P	819	0.53	0.63	0.70
1WLY	2430	0.65	0.71	0.68	3OBQ	1195	0.61	0.61	0.84
1WPA	906	0.64	0.66	0.74	3OQY	1772	0.57	0.62	0.76
1X3O	622	0.53	0.52	0.63	3P6J	857	0.57	0.70	0.88
1XY1	124	0.58	0.19	0.47	3PD7	1354	0.70	0.72	0.85
1XY2	62	0.16	0.27	0.55	3PES	1240	0.72	0.73	0.84
1Y6X	669	0.44	0.53	0.46	3PID	3078	0.49	0.56	0.86
1YJO	55	0.36	0.12	0.02	3PIW	1223	0.72	0.75	0.87
1YZM	361	0.51	0.60	0.56	3PKV	1688	0.66	0.68	0.81
1Z21	771	0.63	0.66	0.63	3PSM	729	0.62	0.68	0.80
1ZCE	1100	0.77	0.81	0.73	3PTL	2101	0.61	0.62	0.72
1ZVA	551	0.59	0.56	0.58	3PVE	2656	0.56	0.61	0.46
2A50	3493	0.64	0.48	0.68	3PZ9	2913	0.63	0.76	0.60
2AGK	1867	0.61	0.68	0.44	3PZZ	76	0.47	0.25	0.85
2AH1	7215	0.65	0.57	0.67	3Q2X	43	0.29	0.59	0.76
2B0A	1454	0.66	0.68	0.72	3Q6L	1022	0.71	0.67	0.75
2BCM	3002	0.51	0.62	0.85	3QDS	2234	0.71	0.72	0.71
2BF9	287	0.39	0.52	0.70	3QPA	1348	0.43	0.44	0.71
2BRF	735	0.76	0.78	0.86	3R6D	1550	0.31	0.69	0.59
2C71	1446	0.59	0.61	0.83	3R87	1007	0.39	0.51	0.53
2CE0	714	0.62	0.65	0.90	3RQ9	1174	0.32	0.47	0.66
2CG7	536	0.47	0.54	0.79	3RY0	964	0.66	0.65	0.53
2COV	4366	0.76	0.83	0.78	3RZY	985	0.69	0.69	0.64
2CWS	1624	0.63	0.60	0.78	3S0A	884	0.55	0.61	0.61
2D5W	9772	0.71	0.75	0.75	3SD2	527	0.38	0.52	0.71
2DKO	1933	0.71	0.72	0.72	3SEB	1948	0.61	0.71	0.57
2DPL	4454	0.49	0.53	0.73	3SED	933	0.70	0.71	0.72
2DSX	386	0.36	0.44	0.56	3SO6	1119	0.69	0.75	0.01
2E10	3416	0.50	0.64	0.61	3SR3	4891	0.69	0.69	0.45
2E3H	589	0.70	0.73	0.38	3SUK	1761	0.62	0.65	0.59
2EAQ	705	0.63	0.61	0.58	3SZH	5074	0.74	0.80	0.44
2EHP	1875	0.75	0.74	0.74	3T0H	1627	0.78	0.81	0.65
2EHS	590	0.55	0.71	0.38	3T3K	922	0.56	0.68	0.62
2ERW	385	0.47	0.50	0.32	3T47	1116	0.54	0.62	0.74
2ETX	3018	0.56	0.61	0.58	3TDN	2703	0.55	0.55	0.58
2FB6	766	0.63	0.65	0.52	3TOW	1193	0.53	0.66	0.66
2FG1	1021	0.55	0.65	0.68	3TUA	1510	0.63	0.66	0.70

Table 7.12 (cont'd)

PDB ID	N	RF	GBT	CNN	PDB ID	N	RF	GBT	CNN
2FN9	4362	0.37	0.60	0.61	3TYS	556	0.67	0.68	0.71
2FQ3	721	0.67	0.75	0.76	3U6G	1658	0.52	0.51	0.60
2G69	744	0.60	0.61	0.87	3U97	524	0.57	0.66	0.27
2G7O	537	0.52	0.63	0.89	3UCI	536	0.44	0.51	0.56
2G7S	1258	0.60	0.60	0.81	3UR8	5033	0.63	0.66	0.83
2GKG	706	0.63	0.60	0.70	3US6	1156	0.62	0.64	0.01
2GOM	987	0.61	0.70	0.92	3V1A	319	0.36	0.36	0.76
2GXG	1132	0.67	0.75	0.86	3V75	1974	0.63	0.65	0.83
2GZQ	1402	0.59	0.60	0.90	3VN0	1469	0.69	0.76	0.76
2HQK	1582	0.76	0.76	0.90	3VOR	1077	0.41	0.50	0.81
2HYK	1832	0.60	0.65	0.85	3VUB	787	0.64	0.70	0.78
2I24	872	0.52	0.52	0.91	3VVV	869	0.62	0.69	0.84
2I49	3109	0.78	0.77	0.90	3VZ9	1366	0.70	0.72	0.66
2IBL	815	0.46	0.53	0.88	3W4Q	5406	0.66	0.73	0.65
2IGD	431	0.58	0.68	0.82	3ZBD	1718	0.54	0.54	0.78
2IMF	1564	0.62	0.62	0.47	3ZIT	1192	0.51	0.54	0.71
2IP6	702	0.62	0.67	0.64	3ZRX	1654	0.38	0.67	0.60
2IVY	727	0.47	0.59	0.62	3ZSL	925	0.61	0.64	0.69
2J32	1935	0.79	0.78	0.70	3ZZP	585	0.40	0.46	0.56
2J9W	1626	0.66	0.68	0.73	3ZZY	1741	0.64	0.69	0.69
2JKU	229	0.57	0.63	0.35	4A02	1281	0.62	0.65	0.75
2JLI	708	0.58	0.54	0.73	4ACJ	1210	0.64	0.67	0.75
2JLJ	889	0.66	0.70	0.68	4AE7	1458	0.64	0.74	0.61
2MCM	735	0.71	0.73	0.60	4AM1	2605	0.64	0.67	0.56
2NLS	269	0.45	0.49	0.70	4ANN	1180	0.53	0.60	0.72
2NR7	1556	0.71	0.70	0.66	4AVR	1437	0.62	0.61	0.64
2NUH	806	0.64	0.72	0.19	4AXY	317	0.45	0.64	0.75
2O6X	2415	0.76	0.82	0.63	4B6G	4504	0.78	0.76	0.84
2OA2	970	0.54	0.53	0.92	4B9G	2226	0.79	0.81	0.83
2OHW	2074	0.55	0.62	0.81	4DD5	2618	0.63	0.66	0.87
2OKT	2587	0.56	0.59	0.89	4DKN	3356	0.76	0.77	0.88
2OL9	51	0.65	0.51	0.84	4DND	755	0.66	0.73	0.85
2PKT	666	0.06	0.17	0.76	4DPZ	865	0.65	0.66	0.83
2PLT	719	0.62	0.67	0.70	4DQ7	2526	0.58	0.69	0.78
2PMR	590	0.63	0.66	0.63	4DT4	1163	0.71	0.73	0.73
2POF	3418	0.58	0.66	0.85	4EK3	2147	0.70	0.72	0.73
2PPN	701	0.50	0.68	0.83	4ERY	2357	0.70	0.74	0.83
2PSF	4983	0.54	0.55	0.79	4ES1	737	0.63	0.64	0.81
2PTH	1437	0.68	0.72	0.79	4EUG	1789	0.59	0.66	0.79
2Q4N	9496	0.45	0.39	0.85	4F01	3374	0.55	0.54	0.77
2Q52	26784	0.63	0.62	0.77	4F3J	1116	0.58	0.62	0.53
2QJL	734	0.61	0.60	0.42	4FR9	956	0.61	0.64	0.62

Table 7.12 (cont'd)

PDB ID	N	RF	GBT	CNN	PDB ID	N	RF	GBT	CNN
2R16	1262	0.52	0.53	0.50	4G14	39	0.28	0.50	0.55
2R6Q	903	0.59	0.53	0.57	4G2E	1178	0.73	0.73	0.76
2RB8	723	0.61	0.64	0.42	4G5X	4002	0.74	0.75	0.65
2RE2	1559	0.66	0.66	0.54	4G6C	4814	0.47	0.60	0.61
2RFR	1019	0.54	0.58	0.66	4G7X	1315	0.49	0.56	0.80
2V9V	986	0.64	0.61	0.63	4GA2	873	0.51	0.55	0.55
2VE8	3967	0.65	0.59	0.66	4GMQ	678	0.56	0.72	0.54
2VH7	749	0.74	0.70	0.82	4GS3	737	0.56	0.60	0.56
2VIM	781	0.62	0.61	0.75	4H4J	1470	0.69	0.80	0.70
2VPA	1524	0.63	0.68	0.61	4H89	1127	0.55	0.61	0.62
2VQ4	800	0.72	0.76	0.78	4HDE	1288	0.73	0.79	0.70
2VY8	1058	0.71	0.74	0.63	4HJP	2112	0.65	0.70	0.76
2VYO	1589	0.53	0.65	0.61	4HWM	799	0.50	0.57	0.81
2W1V	4223	0.68	0.72	0.72	4IL7	527	0.35	0.43	0.74
2W2A	2918	0.56	0.62	0.63	4J11	2658	0.47	0.58	0.94
2W6A	826	0.66	0.76	0.69	4J5O	1406	0.64	0.63	0.91
2WJ5	630	0.49	0.53	0.77	4J5Q	1062	0.73	0.75	0.87
2WUJ	828	0.55	0.55	0.55	4J78	2443	0.71	0.75	0.86
2WW7	915	0.35	0.43	0.61	4JG2	1294	0.70	0.73	0.88
2WWE	54	0.23	0.22	0.12	4JVU	1615	0.69	0.68	0.89
2X1Q	1852	0.58	0.53	0.77	4JYP	4063	0.70	0.78	0.93
2X25	1289	0.65	0.68	0.80	4KEF	1002	0.65	0.62	0.68
2X3M	1267	0.66	0.70	0.75	5CYT	800	0.68	0.70	0.74
					6RXN	345	0.56	0.71	0.82

Several proteins have low Pearson correlation coefficients indicating a poor model prediction. In these cases we see that if one model performs poorly the other models generally perform satisfactorily. Taking the consensus of the maximum correlation coefficient for each protein among the three machine learning methods results in an average all heavy atom correlation coefficient of 0.73 and an average C_{α} only correlation coefficient of 0.72. This result is similar to that of the parameter-optimized FRI (opFRI) reported in earlier work by Opron *et al* [3].

7.4.1.3 Relative feature importance

Ensemble methods provide relative feature importance of the features used in the resulting models. This is an important tool to help understand which features are most significant in a model. Figure 7.8 shows the individual feature importance for the random forest averaged over the protein superset.

Since several of the features are related, Figure 7.9 provides a plot of the aggregated feature importance. The feature importance of the individual angle, secondary, MWCG, atom type, protein size, amino acid, and packing density features are all summed together to illustrate the overall effect of each feature type.

Figure 7.8 shows the most important MWCG feature is the carbon-carbon interaction. This MWCG feature uses a Lorentz radial basis function as with $\eta = 16$ and $\nu = 3$ as detailed in Section 5.3. The remaining eight MWCG features all rank similarly with the carbon-oxygen interaction ranked as the second most significant MWCG feature. This result validates that the model benefits from the multi-scale property of the MWCG feature, which uses three different kernels to capture interactions at various length scales. Since all MWCG have significance in the feature ranking it follows that the element specific property of the MWCG method is also a meaningful model feature.

Figure 7.8 shows that that the individual MWCG, amino acid type, and packing density feature have low relative importance, however, considering their aggregate importance as seen in Figure 7.9, we see that they contribute to the model. Figure 7.9 shows that the medium density protein packing density feature was twice as important to the model as the short and long density features. The medium packing density may be capturing semi-local side chain interactions which are important in protein flexibility. The short packing

density likely captures only adjacent backbone information while the long packing density is only adding weak atomic interaction information to the model. Protein resolution is the most significant relative feature followed by MWCG features and the STRIDE generated residue solvent accessible area feature. This also highlights the importance of the quality of X-ray crystal structures and difficulty in cross-protein B factor prediction. Protein angles, secondary structures, and size play a less significant role in the model compared to the other features. Atom type has the lowest significance relative to the other features implemented in the model. Not surprisingly, we see that global features such as resolution and R-value are important components in the ensemble model. The global feature of protein size has a small role in the model.

Care must be taken to use feature ranking to understand feature importance. The feature ranking provided by these models is a relative ordering of features that the models find most important. So features with high correlation may be redundant giving one of them a lower rank even though they may have significant prediction power. For example, R-value highly correlates with resolution so it is likely a meaningful feature. However, the use of resolution reduces the relative importance ranking of R-value in the model.

7.5 ASPH & ESPH B Factor Prediction

7.5.1 Least Squares Fitting

The Pearson correlation coefficients using least squares fitting for C_α B factor prediction of small, medium, and large protein subsets are provided in tables 7.17, 7.18, and 7.19 respectively. Results for the all proteins in the dataset are provided in table 7.21. The average Pearson correlation coefficients for small, medium, large, and superset data sets

is provided in table 7.20. Table 7.20 includes fitting results using only Bottleneck, only Wasserstein, and using both Bottleneck and Wasserstein metrics. Results using only an exponential kernel, only a lorentz kernel, or both an exponential and lorentz kernel for fitting are also included. All results reported here PH features generated with a cutoff of 11Å and include three pairwise interactions (carbon-carbon, carbon-nitrogen, carbon-oxygen).

7.5.2 Machine Learning

ASPH and ESPH Pearson correlation coefficients using boosted gradient (GBT), convolutional neural network (CNN), and consensus method (CON) for C_α B factor prediction of small, medium, and large protein subsets is provided in tables 7.14, 7.15, and 7.16 respectively. Parameters for GBT and CNN methods can be found in Tables 5.7 and 5.8. The global and local features used for training and testing are provided in chapter 5. Results for all proteins are provided in table 7.22. The average Pearson correlation coefficients for small, medium, large, and superset data sets is provided in table 7.13. All results reported here use a cutoff of 11Å and include three pairwise interactions (carbon-carbon, carbon-nitrogen, carbon-oxygen). Kernel parameters for both exponential and lorentz kernels are provided in Table 5.4. Results from previously existing C_α B factor prediction methods are included for comparison in Table 7.13. Overall both GBT and CNN algorithms perform similarly. As expected, the CNN method out performs the GBT with average correlation coefficients over the superset of 0.60 and 0.59 respectively. The consensus method improves upon both results with an average Pearson correlation coefficient of 0.61 over the superset. Table 7.13 shows that the blind prediction machine learning models perform better than fitting models GNM and NMA and similar to the pFRI fitting model.

Table 7.21: Pearson correlation coefficients of persistent homology based least squares fitting C_α B factor prediction of all proteins using 11Å cutoff. Two Bottleneck (B) and Wasserstein (W) metrics using various kernel choices are included.

PDB ID	B & W			B			W			
	N	Exp	Lor	Both	Exp	Lor	Both	Exp	Lor	Both
1ABA	87	0.67	0.67	0.76	0.54	0.62	0.68	0.56	0.63	0.70
1AHO	66	0.75	0.78	0.88	0.72	0.73	0.79	0.53	0.65	0.75
1AIE	31	0.97	0.88	0.99	0.78	0.64	0.90	0.90	0.77	0.96
1AKG	16	0.82	0.66	1.00	0.60	0.53	0.72	0.53	0.56	0.87
1ATG	231	0.50	0.50	0.61	0.45	0.47	0.53	0.38	0.48	0.51
1BGF	124	0.75	0.70	0.82	0.64	0.54	0.75	0.68	0.61	0.75
1BX7	51	0.86	0.74	0.89	0.79	0.68	0.82	0.81	0.69	0.82
1BYI	238	0.50	0.51	0.58	0.41	0.46	0.49	0.44	0.48	0.54
1CCR	109	0.65	0.66	0.71	0.53	0.56	0.65	0.43	0.58	0.63
1CYO	88	0.71	0.69	0.78	0.66	0.58	0.68	0.65	0.59	0.67
1DF4	57	0.93	0.92	0.97	0.92	0.89	0.95	0.88	0.91	0.94
1E5K	188	0.67	0.68	0.74	0.66	0.67	0.68	0.63	0.67	0.69
1ES5	260	0.58	0.57	0.65	0.51	0.55	0.58	0.44	0.56	0.60
1ETL	12	1.00	1.00	1.00	0.68	0.87	1.00	0.95	0.98	1.00
1ETM	12	1.00	1.00	1.00	0.45	0.74	0.86	0.70	0.83	1.00
1ETN	12	1.00	1.00	1.00	0.96	0.92	0.99	0.70	0.92	1.00
1EW4	106	0.58	0.60	0.73	0.52	0.51	0.55	0.55	0.55	0.62
1F8R	1932	0.61	0.63	0.70	0.59	0.62	0.63	0.50	0.62	0.65
1FF4	65	0.77	0.72	0.80	0.70	0.65	0.75	0.68	0.68	0.76
1FK5	93	0.53	0.59	0.71	0.49	0.50	0.58	0.49	0.50	0.55
1GCO	1044	0.63	0.64	0.66	0.59	0.63	0.63	0.53	0.63	0.65
1GK7	39	0.95	0.94	0.98	0.91	0.93	0.95	0.88	0.92	0.94
1GVD	56	0.75	0.68	0.84	0.67	0.63	0.69	0.61	0.62	0.66
1GXU	89	0.75	0.78	0.82	0.72	0.61	0.75	0.69	0.72	0.77
1H6V	2927	0.29	0.31	0.33	0.28	0.29	0.30	0.23	0.29	0.30
1HJE	13	1.00	1.00	1.00	0.72	0.79	1.00	0.67	0.57	1.00
1I71	83	0.44	0.66	0.76	0.41	0.46	0.56	0.38	0.58	0.59
1IDP	441	0.48	0.47	0.55	0.43	0.45	0.47	0.39	0.46	0.48
1IFR	113	0.65	0.59	0.73	0.56	0.54	0.65	0.47	0.53	0.62
1K8U	87	0.72	0.74	0.85	0.67	0.64	0.71	0.65	0.67	0.75
1KMM	1499	0.57	0.54	0.59	0.49	0.53	0.54	0.36	0.53	0.57
1KNG	144	0.52	0.51	0.61	0.43	0.47	0.51	0.43	0.50	0.53
1KR4	107	0.57	0.48	0.60	0.39	0.47	0.53	0.45	0.45	0.54
1KYC	15	0.96	0.99	1.00	0.92	0.93	0.99	0.88	0.88	1.00
1LR7	73	0.61	0.62	0.71	0.57	0.55	0.63	0.46	0.56	0.58
1MF7	194	0.56	0.59	0.67	0.55	0.57	0.59	0.50	0.58	0.59
1N7E	95	0.67	0.71	0.80	0.54	0.68	0.72	0.54	0.63	0.73
1NKD	59	0.73	0.69	0.89	0.56	0.58	0.63	0.55	0.65	0.75

Table 7.21 (cont'd)

PDB ID	B & W			B			W			
	N	Exp	Lor	Both	Exp	Lor	Both	Exp	Lor	Both
1NLS	238	0.81	0.78	0.86	0.75	0.65	0.83	0.80	0.72	0.82
1NNX	93	0.84	0.84	0.88	0.81	0.79	0.83	0.81	0.81	0.86
1NOA	113	0.63	0.65	0.72	0.60	0.57	0.63	0.53	0.57	0.59
1NOT	13	1.00	1.00	1.00	0.82	0.86	1.00	0.86	0.81	1.00
1O06	22	0.98	0.97	1.00	0.96	0.92	0.97	0.97	0.94	0.98
1O08	221	0.46	0.48	0.56	0.44	0.42	0.50	0.37	0.45	0.48
1OB4	5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1OB7	5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1OPD	85	0.35	0.29	0.57	0.25	0.21	0.36	0.29	0.19	0.36
1P9I	29	0.89	0.88	0.98	0.87	0.82	0.92	0.87	0.84	0.89
1PEF	18	0.96	0.97	1.00	0.88	0.94	0.96	0.92	0.94	0.96
1PEN	16	0.96	0.90	1.00	0.60	0.67	0.83	0.47	0.73	0.94
1PMY	123	0.71	0.70	0.76	0.62	0.59	0.67	0.68	0.69	0.71
1PZ4	113	0.88	0.82	0.93	0.86	0.74	0.89	0.85	0.76	0.88
1Q9B	44	0.79	0.76	0.94	0.58	0.59	0.69	0.69	0.57	0.71
1QAU	112	0.59	0.61	0.66	0.57	0.55	0.58	0.55	0.57	0.58
1QKI	3912	0.38	0.42	0.45	0.34	0.38	0.41	0.32	0.38	0.40
1QTO	122	0.59	0.59	0.65	0.48	0.46	0.53	0.55	0.52	0.56
1R29	122	0.71	0.56	0.76	0.55	0.35	0.69	0.69	0.43	0.72
1R7J	90	0.88	0.86	0.91	0.83	0.76	0.87	0.81	0.79	0.86
1RJU	36	0.81	0.74	0.91	0.75	0.69	0.81	0.62	0.65	0.72
1RRO	108	0.39	0.35	0.56	0.31	0.23	0.45	0.33	0.19	0.45
1SAU	123	0.76	0.75	0.81	0.70	0.73	0.75	0.68	0.74	0.76
1TGR	111	0.77	0.76	0.83	0.72	0.70	0.74	0.74	0.73	0.75
1TZV	157	0.76	0.78	0.83	0.73	0.71	0.77	0.69	0.70	0.74
1U06	55	0.50	0.52	0.72	0.37	0.36	0.52	0.46	0.39	0.55
1U7I	259	0.71	0.71	0.73	0.62	0.68	0.70	0.53	0.67	0.71
1U9C	220	0.66	0.65	0.74	0.61	0.57	0.64	0.61	0.60	0.67
1UHA	82	0.70	0.75	0.82	0.69	0.68	0.74	0.67	0.69	0.73
1UKU	102	0.80	0.81	0.84	0.78	0.80	0.80	0.74	0.80	0.80
1ULR	87	0.56	0.53	0.68	0.49	0.50	0.59	0.44	0.50	0.61
1UOY	64	0.73	0.72	0.83	0.65	0.66	0.69	0.65	0.69	0.73
1USE	47	0.66	0.75	0.91	0.50	0.52	0.72	0.46	0.53	0.64
1USM	77	0.62	0.61	0.81	0.57	0.53	0.66	0.61	0.58	0.65
1UTG	70	0.57	0.53	0.68	0.51	0.49	0.60	0.49	0.49	0.56
1V05	96	0.67	0.66	0.72	0.60	0.61	0.65	0.52	0.61	0.65
1V70	105	0.64	0.65	0.75	0.56	0.60	0.66	0.51	0.58	0.62
1VRZ	13	1.00	1.00	1.00	0.92	0.92	1.00	0.77	0.85	1.00
1W2L	97	0.72	0.72	0.79	0.60	0.63	0.69	0.56	0.61	0.69
1WBE	206	0.53	0.47	0.63	0.43	0.38	0.55	0.36	0.42	0.48
1WHI	122	0.57	0.55	0.63	0.42	0.44	0.57	0.34	0.43	0.55

Table 7.21 (cont'd)

PDB ID	B & W			B			W			
	N	Exp	Lor	Both	Exp	Lor	Both	Exp	Lor	Both
1WLY	322	0.62	0.64	0.67	0.59	0.62	0.63	0.54	0.62	0.64
1WPA	107	0.70	0.69	0.79	0.61	0.52	0.71	0.66	0.56	0.70
1X3O	80	0.66	0.66	0.72	0.62	0.60	0.65	0.62	0.64	0.67
1XY1	16	0.97	0.96	1.00	0.73	0.66	0.87	0.81	0.89	0.99
1XY2	8	1.00	1.00	1.00	0.99	0.95	1.00	0.91	0.91	1.00
1Y6X	86	0.56	0.53	0.62	0.50	0.49	0.59	0.50	0.52	0.56
1YJO	6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1YZM	46	0.87	0.90	0.95	0.82	0.72	0.88	0.86	0.84	0.90
1Z21	96	0.70	0.73	0.82	0.61	0.63	0.64	0.64	0.69	0.72
1ZCE	139	0.84	0.83	0.88	0.83	0.77	0.85	0.81	0.78	0.82
1ZVA	75	0.85	0.85	0.94	0.84	0.78	0.92	0.83	0.81	0.86
2A50	469	0.64	0.63	0.70	0.54	0.60	0.67	0.41	0.58	0.67
2AGK	233	0.65	0.65	0.69	0.61	0.64	0.65	0.55	0.63	0.67
2AH1	939	0.45	0.47	0.49	0.42	0.45	0.46	0.33	0.46	0.48
2B0A	191	0.59	0.60	0.69	0.50	0.58	0.62	0.48	0.59	0.63
2BCM	415	0.46	0.41	0.50	0.39	0.39	0.40	0.35	0.39	0.45
2BF9	35	0.94	0.73	0.97	0.70	0.65	0.78	0.89	0.71	0.92
2BRF	103	0.74	0.73	0.76	0.74	0.71	0.74	0.72	0.72	0.75
2C71	225	0.45	0.38	0.56	0.29	0.33	0.42	0.23	0.30	0.48
2CE0	109	0.77	0.79	0.86	0.75	0.73	0.80	0.71	0.77	0.79
2CG7	110	0.32	0.44	0.63	0.29	0.31	0.36	0.30	0.33	0.41
2COV	534	0.66	0.64	0.70	0.63	0.64	0.67	0.57	0.64	0.67
2CWS	235	0.59	0.55	0.66	0.53	0.52	0.54	0.40	0.52	0.55
2D5W	1214	0.52	0.52	0.54	0.49	0.52	0.52	0.41	0.52	0.53
2DKO	253	0.75	0.72	0.79	0.72	0.69	0.75	0.68	0.69	0.72
2DPL	565	0.35	0.36	0.41	0.30	0.32	0.35	0.24	0.33	0.37
2DSX	52	0.54	0.50	0.78	0.37	0.30	0.56	0.41	0.36	0.55
2E10	439	0.60	0.59	0.65	0.51	0.58	0.61	0.43	0.57	0.62
2E3H	81	0.66	0.71	0.82	0.62	0.69	0.76	0.56	0.69	0.78
2EAQ	89	0.81	0.77	0.86	0.78	0.72	0.81	0.77	0.76	0.82
2EHP	246	0.63	0.65	0.71	0.58	0.62	0.65	0.52	0.62	0.64
2EHS	75	0.75	0.73	0.81	0.72	0.71	0.74	0.69	0.71	0.73
2ERW	53	0.62	0.41	0.84	0.33	0.26	0.60	0.31	0.28	0.49
2ETX	390	0.54	0.54	0.57	0.52	0.53	0.56	0.47	0.51	0.54
2FB6	129	0.71	0.66	0.76	0.67	0.63	0.69	0.65	0.63	0.74
2FG1	176	0.55	0.56	0.62	0.54	0.52	0.58	0.52	0.54	0.57
2FN9	560	0.51	0.49	0.62	0.44	0.47	0.55	0.41	0.46	0.55
2FQ3	85	0.78	0.76	0.82	0.75	0.75	0.79	0.68	0.75	0.78
2G69	99	0.59	0.65	0.76	0.42	0.50	0.66	0.47	0.45	0.60
2G7O	68	0.89	0.91	0.95	0.85	0.79	0.88	0.76	0.82	0.87
2G7S	206	0.63	0.60	0.66	0.59	0.58	0.63	0.54	0.59	0.63

Table 7.21 (cont'd)

PDB ID	B & W			B			W			
	N	Exp	Lor	Both	Exp	Lor	Both	Exp	Lor	Both
2GKG	150	0.77	0.71	0.83	0.74	0.65	0.78	0.76	0.67	0.78
2GOM	121	0.47	0.52	0.64	0.42	0.42	0.45	0.44	0.47	0.53
2GXG	140	0.74	0.72	0.79	0.71	0.68	0.72	0.69	0.68	0.73
2GZQ	203	0.45	0.40	0.60	0.38	0.34	0.48	0.24	0.29	0.31
2HQK	232	0.80	0.79	0.83	0.70	0.74	0.80	0.68	0.76	0.81
2HYK	237	0.59	0.58	0.63	0.51	0.55	0.59	0.43	0.54	0.60
2I24	113	0.47	0.44	0.69	0.40	0.40	0.48	0.45	0.40	0.49
2I49	399	0.54	0.53	0.62	0.43	0.51	0.56	0.41	0.49	0.58
2IBL	108	0.69	0.71	0.75	0.66	0.67	0.70	0.65	0.68	0.71
2IGD	61	0.67	0.72	0.84	0.61	0.64	0.74	0.61	0.66	0.74
2IMF	203	0.61	0.65	0.71	0.59	0.56	0.60	0.59	0.59	0.64
2IP6	87	0.72	0.66	0.82	0.66	0.58	0.73	0.64	0.64	0.78
2IVY	89	0.43	0.53	0.69	0.35	0.45	0.48	0.34	0.42	0.57
2J32	244	0.77	0.72	0.85	0.73	0.68	0.77	0.73	0.68	0.77
2J9W	203	0.59	0.60	0.70	0.55	0.59	0.64	0.51	0.59	0.62
2JKU	38	0.89	0.75	0.95	0.85	0.65	0.88	0.83	0.60	0.88
2JLI	112	0.87	0.81	0.90	0.82	0.70	0.85	0.85	0.78	0.86
2JLJ	121	0.78	0.75	0.80	0.71	0.65	0.74	0.74	0.71	0.76
2MCM	112	0.80	0.80	0.85	0.78	0.77	0.81	0.75	0.77	0.82
2NLS	36	0.75	0.66	0.88	0.61	0.32	0.76	0.49	0.47	0.69
2NR7	193	0.75	0.75	0.79	0.74	0.72	0.76	0.71	0.73	0.77
2NUH	104	0.77	0.74	0.85	0.73	0.63	0.81	0.75	0.66	0.80
2O6X	309	0.74	0.75	0.78	0.70	0.73	0.75	0.65	0.73	0.75
2OA2	140	0.63	0.64	0.70	0.55	0.49	0.60	0.60	0.63	0.67
2OHW	257	0.35	0.39	0.48	0.29	0.32	0.35	0.27	0.34	0.38
2OKT	377	0.43	0.37	0.49	0.31	0.36	0.40	0.22	0.33	0.46
2OL9	6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2OLX	4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2PKT	93	0.44	0.39	0.69	0.40	0.35	0.55	0.36	0.36	0.43
2PLT	98	0.66	0.63	0.72	0.57	0.59	0.67	0.52	0.59	0.66
2PMR	83	0.69	0.68	0.80	0.59	0.62	0.68	0.65	0.65	0.69
2POF	428	0.62	0.56	0.66	0.48	0.55	0.60	0.44	0.54	0.63
2PPN	122	0.57	0.61	0.74	0.51	0.59	0.63	0.44	0.57	0.63
2PSF	608	0.43	0.45	0.53	0.41	0.44	0.45	0.37	0.42	0.44
2PTH	193	0.71	0.71	0.77	0.65	0.70	0.73	0.61	0.69	0.72
2Q4N	1208	0.65	0.62	0.68	0.58	0.55	0.59	0.55	0.57	0.61
2Q52	3296	0.65	0.66	0.70	0.62	0.56	0.64	0.63	0.57	0.65
2QJL	107	0.45	0.52	0.63	0.42	0.46	0.50	0.41	0.49	0.51
2R16	185	0.50	0.51	0.66	0.46	0.45	0.51	0.45	0.46	0.52
2R6Q	149	0.71	0.72	0.76	0.66	0.68	0.70	0.62	0.65	0.67
2RB8	93	0.81	0.78	0.84	0.78	0.75	0.80	0.74	0.76	0.81

Table 7.21 (cont'd)

PDB ID	B & W			B			W			
	N	Exp	Lor	Both	Exp	Lor	Both	Exp	Lor	Both
2RE2	249	0.64	0.65	0.70	0.57	0.59	0.61	0.59	0.60	0.63
2RFR	166	0.73	0.66	0.80	0.68	0.57	0.74	0.72	0.59	0.74
2V9V	149	0.60	0.51	0.66	0.53	0.48	0.56	0.55	0.50	0.62
2VE8	515	0.46	0.48	0.55	0.42	0.41	0.44	0.40	0.43	0.47
2VH7	94	0.59	0.54	0.68	0.52	0.49	0.63	0.42	0.49	0.54
2VIM	114	0.38	0.33	0.52	0.29	0.28	0.41	0.24	0.31	0.40
2VPA	217	0.73	0.75	0.78	0.72	0.71	0.73	0.68	0.73	0.74
2VQ4	106	0.56	0.54	0.64	0.43	0.49	0.56	0.35	0.46	0.58
2VY8	162	0.47	0.46	0.58	0.38	0.42	0.46	0.38	0.42	0.49
2VYO	207	0.68	0.70	0.77	0.64	0.66	0.72	0.59	0.68	0.70
2W1V	551	0.69	0.67	0.77	0.63	0.63	0.70	0.56	0.64	0.68
2W2A	350	0.60	0.59	0.65	0.57	0.56	0.59	0.54	0.57	0.60
2W6A	139	0.59	0.59	0.64	0.51	0.52	0.54	0.52	0.56	0.60
2WJ5	110	0.63	0.55	0.79	0.59	0.52	0.68	0.59	0.53	0.64
2WUJ	103	0.69	0.68	0.79	0.62	0.52	0.65	0.67	0.59	0.71
2WW7	161	0.44	0.48	0.60	0.40	0.42	0.50	0.33	0.43	0.49
2WWE	120	0.71	0.71	0.83	0.62	0.62	0.75	0.61	0.58	0.73
2X1Q	240	0.48	0.44	0.54	0.38	0.39	0.46	0.34	0.37	0.47
2X25	167	0.62	0.61	0.73	0.56	0.57	0.64	0.57	0.57	0.64
2X3M	175	0.61	0.61	0.69	0.60	0.55	0.64	0.57	0.57	0.60
2X5Y	185	0.67	0.63	0.71	0.60	0.59	0.64	0.53	0.58	0.69
2X9Z	266	0.50	0.42	0.54	0.37	0.38	0.42	0.38	0.39	0.51
2XHF	310	0.62	0.62	0.67	0.58	0.56	0.60	0.55	0.62	0.63
2Y0T	111	0.69	0.68	0.83	0.60	0.61	0.68	0.56	0.64	0.70
2Y72	183	0.71	0.71	0.78	0.69	0.69	0.72	0.66	0.70	0.71
2Y7L	323	0.68	0.70	0.72	0.66	0.68	0.69	0.58	0.69	0.69
2Y9F	149	0.75	0.72	0.78	0.65	0.69	0.71	0.58	0.70	0.74
2YLB	418	0.55	0.52	0.63	0.46	0.49	0.52	0.34	0.49	0.59
2YNY	326	0.63	0.67	0.75	0.60	0.62	0.63	0.56	0.63	0.66
2ZCM	348	0.42	0.39	0.49	0.34	0.35	0.40	0.24	0.32	0.43
2ZU1	360	0.61	0.61	0.68	0.53	0.58	0.63	0.45	0.58	0.63
3A0M	146	0.74	0.76	0.84	0.68	0.70	0.72	0.61	0.73	0.78
3A7L	128	0.69	0.61	0.78	0.52	0.45	0.59	0.62	0.54	0.67
3AMC	614	0.54	0.53	0.64	0.47	0.50	0.54	0.37	0.51	0.57
3AUB	124	0.36	0.41	0.53	0.31	0.26	0.41	0.32	0.32	0.37
3B5O	249	0.55	0.58	0.66	0.52	0.56	0.63	0.46	0.55	0.57
3BA1	312	0.67	0.66	0.72	0.64	0.65	0.68	0.60	0.65	0.70
3BED	262	0.61	0.55	0.67	0.53	0.53	0.56	0.44	0.53	0.61
3BQX	136	0.52	0.50	0.54	0.47	0.48	0.51	0.41	0.46	0.51
3BZQ	99	0.57	0.62	0.69	0.50	0.55	0.61	0.47	0.55	0.59
3BZZ	103	0.60	0.63	0.68	0.51	0.58	0.61	0.45	0.50	0.59

Table 7.21 (cont'd)

PDB ID	B & W			B			W			
	N	Exp	Lor	Both	Exp	Lor	Both	Exp	Lor	Both
3DRF	567	0.32	0.32	0.38	0.27	0.29	0.33	0.22	0.30	0.34
3DWV	359	0.67	0.63	0.69	0.62	0.62	0.66	0.54	0.62	0.65
3E5T	268	0.55	0.52	0.60	0.51	0.51	0.56	0.38	0.50	0.55
3E7R	40	0.81	0.86	0.96	0.78	0.77	0.81	0.73	0.82	0.88
3EUR	150	0.49	0.46	0.53	0.39	0.43	0.47	0.31	0.42	0.47
3F2Z	148	0.76	0.78	0.84	0.75	0.76	0.78	0.69	0.77	0.78
3F7E	261	0.66	0.65	0.71	0.61	0.64	0.65	0.47	0.63	0.69
3FCN	185	0.60	0.65	0.75	0.56	0.59	0.64	0.54	0.59	0.67
3FE7	89	0.69	0.65	0.76	0.58	0.60	0.67	0.54	0.63	0.70
3FKE	250	0.47	0.42	0.52	0.40	0.36	0.49	0.34	0.36	0.45
3FMY	75	0.71	0.69	0.79	0.66	0.64	0.70	0.66	0.66	0.71
3FOD	48	0.48	0.47	0.82	0.42	0.33	0.55	0.38	0.35	0.48
3FSO	238	0.82	0.82	0.85	0.77	0.74	0.77	0.77	0.81	0.82
3FTD	257	0.60	0.57	0.67	0.49	0.52	0.59	0.41	0.52	0.60
3G1S	418	0.44	0.51	0.68	0.41	0.45	0.51	0.38	0.45	0.49
3GBW	170	0.77	0.78	0.84	0.64	0.74	0.79	0.51	0.71	0.81
3GHJ	129	0.71	0.71	0.81	0.65	0.67	0.72	0.65	0.68	0.72
3HFO	216	0.75	0.72	0.82	0.70	0.63	0.75	0.65	0.69	0.74
3HHP	1314	0.61	0.62	0.68	0.57	0.59	0.62	0.52	0.59	0.63
3HNY	170	0.59	0.56	0.64	0.47	0.52	0.57	0.42	0.49	0.56
3HP4	201	0.60	0.61	0.72	0.57	0.54	0.64	0.43	0.56	0.62
3HWU	155	0.60	0.69	0.81	0.57	0.61	0.63	0.50	0.61	0.68
3HYD	8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3HZ8	200	0.58	0.59	0.66	0.55	0.53	0.56	0.52	0.54	0.58
3I2V	127	0.57	0.58	0.66	0.51	0.53	0.61	0.40	0.48	0.53
3I2Z	140	0.58	0.59	0.65	0.52	0.54	0.56	0.56	0.57	0.61
3I4O	154	0.63	0.64	0.73	0.58	0.59	0.60	0.56	0.63	0.66
3I7M	145	0.58	0.62	0.71	0.53	0.55	0.58	0.49	0.58	0.64
3IHS	173	0.62	0.67	0.74	0.58	0.54	0.60	0.58	0.60	0.62
3IVV	168	0.80	0.80	0.89	0.75	0.76	0.83	0.68	0.74	0.79
3K6Y	227	0.53	0.53	0.60	0.48	0.49	0.52	0.42	0.50	0.55
3KBE	166	0.62	0.61	0.65	0.57	0.60	0.62	0.52	0.60	0.61
3KGK	190	0.79	0.80	0.84	0.77	0.79	0.81	0.68	0.79	0.80
3KZD	94	0.79	0.72	0.83	0.55	0.68	0.77	0.47	0.66	0.78
3L41	219	0.61	0.62	0.71	0.59	0.60	0.66	0.57	0.59	0.67
3LAA	176	0.70	0.66	0.80	0.68	0.56	0.76	0.69	0.60	0.77
3LAX	118	0.81	0.81	0.86	0.80	0.76	0.83	0.77	0.78	0.82
3LG3	846	0.40	0.38	0.41	0.36	0.37	0.40	0.32	0.37	0.41
3LJI	270	0.53	0.53	0.62	0.47	0.52	0.58	0.45	0.52	0.56
3M3P	244	0.47	0.44	0.69	0.40	0.40	0.58	0.25	0.35	0.48
3M8J	178	0.74	0.72	0.75	0.69	0.69	0.73	0.67	0.70	0.73

Table 7.21 (cont'd)

PDB ID	B & W			B			W			
	N	Exp	Lor	Both	Exp	Lor	Both	Exp	Lor	Both
3M9J	250	0.57	0.56	0.59	0.53	0.54	0.56	0.39	0.53	0.56
3M9Q	190	0.53	0.52	0.59	0.50	0.51	0.53	0.46	0.50	0.51
3MAB	180	0.57	0.56	0.62	0.52	0.47	0.55	0.56	0.51	0.56
3MD4	13	1.00	1.00	1.00	0.91	0.94	1.00	0.93	0.99	1.00
3MD5	14	1.00	1.00	1.00	0.98	0.93	1.00	0.94	0.92	1.00
3MEA	170	0.58	0.58	0.68	0.57	0.57	0.64	0.48	0.57	0.59
3MGN	277	0.33	0.32	0.47	0.26	0.28	0.30	0.16	0.29	0.39
3MRE	446	0.40	0.38	0.45	0.32	0.36	0.40	0.24	0.35	0.41
3N11	325	0.43	0.45	0.51	0.42	0.44	0.45	0.38	0.44	0.45
3NE0	208	0.77	0.79	0.84	0.75	0.70	0.77	0.70	0.76	0.82
3NGG	97	0.80	0.81	0.85	0.72	0.74	0.78	0.74	0.76	0.80
3NPV	500	0.44	0.44	0.50	0.40	0.42	0.44	0.36	0.43	0.47
3NVG	6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3NZL	70	0.68	0.61	0.84	0.53	0.49	0.66	0.59	0.55	0.67
3O0P	197	0.62	0.64	0.71	0.59	0.62	0.64	0.53	0.62	0.64
3O5P	147	0.64	0.60	0.71	0.55	0.57	0.60	0.53	0.56	0.64
3OBQ	150	0.59	0.59	0.66	0.46	0.49	0.58	0.53	0.56	0.58
3OQY	236	0.71	0.66	0.73	0.63	0.64	0.70	0.60	0.64	0.72
3P6J	145	0.75	0.73	0.81	0.69	0.71	0.73	0.61	0.71	0.75
3PD7	216	0.65	0.66	0.72	0.62	0.60	0.65	0.60	0.61	0.65
3PES	166	0.70	0.72	0.79	0.58	0.63	0.70	0.52	0.60	0.66
3PID	387	0.50	0.49	0.56	0.44	0.48	0.53	0.37	0.46	0.51
3PIW	161	0.66	0.67	0.78	0.60	0.63	0.70	0.56	0.63	0.72
3PKV	229	0.50	0.52	0.63	0.43	0.48	0.53	0.35	0.50	0.57
3PSM	94	0.83	0.78	0.88	0.79	0.77	0.83	0.68	0.76	0.79
3PTL	289	0.50	0.50	0.53	0.49	0.49	0.50	0.43	0.49	0.50
3PVE	363	0.45	0.45	0.59	0.37	0.39	0.44	0.41	0.42	0.45
3PZ9	357	0.51	0.45	0.57	0.36	0.38	0.42	0.34	0.39	0.50
3PZZ	12	1.00	1.00	1.00	0.95	0.90	1.00	0.94	0.80	1.00
3Q2X	6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3Q6L	131	0.39	0.44	0.56	0.33	0.31	0.37	0.34	0.37	0.42
3QDS	284	0.63	0.62	0.69	0.59	0.59	0.65	0.51	0.59	0.64
3QPA	212	0.68	0.66	0.78	0.45	0.45	0.47	0.59	0.59	0.65
3R6D	222	0.65	0.66	0.73	0.62	0.63	0.65	0.53	0.64	0.69
3R87	148	0.48	0.47	0.55	0.41	0.44	0.48	0.40	0.45	0.47
3RQ9	165	0.51	0.47	0.61	0.41	0.44	0.52	0.39	0.45	0.56
3RY0	128	0.44	0.45	0.54	0.40	0.40	0.47	0.41	0.42	0.47
3RZY	151	0.65	0.65	0.84	0.59	0.54	0.65	0.57	0.51	0.59
3S0A	132	0.39	0.43	0.52	0.33	0.34	0.38	0.32	0.31	0.37
3SD2	100	0.65	0.67	0.77	0.64	0.63	0.69	0.56	0.63	0.67
3SEB	238	0.63	0.66	0.77	0.62	0.61	0.68	0.61	0.62	0.67

Table 7.21 (cont'd)

PDB ID	B & W			B			W			
	N	Exp	Lor	Both	Exp	Lor	Both	Exp	Lor	Both
3SED	126	0.39	0.45	0.55	0.28	0.29	0.38	0.33	0.33	0.40
3SO6	157	0.67	0.71	0.78	0.63	0.69	0.73	0.55	0.64	0.70
3SR3	657	0.45	0.44	0.48	0.43	0.41	0.45	0.39	0.43	0.44
3SUK	254	0.53	0.54	0.64	0.46	0.48	0.54	0.47	0.49	0.57
3SZH	753	0.53	0.53	0.57	0.51	0.51	0.52	0.45	0.52	0.53
3T0H	209	0.76	0.73	0.78	0.72	0.69	0.74	0.68	0.71	0.76
3T3K	122	0.66	0.66	0.72	0.55	0.62	0.68	0.48	0.60	0.68
3T47	145	0.54	0.54	0.78	0.45	0.45	0.62	0.43	0.47	0.54
3TDN	359	0.47	0.43	0.53	0.43	0.42	0.44	0.38	0.43	0.49
3TOW	155	0.66	0.65	0.74	0.58	0.61	0.66	0.53	0.60	0.65
3TUA	226	0.57	0.55	0.63	0.52	0.50	0.55	0.45	0.52	0.54
3TYS	78	0.78	0.58	0.86	0.67	0.48	0.73	0.70	0.46	0.75
3U6G	276	0.44	0.39	0.54	0.39	0.37	0.45	0.27	0.35	0.48
3U97	85	0.78	0.78	0.84	0.77	0.73	0.80	0.77	0.76	0.80
3UCI	72	0.67	0.64	0.72	0.48	0.53	0.57	0.55	0.56	0.63
3UR8	637	0.52	0.53	0.60	0.49	0.51	0.55	0.45	0.52	0.53
3US6	159	0.60	0.56	0.67	0.55	0.49	0.62	0.53	0.46	0.59
3V1A	59	0.74	0.57	0.95	0.51	0.53	0.77	0.39	0.46	0.68
3V75	294	0.50	0.49	0.57	0.48	0.46	0.53	0.47	0.47	0.53
3VN0	193	0.87	0.88	0.90	0.86	0.87	0.88	0.79	0.88	0.89
3VOR	219	0.64	0.58	0.70	0.56	0.52	0.63	0.53	0.55	0.63
3VUB	101	0.65	0.60	0.71	0.60	0.56	0.61	0.61	0.57	0.64
3VVV	112	0.64	0.64	0.79	0.55	0.48	0.65	0.57	0.49	0.58
3VZ9	163	0.65	0.64	0.70	0.60	0.55	0.63	0.60	0.60	0.67
3W4Q	826	0.61	0.60	0.68	0.56	0.59	0.61	0.47	0.60	0.64
3ZBD	213	0.36	0.47	0.74	0.24	0.28	0.34	0.25	0.31	0.36
3ZIT	157	0.51	0.47	0.59	0.36	0.39	0.47	0.47	0.41	0.52
3ZRX	241	0.56	0.56	0.63	0.49	0.52	0.53	0.46	0.52	0.56
3ZSL	165	0.39	0.39	0.54	0.28	0.22	0.40	0.31	0.24	0.37
3ZZP	74	0.40	0.30	0.47	0.19	0.27	0.31	0.12	0.22	0.40
3ZZY	226	0.65	0.67	0.69	0.63	0.63	0.64	0.59	0.63	0.64
4A02	169	0.61	0.56	0.66	0.49	0.52	0.57	0.31	0.51	0.60
4ACJ	182	0.55	0.59	0.75	0.55	0.58	0.61	0.51	0.59	0.60
4AE7	189	0.69	0.67	0.74	0.63	0.61	0.65	0.63	0.65	0.69
4AM1	359	0.57	0.54	0.59	0.53	0.52	0.53	0.46	0.53	0.55
4ANN	210	0.50	0.48	0.57	0.42	0.43	0.48	0.36	0.42	0.47
4AVR	189	0.57	0.57	0.70	0.53	0.51	0.59	0.49	0.53	0.57
4AXY	56	0.55	0.60	0.76	0.47	0.48	0.63	0.47	0.50	0.62
4B6G	559	0.70	0.71	0.75	0.67	0.69	0.72	0.60	0.69	0.73
4B9G	292	0.81	0.82	0.85	0.78	0.80	0.81	0.71	0.82	0.83
4DD5	412	0.60	0.63	0.71	0.57	0.59	0.63	0.51	0.61	0.66

Table 7.21 (cont'd)

PDB ID	B & W			B			W			
	N	Exp	Lor	Both	Exp	Lor	Both	Exp	Lor	Both
4DKN	423	0.59	0.58	0.63	0.52	0.54	0.56	0.42	0.55	0.61
4DND	93	0.75	0.66	0.82	0.67	0.64	0.75	0.61	0.64	0.74
4DPZ	113	0.68	0.70	0.79	0.65	0.64	0.67	0.62	0.64	0.69
4DQ7	338	0.45	0.46	0.51	0.37	0.44	0.49	0.29	0.40	0.46
4DT4	170	0.76	0.74	0.78	0.70	0.68	0.72	0.70	0.70	0.73
4EK3	313	0.58	0.63	0.65	0.55	0.56	0.58	0.53	0.59	0.60
4ERY	318	0.61	0.60	0.67	0.59	0.59	0.64	0.52	0.59	0.65
4ES1	96	0.76	0.77	0.86	0.69	0.73	0.78	0.57	0.74	0.83
4EUG	225	0.61	0.61	0.67	0.54	0.60	0.62	0.51	0.58	0.62
4F01	459	0.38	0.37	0.47	0.32	0.34	0.37	0.22	0.34	0.39
4F3J	143	0.57	0.63	0.66	0.52	0.59	0.61	0.47	0.58	0.60
4FR9	145	0.65	0.62	0.78	0.63	0.58	0.70	0.58	0.57	0.64
4G14	5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4G2E	155	0.75	0.64	0.85	0.59	0.61	0.74	0.68	0.61	0.80
4G5X	584	0.71	0.69	0.80	0.69	0.64	0.74	0.64	0.67	0.72
4G6C	676	0.43	0.44	0.50	0.40	0.44	0.46	0.24	0.43	0.45
4G7X	216	0.53	0.47	0.61	0.41	0.31	0.47	0.51	0.37	0.53
4GA2	183	0.55	0.56	0.70	0.52	0.53	0.57	0.49	0.53	0.60
4GMQ	94	0.73	0.77	0.84	0.68	0.66	0.72	0.67	0.63	0.72
4GS3	90	0.65	0.68	0.74	0.60	0.64	0.68	0.51	0.66	0.70
4H4J	278	0.67	0.67	0.82	0.63	0.64	0.75	0.57	0.66	0.69
4H89	175	0.39	0.50	0.67	0.33	0.37	0.39	0.35	0.40	0.42
4HDE	167	0.63	0.55	0.75	0.59	0.52	0.69	0.59	0.51	0.67
4HJP	308	0.62	0.61	0.65	0.57	0.55	0.59	0.58	0.58	0.62
4HWM	129	0.69	0.66	0.71	0.66	0.60	0.68	0.68	0.63	0.70
4IL7	99	0.63	0.63	0.65	0.60	0.59	0.62	0.57	0.61	0.62
4J11	377	0.66	0.63	0.68	0.62	0.61	0.63	0.63	0.61	0.66
4J5O	268	0.77	0.76	0.82	0.71	0.62	0.77	0.75	0.66	0.77
4J5Q	162	0.65	0.63	0.75	0.57	0.56	0.66	0.59	0.57	0.64
4J78	305	0.48	0.48	0.56	0.43	0.44	0.50	0.38	0.47	0.53
4JG2	202	0.63	0.63	0.74	0.61	0.61	0.64	0.58	0.60	0.63
4JVU	207	0.67	0.64	0.75	0.57	0.58	0.66	0.59	0.60	0.67
4JYP	550	0.59	0.60	0.69	0.52	0.57	0.61	0.38	0.58	0.61
4KEF	145	0.52	0.49	0.65	0.40	0.42	0.49	0.27	0.45	0.56
5CYT	103	0.53	0.52	0.65	0.49	0.46	0.54	0.43	0.48	0.50
6RXN	45	0.74	0.63	0.86	0.59	0.48	0.76	0.49	0.49	0.76

Table 7.22: Persistent homology based Pearson correlation coefficients for cross protein C_α atom blind B factor prediction obtained by boosted gradient (GBT), convolutional neural network (CNN), and consensus method (CON) for the Superset.

PDB ID	N	GBT	CNN	CON	PDB ID	N	GBT	CNN	CON
1ABA	87	0.73	0.71	0.74	2X5Y	185	0.76	0.68	0.76
1AHO	66	0.66	0.66	0.7	2X9Z	266	0.49	0.52	0.52
1AIE	31	0.75	0.7	0.78	2XHF	310	0.58	0.57	0.58
1AKG	16	0.27	0.32	0.29	2Y0T	111	0.71	0.71	0.74
1ATG	231	0.55	0.51	0.55	2Y72	183	0.65	0.71	0.69
1BGF	124	0.61	0.58	0.62	2Y7L	323	0.66	0.66	0.68
1BX7	51	0.74	0.74	0.76	2Y9F	149	0.74	0.75	0.76
1BYI	238	0.61	0.5	0.6	2YLB	418	0.67	0.66	0.7
1CCR	109	0.55	0.6	0.59	2YNY	326	0.65	0.71	0.69
1CYO	88	0.64	0.7	0.68	2ZCM	348	0.33	0.38	0.36
1DF4	57	0.85	0.85	0.88	2ZU1	360	0.66	0.66	0.68
1E5K	188	0.74	0.72	0.74	3A0M	146	0.53	0.6	0.59
1ES5	260	0.65	0.62	0.66	3A7L	128	0.44	0.61	0.53
1ETL	12	0.37	0.82	0.55	3AMC	614	0.68	0.64	0.69
1ETM	12	0.37	0.63	0.43	3AUB	124	0.5	0.5	0.55
1ETN	12	0.07	0.48	0.13	3B5O	249	0.49	0.55	0.52
1EW4	106	0.59	0.6	0.61	3BA1	312	0.62	0.59	0.63
1F8R	1932	0.52	0.54	0.54	3BED	262	0.45	0.53	0.5
1FF4	65	0.61	0.66	0.64	3BQX	136	0.56	0.55	0.58
1FK5	93	0.59	0.6	0.61	3BZQ	99	0.45	0.53	0.49
1GCO	1044	0.47	0.47	0.5	3BZZ	103	0.38	0.51	0.44
1GK7	39	0.77	0.9	0.82	3DRF	567	0.51	0.45	0.52
1GVD	56	0.71	0.55	0.69	3DWV	359	0.63	0.55	0.63
1GXU	89	0.67	0.68	0.69	3E5T	268	0.44	0.48	0.46
1H6V	2927	0.26	0.34	0.34	3E7R	40	0.72	0.66	0.77
1HJE	13	0.84	0.75	0.9	3EUR	150	0.36	0.42	0.38
1I71	83	0.53	0.58	0.56	3F2Z	148	0.73	0.76	0.75
1IDP	441	0.62	0.6	0.63	3F7E	261	0.65	0.69	0.68
1IFR	113	0.7	0.64	0.7	3FCN	185	0.63	0.65	0.66
1K8U	87	0.57	0.6	0.59	3FE7	89	0.52	0.55	0.54
1KMM	1499	0.64	0.51	0.63	3FKE	250	0.51	0.51	0.54
1KNG	144	0.5	0.52	0.51	3FMY	75	0.65	0.67	0.68
1KR4	107	0.56	0.71	0.63	3FOD	48	0.45	0.57	0.54
1KYC	15	0.62	0.69	0.66	3FSO	238	0.72	0.75	0.74
1LR7	73	0.62	0.61	0.64	3FTD	257	0.64	0.68	0.67
1MF7	194	0.65	0.66	0.67	3G1S	418	0.6	0.57	0.61
1N7E	95	0.63	0.58	0.65	3GBW	170	0.74	0.74	0.75
1NKD	59	0.7	0.7	0.72	3GHJ	129	0.58	0.56	0.59

Table 7.22 (cont'd)

PDB ID	N	GBT	CNN	CON	PDB ID	N	GBT	CNN	CON
1NLS	238	0.55	0.57	0.57	3HFO	216	0.51	0.57	0.54
1NNX	93	0.78	0.79	0.8	3HHP	1314	0.61	0.65	0.65
1NOA	113	0.55	0.53	0.56	3HNY	170	0.61	0.6	0.62
1NOT	13	0.69	0.96	0.8	3HP4	201	0.56	0.58	0.58
1O06	22	0.94	0.93	0.95	3HWU	155	0.58	0.65	0.62
1O08	221	0.49	0.47	0.49	3HYD	8	0.99	0.74	0.99
1OPD	85	0.42	0.34	0.41	3HZ8	200	0.45	0.54	0.48
1P9I	29	0.73	0.73	0.74	3I2V	127	0.44	0.52	0.48
1PEF	18	0.79	0.82	0.82	3I2Z	140	0.6	0.6	0.6
1PEN	16	0.36	0.74	0.44	3I4O	154	0.62	0.72	0.66
1PMY	123	0.59	0.7	0.65	3I7M	145	0.44	0.57	0.49
1PZ4	113	0.72	0.8	0.77	3IHS	173	0.61	0.62	0.64
1Q9B	44	0.59	0.85	0.67	3IVV	168	0.83	0.82	0.84
1QAU	112	0.51	0.59	0.57	3K6Y	227	0.56	0.57	0.58
1QKI	3912	0.34	0.45	0.38	3KBE	166	0.56	0.64	0.6
1QTO	122	0.53	0.48	0.54	3K GK	190	0.76	0.8	0.78
1R29	122	0.56	0.59	0.59	3KZD	94	0.55	0.67	0.6
1R7J	90	0.71	0.77	0.75	3L4I	219	0.61	0.64	0.64
1RJU	36	0.6	0.46	0.58	3LAA	176	0.35	0.49	0.42
1RRO	108	0.4	0.45	0.43	3LAX	118	0.74	0.69	0.74
1SAU	123	0.54	0.66	0.59	3LG3	846	0.45	0.51	0.5
1TGR	111	0.66	0.69	0.69	3LJI	270	0.57	0.55	0.58
1TZV	157	0.74	0.77	0.76	3M3P	244	0.53	0.59	0.57
1U06	55	0.44	0.4	0.45	3M8J	178	0.72	0.71	0.74
1U7I	259	0.71	0.74	0.74	3M9J	250	0.56	0.52	0.56
1U9C	220	0.57	0.59	0.59	3M9Q	190	0.4	0.48	0.45
1UHA	82	0.71	0.74	0.73	3MAB	180	0.63	0.63	0.65
1UKU	102	0.75	0.76	0.77	3MD4	13	0.88	0.96	0.96
1ULR	87	0.54	0.53	0.56	3MEA	170	0.62	0.63	0.63
1UOY	64	0.72	0.7	0.76	3MGN	277	0.08	0.09	0.09
1USE	47	0.05	0.32	0.12	3MRE	446	0.54	0.54	0.57
1USM	77	0.73	0.72	0.75	3N11	325	0.51	0.47	0.52
1UTG	70	0.62	0.64	0.66	3NE0	208	0.67	0.73	0.71
1V05	96	0.6	0.64	0.63	3NGG	97	0.72	0.75	0.75
1V70	105	0.63	0.62	0.64	3NPV	500	0.51	0.5	0.54
1VRZ	13	0.54	0.34	0.54	3NVG	6	0.51	0.63	0.71
1W2L	97	0.43	0.5	0.47	3NZL	70	0.56	0.58	0.57
1WBE	206	0.6	0.56	0.6	3O0P	197	0.68	0.72	0.71
1WHI	122	0.59	0.56	0.6	3O5P	147	0.6	0.59	0.61
1WLY	322	0.64	0.62	0.66	3OBQ	150	0.59	0.57	0.59
1WPA	107	0.65	0.65	0.67	3OQY	236	0.66	0.59	0.66
1X3O	80	0.41	0.43	0.44	3P6J	145	0.66	0.72	0.69

Table 7.22 (cont'd)

PDB ID	N	GBT	CNN	CON	PDB ID	N	GBT	CNN	CON
1XY1	16	0.82	0.75	0.83	3PD7	216	0.68	0.7	0.71
1XY2	8	0.79	0.82	0.81	3PES	166	0.56	0.54	0.57
1Y6X	86	0.5	0.46	0.51	3PID	387	0.48	0.3	0.45
1YJO	6	0.7	-0.06	0.57	3PIW	161	0.72	0.77	0.75
1YZM	46	0.69	0.64	0.7	3PKV	229	0.52	0.51	0.53
1Z21	96	0.68	0.65	0.69	3PSM	94	0.8	0.77	0.82
1ZCE	139	0.7	0.74	0.73	3PTL	289	0.53	0.55	0.55
1ZVA	75	0.7	0.7	0.71	3PVE	363	0.61	0.61	0.63
2A50	469	0.6	0.54	0.6	3PZ9	357	0.61	0.58	0.63
2AGK	233	0.67	0.63	0.67	3PZZ	12	0.94	0.85	0.93
2AH1	939	0.48	0.55	0.54	3Q2X	6	0.95	0.72	0.93
2B0A	191	0.62	0.59	0.63	3Q6L	131	0.47	0.53	0.52
2BCM	415	0.5	0.51	0.52	3QDS	284	0.62	0.62	0.63
2BF9	35	0.48	0.79	0.58	3QPA	212	0.55	0.67	0.59
2BRF	103	0.72	0.77	0.75	3R6D	222	0.65	0.74	0.69
2C71	225	0.57	0.6	0.6	3R87	148	0.47	0.45	0.48
2CE0	109	0.6	0.66	0.64	3RQ9	165	0.46	0.4	0.46
2CG7	110	0.3	0.32	0.32	3RY0	128	0.41	0.49	0.46
2COV	534	0.74	0.72	0.75	3RZY	151	0.65	0.62	0.66
2CWS	235	0.61	0.47	0.6	3S0A	132	0.53	0.49	0.54
2D5W	1214	0.54	0.64	0.59	3SD2	100	0.56	0.56	0.57
2DKO	253	0.78	0.78	0.8	3SEB	238	0.63	0.6	0.63
2DPL	565	0.41	0.36	0.42	3SED	126	0.53	0.52	0.55
2DSX	52	0.34	0.34	0.36	3SO6	157	0.65	0.65	0.66
2OCT	439	0.64	0.67	0.67	3SR3	657	0.5	0.46	0.5
2E3H	81	0.65	0.68	0.67	3SUK	254	0.58	0.59	0.6
2EAQ	89	0.57	0.63	0.61	3SZH	753	0.69	0.67	0.71
2EHP	246	0.66	0.62	0.67	3T0H	209	0.71	0.7	0.73
2EHS	75	0.62	0.67	0.65	3T3K	122	0.76	0.76	0.78
2ERW	53	0.12	0.24	0.16	3T47	145	0.51	0.62	0.57
2ETX	390	0.49	0.48	0.51	3TDN	359	0.47	0.49	0.49
2FB6	129	0.73	0.75	0.75	3TOW	155	0.61	0.63	0.63
2FG1	176	0.57	0.61	0.59	3TUA	226	0.62	0.56	0.63
2FN9	560	0.57	0.54	0.58	3TYS	78	0.66	0.74	0.72
2FQ3	85	0.77	0.82	0.81	3U6G	276	0.53	0.46	0.52
2G69	99	0.62	0.5	0.6	3U97	85	0.67	0.72	0.71
2G7O	68	0.72	0.86	0.8	3UCI	72	0.42	0.42	0.43
2G7S	206	0.55	0.58	0.58	3UR8	637	0.64	0.6	0.64
2GKG	150	0.56	0.64	0.59	3US6	159	0.61	0.63	0.64
2GOM	121	0.69	0.59	0.69	3V1A	59	0.57	0.27	0.55
2GXG	140	0.65	0.67	0.68	3V75	294	0.49	0.56	0.53
2GZQ	203	0.34	0.4	0.37	3VN0	193	0.85	0.85	0.86

Table 7.22 (cont'd)

PDB ID	N	GBT	CNN	CON	PDB ID	N	GBT	CNN	CON
2HQK	232	0.77	0.77	0.78	3VOR	219	0.47	0.48	0.48
2HYK	237	0.65	0.63	0.65	3VUB	101	0.59	0.55	0.59
2I24	113	0.44	0.46	0.46	3VVV	112	0.56	0.57	0.57
2I49	399	0.65	0.61	0.66	3VZ9	163	0.72	0.64	0.72
2IBL	108	0.65	0.66	0.67	3W4Q	826	0.65	0.6	0.66
2IGD	61	0.57	0.56	0.58	3ZBD	213	0.55	0.49	0.55
2IMF	203	0.53	0.58	0.56	3ZIT	157	0.52	0.42	0.5
2IP6	87	0.6	0.66	0.63	3ZRX	241	0.54	0.6	0.58
2IVY	89	0.51	0.45	0.51	3ZSL	165	0.49	0.57	0.53
2J32	244	0.75	0.79	0.79	3ZZP	74	0.38	0.48	0.42
2J9W	203	0.64	0.58	0.64	3ZZY	226	0.65	0.65	0.68
2JKU	38	0.57	0.71	0.66	4A02	169	0.59	0.65	0.62
2JLI	112	0.62	0.68	0.65	4ACJ	182	0.62	0.66	0.64
2JLJ	121	0.71	0.71	0.74	4AE7	189	0.65	0.7	0.68
2MCM	112	0.71	0.77	0.75	4AM1	359	0.54	0.52	0.55
2NLS	36	0.23	0.47	0.29	4ANN	210	0.44	0.43	0.45
2NR7	193	0.78	0.76	0.79	4AVR	189	0.56	0.53	0.56
2NUH	104	0.72	0.56	0.7	4AXY	56	0.59	0.65	0.62
2O6X	309	0.76	0.76	0.78	4B6G	559	0.69	0.68	0.71
2OA2	140	0.54	0.55	0.56	4B9G	292	0.74	0.74	0.76
2OHW	257	0.56	0.46	0.54	4DD5	412	0.61	0.62	0.63
2OKT	377	0.42	0.42	0.43	4DKN	423	0.66	0.64	0.68
2OL9	6	0.94	0.85	0.94	4DND	93	0.62	0.67	0.65
2PKT	93	0.01	-0.04	-0.01	4DPZ	113	0.7	0.74	0.72
2PLT	98	0.52	0.53	0.54	4DQ7	338	0.55	0.6	0.57
2PMR	83	0.6	0.63	0.63	4DT4	170	0.67	0.69	0.69
2POF	428	0.62	0.6	0.66	4EK3	313	0.6	0.58	0.61
2PPN	122	0.64	0.54	0.63	4ERY	318	0.57	0.59	0.59
2PSF	608	0.42	0.42	0.43	4ES1	96	0.69	0.69	0.71
2PTH	193	0.69	0.7	0.71	4EUG	225	0.56	0.55	0.58
2Q4N	1208	0.44	0.43	0.45	4F01	459	0.35	0.26	0.33
2Q52	3296	0.55	0.28	0.52	4F3J	143	0.58	0.63	0.62
2QJL	107	0.54	0.57	0.56	4FR9	145	0.6	0.56	0.61
2R16	185	0.44	0.49	0.46	4G14	5	-0.28	0.45	0.04
2R6Q	149	0.63	0.62	0.65	4G2E	155	0.75	0.72	0.76
2RB8	93	0.67	0.7	0.7	4G5X	584	0.71	0.73	0.74
2RE2	249	0.65	0.66	0.68	4G6C	676	0.56	0.54	0.58
2RFR	166	0.61	0.69	0.66	4G7X	216	0.45	0.4	0.45
2V9V	149	0.53	0.52	0.54	4GA2	183	0.61	0.53	0.61
2VE8	515	0.55	0.55	0.58	4GMQ	94	0.76	0.67	0.76
2VH7	94	0.75	0.56	0.73	4GS3	90	0.61	0.56	0.61
2VIM	114	0.44	0.47	0.47	4H4J	278	0.75	0.74	0.77

Table 7.22 (cont'd)

PDB ID	N	GBT	CNN	CON	PDB ID	N	GBT	CNN	CON
2VPA	217	0.66	0.75	0.71	4H89	175	0.53	0.58	0.56
2VQ4	106	0.7	0.75	0.72	4HDE	167	0.66	0.72	0.7
2VY8	162	0.77	0.68	0.76	4HJP	308	0.68	0.6	0.67
2VYO	207	0.6	0.63	0.63	4HWM	129	0.54	0.6	0.57
2W1V	551	0.64	0.69	0.66	4IL7	99	0.55	0.55	0.56
2W2A	350	0.59	0.6	0.61	4J11	377	0.58	0.49	0.58
2W6A	139	0.71	0.69	0.72	4J5O	268	0.67	0.68	0.69
2WJ5	110	0.45	0.53	0.48	4J5Q	162	0.72	0.74	0.74
2WUJ	103	0.35	0.54	0.45	4J78	305	0.63	0.6	0.64
2WW7	161	0.36	0.35	0.37	4JG2	202	0.72	0.72	0.73
2WWE	120	0.49	0.55	0.53	4JVU	207	0.7	0.7	0.72
2X1Q	240	0.44	0.5	0.47	4JYP	550	0.59	0.67	0.65
2X25	167	0.5	0.57	0.55	4KEF	145	0.48	0.53	0.51
2X3M	175	0.64	0.65	0.65	5CYT	103	0.39	0.34	0.39
					6RXN	45	0.59	0.6	0.61

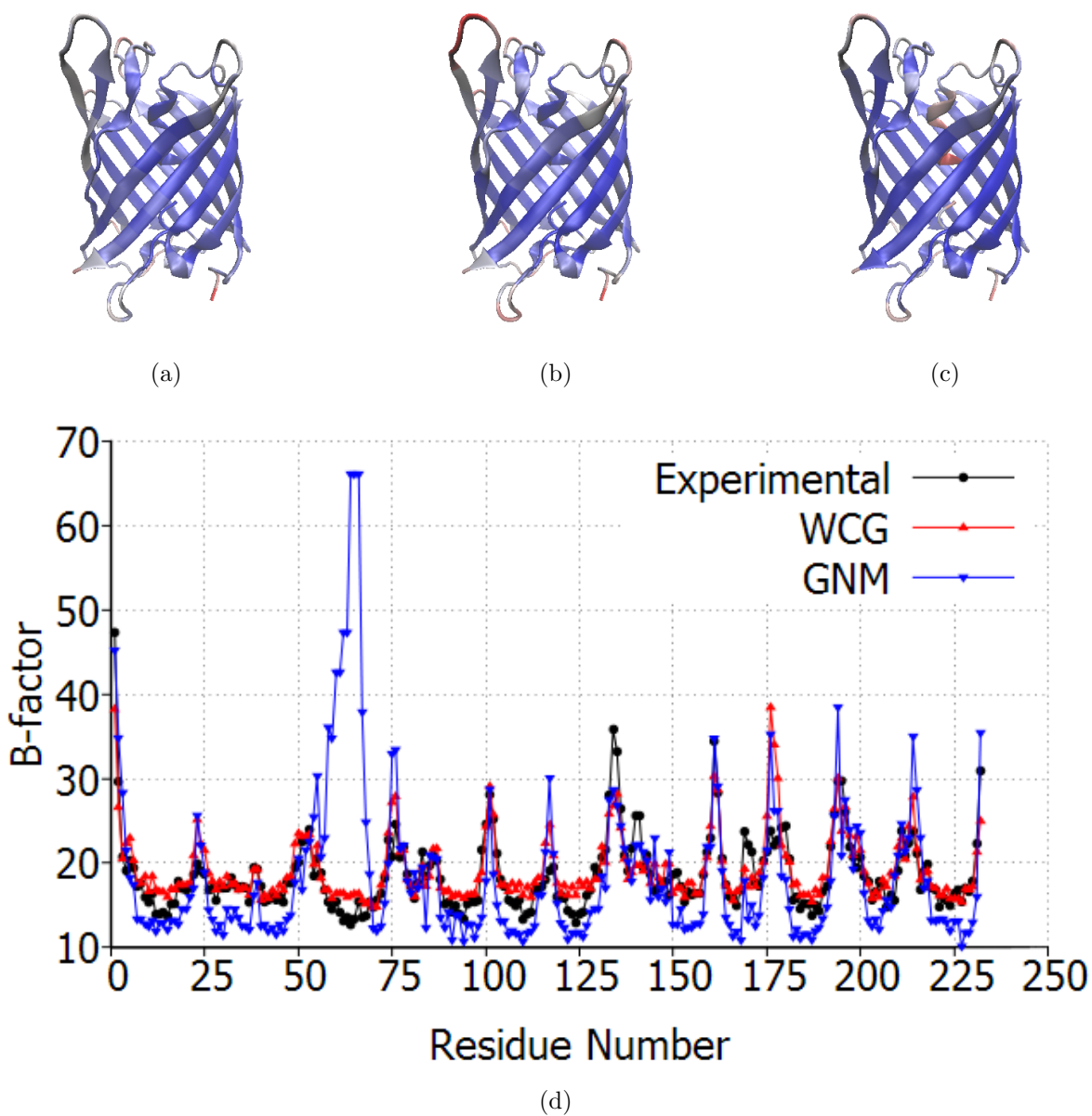


Figure 7.6: A visual comparison of experimental B factors (a), WCG predicted B factors (b), and GNM predicted B factors (c) for the engineered cyan fluorescent protein, mTFP1 (PDB ID:2HQB). (d) The experimental (Exp) and predicted B factor values plotted per residue for PDB ID 2HQB. The GNM is for the GNM method with a cutoff distance of 7 Å. WCG is parametrized using CC, CN, CO kernels of the exponential type with fixed parameters $\kappa = 1$, and $\eta = 3$ Å. Figure originally published in Bramer *et al* [1].

Table 7.1: Correlation coefficients for B factor prediction obtained by optimal FRI (opFRI), parameter free FRI (pfFRI), and Gaussian normal mode (GNM) for small-size structures. Results for opFRI, pfFRI are taken from *Opron et al* [3]. GNM and NMA values are taken from the coarse-grained (C_α) results reported in *Park et al* [4]. MWCG results are parameter free and use all C, N, and O to predict C_α . MWCG results originally published in *Bramer et al* [1].

PDB ID	N	MWCG	opFRI	pfFRI	GNM	NMA
1AIE	31	0.969	0.588	0.416	0.155	0.712
1AKG	16	0.945	0.373	0.35	0.185	-0.229
1BX7	51	0.896	0.726	0.623	0.706	0.868
1ETL	12	0.932	0.71	0.609	0.628	0.355
1ETM	12	0.941	0.544	0.393	0.432	0.027
1ETN	12	0.949	0.089	0.023	-0.274	-0.573
1FF4	65	0.933	0.718	0.613	0.674	0.555
1GK7	39	0.984	0.845	0.773	0.821	0.822
1GVD	52	0.849	0.781	0.732	0.591	0.570
1HJE	13	0.931	0.811	0.686	0.616	0.562
1KYC	15	0.971	0.796	0.763	0.754	0.784
1NOT	13	0.937	0.746	0.622	0.523	0.567
1O06	20	0.988	0.91	0.874	0.844	0.900
1OB4	16	1.000	0.776	0.763	0.750	0.930
1OB7	16	1.000	0.737	0.545	0.652	0.952
1P9I	29	0.841	0.754	0.742	0.625	0.603
1PEF	18	0.989	0.888	0.826	0.808	0.888
1PEN	16	0.957	0.516	0.465	0.270	0.056
1Q9B	43	0.957	0.746	0.726	0.656	0.646
1RJU	36	0.805	0.517	0.447	0.431	0.235
1U06	55	0.774	0.474	0.429	0.434	0.377
1UOY	64	0.769	0.713	0.653	0.671	0.628
1USE	40	0.960	0.438	0.146	-0.142	-0.399
1VRZ	21	0.995	0.792	0.695	0.677	-0.203
1XY2	8	1.000	0.619	0.57	0.562	0.458
1YJO	6	1.000	0.375	0.333	0.434	0.445
1YZM	46	0.970	0.842	0.834	0.901	0.939
2DSX	52	0.704	0.337	0.333	0.127	0.433
2JKU	35	0.926	0.805	0.695	0.656	0.850
2NLS	36	0.937	0.605	0.559	0.530	0.088
2OL9	6	1.000	0.909	0.904	0.689	0.886
2OLX	4	1.000	0.917	0.888	0.885	0.776
6RXN	45	0.583	0.614	0.574	0.594	0.304

Table 7.2: Correlation coefficients for B factor prediction obtained by optimal FRI (opFRI), parameter free FRI (pfFRI) and Gaussian normal mode (GNM) for medium-size structures. Results for opFRI, pfFRI are taken from *Opron et al* [3]. GNM and NMA values are taken from the coarse-grained (C_α) results reported in *Park et al* [4]. MWCG results are parameter free and use all C, N, and O to predict C_α . MWCG results originally published in *Bramer et al* [1].

PDB ID	N	MWCG	opFRI	pfFRI	GNM	NMA
1ABA	87	0.855	0.727	0.698	0.613	0.057
1CYO	88	0.860	0.751	0.702	0.741	0.774
1FK5	93	0.648	0.590	0.568	0.485	0.362
1GXU	88	0.901	0.748	0.634	0.421	0.581
1I71	83	0.798	0.549	0.516	0.549	0.380
1LR7	73	0.929	0.679	0.657	0.620	0.795
1N7E	95	0.812	0.651	0.609	0.497	0.385
1NNX	93	0.834	0.795	0.789	0.631	0.517
1NOA	113	0.808	0.622	0.604	0.615	0.485
1OPD	85	0.607	0.555	0.409	0.398	0.796
1QAU	112	0.786	0.678	0.672	0.620	0.533
1R7J	90	0.859	0.789	0.621	0.368	0.078
1UHA	83	0.838	0.726	0.665	0.638	0.308
1ULR	87	0.718	0.639	0.594	0.495	0.223
1USM	77	0.819	0.832	0.809	0.798	0.780
1V05	96	0.841	0.629	0.599	0.632	0.389
1W2L	97	0.747	0.691	0.564	0.397	0.432
1X3O	80	0.787	0.600	0.559	0.654	0.453
1Z21	96	0.725	0.662	0.638	0.433	0.289
1ZVA	75	0.911	0.756	0.579	0.690	0.579
2BF9	36	0.714	0.606	0.554	0.680	0.521
2BRF	100	0.873	0.795	0.764	0.710	0.535
2CE0	99	0.824	0.706	0.598	0.529	0.628
2E3H	81	0.794	0.692	0.682	0.605	0.632
2EAQ	89	0.817	0.753	0.690	0.695	0.688
2EHS	75	0.805	0.720	0.713	0.747	0.565
2FQ3	85	0.844	0.719	0.692	0.348	0.508
2IP6	87	0.841	0.654	0.578	0.572	0.826
2MCM	113	0.867	0.789	0.713	0.639	0.643
2NUH	104	0.922	0.835	0.691	0.771	0.685
2PKT	93	0.762	0.162	0.003	-0.193	-0.165
2PLT	99	0.635	0.508	0.484	0.509	0.187
2QJL	99	0.611	0.594	0.584	0.594	0.497
2RB8	93	0.840	0.727	0.614	0.517	0.485
3BZQ	99	0.848	0.532	0.516	0.466	0.351
5CYT	103	0.548	0.441	0.421	0.331	0.102

Table 7.3: Correlation coefficients for B factor prediction obtained by optimal FRI (opFRI), parameter free FRI (pfFRI), and Gaussian normal mode (GNM) for large-size structures. Results for opFRI, pfFRI are taken from Opron *et al* [3]. GNM and NMA values are taken from the coarse-grained (C_α) results reported in Park *et al* [4]. MWCG results are parameter free and use all C, N, and O to predict C_α . MWCG results originally published in Bramer *et al* [1].

PDB ID	N	MWCG	opFRI	pfFRI	GNM	NMA
1AHO	64	0.768	0.698	0.625	0.562	0.339
1ATG	231	0.843	0.613	0.578	0.497	0.154
1BYI	224	0.600	0.543	0.491	0.552	0.133
1CCR	111	0.741	0.580	0.512	0.351	0.530
1E5K	188	0.848	0.746	0.732	0.859	0.620
1EW4	106	0.804	0.650	0.644	0.547	0.447
1IFR	113	0.875	0.697	0.689	0.637	0.330
1NKO	122	0.831	0.619	0.535	0.368	0.322
1NLS	238	0.799	0.669	0.530	0.523	0.385
1O08	221	0.516	0.562	0.333	0.309	0.616
1PMY	123	0.701	0.671	0.654	0.685	0.702
1PZ4	114	0.921	0.828	0.781	0.843	0.844
1QTO	122	0.809	0.543	0.520	0.334	0.725
1RRO	112	0.748	0.435	0.372	0.529	0.546
1UKU	102	0.765	0.665	0.661	0.742	0.720
1V70	105	0.854	0.622	0.492	0.162	0.285
1WBE	204	0.767	0.591	0.577	0.549	0.574
1WHI	122	0.804	0.601	0.539	0.270	0.414
1WPA	107	0.797	0.634	0.577	0.417	0.380
2AGK	233	0.821	0.705	0.694	0.512	0.514
2C71	205	0.773	0.658	0.649	0.560	0.584
2CG7	90	0.738	0.551	0.539	0.379	0.308
2CWS	227	0.756	0.647	0.640	0.696	0.524
2HQK	213	0.897	0.824	0.809	0.365	0.743
2HYK	238	0.728	0.585	0.575	0.510	0.593
2I24	113	0.672	0.593	0.498	0.494	0.441
2IMF	203	0.798	0.652	0.625	0.514	0.401
2PPN	107	0.673	0.677	0.638	0.668	0.468
2R16	176	0.640	0.582	0.495	0.618	0.411
2V9V	135	0.697	0.555	0.548	0.528	0.594
2VIM	104	0.859	0.413	0.393	0.212	0.221
2VPA	204	0.757	0.763	0.755	0.576	0.594
2VYO	210	0.777	0.675	0.648	0.729	0.739
3SEB	238	0.879	0.801	0.712	0.826	0.720
3VUB	101	0.852	0.625	0.610	0.607	0.365

Table 7.5: Average pearson correlation coefficients for C_α B factor prediction with FRI, GNM and NMA for three structure sets from Park *et al.* [4] and a superset of 364 structures. Results for opFRI, pfFRI are taken from Opron *et al* [3]. GNM and NMA values are taken from the coarse-grained (C_α) results reported in Park *et al.* [4] MWCG results are parameter free and use all C, N, and O to predict C_α . MWCG Results originally published in Bramer *et al* [1].

PDB set	MWCG	opFRI[3]	pfFRI[3]	GNM	NMA[4]
Small	0.921	0.667	0.594	0.541[4]	0.480
Medium	0.795	0.664	0.605	0.550[4]	0.482
Large	0.775	0.636	0.591	0.529[4]	0.494
Superset	0.803	0.673	0.626	0.565 [3]	NA

Table 7.6: Pearson Correlation coefficients for C_α , non C_α carbon, nitrogen, oxygen, and sulfur using parameter free MWCG. Only 215 of the 364 proteins contain sulfur atoms. MWCG results originally published in Bramer *et al* [1].

Subset	C_α	Non C_α Carbon	Nitrogen	Oxygen	Sulfur
Average	0.803	0.744	0.812	0.789	0.903
No of proteins	364	364	364	364	215

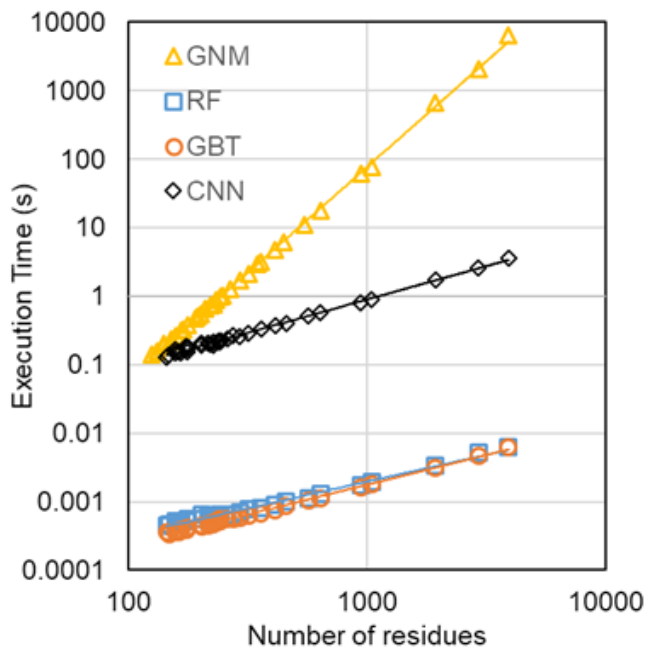


Figure 7.7: CPU Efficiency comparison between GNM [3], RF, GBT, and CNN algorithms for MWCG B factor prediction. Execution times in seconds (s) versus number of residues. A set of 34 proteins, listed in Table 7.7, were used to evaluate the computational complexity. Result originally published in Bramer *et al* [2].

Table 7.7: CPU execution times, in seconds, from efficiency comparison between GNM [3], RF, GBT, and CNN. Results originally reported in Bramer *et al* [2]

PDB	N	GNM[3]	RF	GBT	CNN
3P6J	125	0.141	0.000455	0.000358	0.130
3R87	132	0.156	0.000464	0.000339	0.138
3KBE	140	0.187	0.000505	0.000384	0.149
1TZV	141	0.203	0.000473	0.000365	0.163
2VY8	149	0.219	0.000486	0.000359	0.156
3ZIT	152	0.234	0.000519	0.000365	0.148
2FG1	157	0.265	0.000518	0.000403	0.174
2X3M	166	0.312	0.000526	0.000382	0.182
3LAA	169	0.327	0.000514	0.000405	0.155
3M8J	178	0.375	0.000548	0.000412	0.178
2GZQ	191	0.468	0.000647	0.000454	0.195
4G7X	194	0.499	0.000631	0.000445	0.209
2J9W	200	0.546	0.000554	0.000424	0.208
3TUA	210	0.655	0.000602	0.000472	0.217
1U9C	221	0.733	0.000592	0.000486	0.198
3ZRX	221	0.718	0.000654	0.000515	0.216
3K6Y	227	0.765	0.000619	0.000490	0.189
3OQY	234	0.873	0.000619	0.000502	0.211
2J32	244	0.967	0.000625	0.000556	0.225
3M3P	249	1.029	0.000621	0.000525	0.220
1U7I	267	1.263	0.000647	0.000551	0.237
4B9G	292	1.669	0.000693	0.000574	0.256
4ERY	318	2.122	0.000775	0.000619	0.289
3MGN	348	2.902	0.000655	0.000552	0.267
2ZU1	360	3.136	0.000816	0.000675	0.337
2Q52	412	4.696	0.000900	0.000750	0.369
4F01	448	6.178	0.001016	0.000878	0.401
3DRF	547	11.154	0.001131	0.001033	0.512
3UR8	637	17.409	0.001307	0.001136	0.583
2AH1	939	61.012	0.001716	0.001605	0.800
1GCO	1044	75.801	0.001936	0.001814	0.905
1F8R	1932	654.127	0.003343	0.003163	1.745
1H6V	2927	2085.842	0.005205	0.004739	2.543
1QKI	3912	6365.668	0.006261	0.006198	3.560

Table 7.8: Average Pearson correlation coefficients (PCC) both of all heavy atom and C_α only B factor predictions for small-, medium-, and large-sized protein sets along with the entire superset of the 364 protein dataset. Predictions of random forest (RF), gradient boosted tree (GBT), and convolutional neural network (CNN) are obtained by leave-one-protein-out (blind), while predictions of parameter-free flexibility-rigidity index (pfFRI), Gaussian network model (GNM) and normal mode analysis (NMA) were obtained via the least squares fitting of individual proteins. All machine learning models use all heavy atom information for training. MWCG machine learning B factor prediction results originally reported in Bramer *et al* [2].

Prediction Of Only C_α						
Protein Set	RF	GBT	CNN	pfFRI [3]	GNM [3]	NMA [3]
Small	0.25	0.39	0.53	0.60	0.54	0.48
Medium	0.47	0.59	0.55	0.61	0.55	0.48
Large	0.50	0.57	0.62	0.59	0.53	0.49
Superset	0.49	0.57	0.66	0.63	0.57	NA
Prediction Of All Heavy Atom						
Protein Set	RF	GBT	CNN	pfFRI [3]	GNM [3]	NMA [3]
Small	0.44	0.49	0.56	NA	NA	NA
Medium	0.59	0.64	0.62	NA	NA	NA
Large	0.62	0.65	0.68	NA	NA	NA
Superset	0.59	0.63	0.69	NA	NA	NA

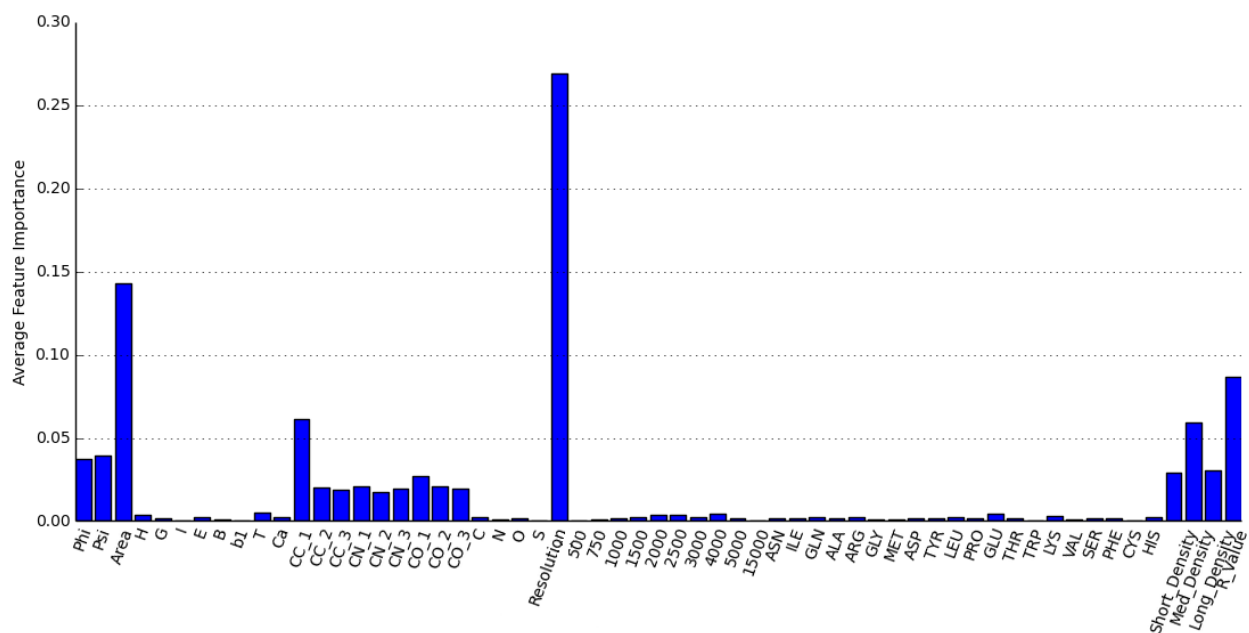


Figure 7.8: Individual feature importance for the MWCG random forest model averaged over the data set. Reported feature selection includes the use heavy atoms in the model. Figure originally published in Bramer *et al* [2].

Table 7.9: Pearson correlation coefficients for cross protein heavy atom blind MWCG B factor prediction obtained by random forest (RF), boosted gradient (GBT), and convolutional neural network (CNN) for the small-sized protein set. Results reported use heavy atoms in both training and prediction. Originally published in Bramer *et al* [2].

PDB ID	N	RF	GBT	CNN
1AIE	235	0.62	0.53	0.60
1AKG	108	0.41	0.51	0.70
1BX7	345	0.55	0.67	0.63
1ETL	76	0.27	0.03	0.48
1ETM	80	0.46	0.13	0.48
1ETN	77	0.33	0.25	0.20
1FF4	477	0.55	0.59	0.76
1GK7	321	0.53	0.73	0.72
1GVD	401	0.66	0.69	0.71
1HJE	73	-0.07	0.46	0.37
1KYC	138	0.43	0.30	0.32
1NOT	96	-0.18	0.81	0.63
1O06	142	0.51	0.64	0.65
1P9I	203	0.73	0.77	0.77
1PEF	153	0.60	0.64	0.76
1PEN	109	0.34	0.24	0.21
1Q9B	303	0.41	0.67	0.75
1RJU	257	0.71	0.75	0.73
1U06	432	0.55	0.68	0.61
1UOY	452	0.55	0.56	0.55
1USE	290	0.25	0.50	0.68
1VRZ	66	0.38	-0.17	0.09
1XY2	62	0.16	0.27	0.55
1YJO	55	0.36	0.12	0.02
1YZM	361	0.51	0.60	0.56
2DSX	386	0.36	0.44	0.56
2JKU	229	0.57	0.63	0.35
2NLS	269	0.45	0.49	0.70
2OL9	51	0.65	0.51	0.84
6RXN	345	0.56	0.71	0.82

Table 7.10: Pearson correlation coefficients for cross protein heavy atom blind MWCG B factor prediction obtained by random forest (RF), boosted gradient (GBT), and convolutional neural network (CNN) for the medium-sized protein set. Results reported use heavy atoms in both training and prediction. Originally published in Bramer *et al* [2].

PDB ID	N	RF	GBT	CNN
1ABA	728	0.74	0.77	0.73
1CYO	697	0.66	0.68	0.76
1FK5	626	0.62	0.71	0.63
1GXU	694	0.65	0.67	0.66
1I71	683	0.57	0.62	0.66
1LR7	522	0.53	0.70	0.71
1N7E	700	0.62	0.65	0.71
1NNX	674	0.69	0.73	0.53
1NOA	778	0.52	0.57	0.57
1OPD	642	0.55	0.60	0.62
1QAU	812	0.57	0.58	0.57
1R7J	729	0.71	0.70	0.65
1UHA	623	0.74	0.80	0.75
1ULR	677	0.69	0.71	0.68
1USM	631	0.59	0.78	0.67
1V05	17	-0.20	0.02	0.60
1W2L	746	0.62	0.68	0.69
1X3O	622	0.53	0.52	0.63
1Z21	771	0.63	0.66	0.63
1ZVA	551	0.59	0.56	0.58
2BF9	287	0.39	0.52	0.70
2BRF	735	0.76	0.78	0.86
2CE0	714	0.62	0.65	0.90
2E3H	589	0.70	0.73	0.38
2EAQ	705	0.63	0.61	0.58
2EHS	590	0.55	0.71	0.38
2FQ3	721	0.67	0.75	0.76
2IP6	702	0.62	0.67	0.64
2MCM	735	0.71	0.73	0.60
2NUH	806	0.64	0.72	0.19
2PKT	666	0.06	0.17	0.76
2PLT	719	0.62	0.67	0.70
2QJL	734	0.61	0.60	0.42
2RB8	723	0.61	0.64	0.42
3BZQ	742	0.60	0.61	0.43
5CYT	800	0.68	0.70	0.74

Table 7.11: Pearson correlation coefficients for cross protein heavy atom blind MWCG B factor prediction obtained by random forest (RF), boosted gradient (GBT), and convolutional neural network (CNN) for the large-sized protein set. Results reported use heavy atoms in both training and prediction. Originally published in Bramer *et al* [2].

PDB ID	N	RF	GBT	CNN
1AHO	482	0.62	0.71	0.76
1ATG	1689	0.61	0.66	0.63
1BYI	1540	0.59	0.63	0.59
1CCR	837	0.70	0.67	0.66
1E5K	1423	0.70	0.73	0.74
1EW4	863	0.70	0.71	0.61
1IFR	878	0.72	0.74	0.73
1NLS	1746	0.61	0.64	0.56
1O08	1722	0.51	0.58	0.55
1PMY	937	0.64	0.65	0.67
1PZ4	874	0.73	0.73	0.74
1QTO	934	0.61	0.55	0.63
1RRO	846	0.56	0.52	0.54
1UKU	873	0.74	0.75	0.70
1V70	784	0.70	0.67	0.62
1WBE	1542	0.59	0.61	0.63
1WHI	937	0.74	0.77	0.71
1WPA	906	0.64	0.66	0.74
2AGK	1867	0.61	0.68	0.44
2C71	1446	0.59	0.61	0.83
2CG7	536	0.47	0.54	0.79
2CWS	1624	0.63	0.60	0.78
2HQB	1582	0.76	0.76	0.90
2HYK	1832	0.60	0.65	0.85
2I24	872	0.52	0.52	0.91
2IMF	1564	0.62	0.62	0.47
2PPN	701	0.50	0.68	0.83
2R16	1262	0.52	0.53	0.50
2V9V	986	0.64	0.61	0.63
2VIM	781	0.62	0.61	0.75
2VPA	1524	0.63	0.68	0.61
2VYO	1589	0.53	0.65	0.61
3SEB	1948	0.61	0.71	0.57
3VUB	787	0.64	0.70	0.78

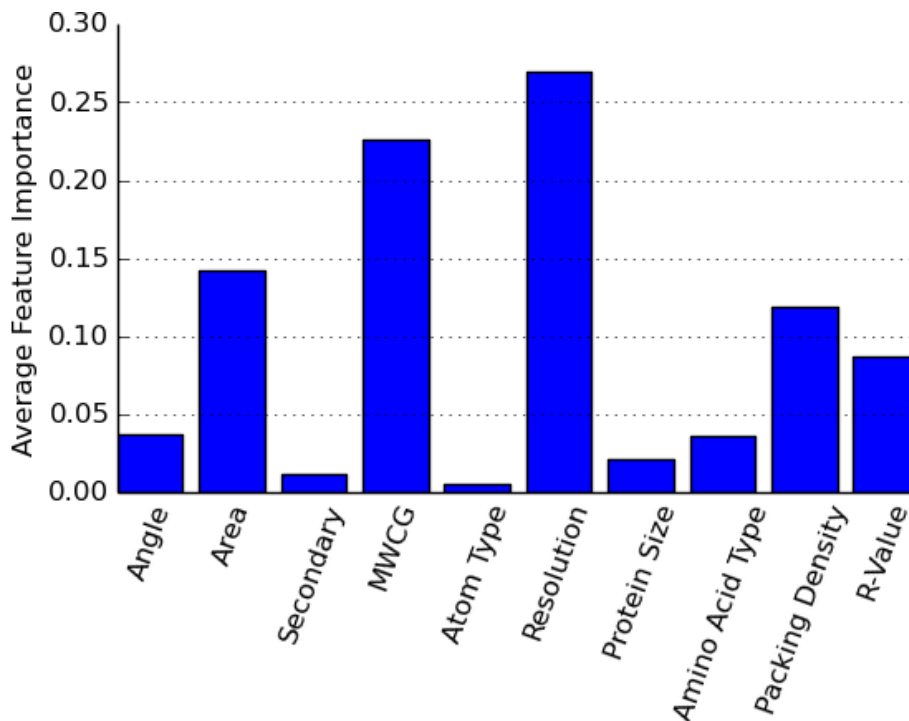


Figure 7.9: Average feature importance for the MWCG random forest model with the angle, secondary, MWCG, atom type, protein size, amino acid, and packing density features aggregated. Reported feature selection includes the use heavy atoms in the model. Figure originally published in Bramer *et al* [2].

Table 7.13: ASPH and ESPH average Pearson correlation coefficients C_{α} B factor predictions for small-, medium-, and large-sized protein sets along with the entire superset of the 364 protein dataset. Gradient boosted tree (GBT), convolutional neural network, and consensus(CON) results are obtained by leave-one-protein-out (blind). Predictions of parameter-free flexibility-rigidity index (pFRI), Gaussian network model (GNM) and normal mode analysis (NMA) were obtained via the least squares fitting of individual proteins.

	CNN	GBT	CON	pFRI	GNM	NMA
Small	0.63	0.58	0.62	0.59	0.54	0.48
Medium	0.60	0.58	0.61	0.61	0.55	0.48
Large	0.58	0.59	0.58	0.59	0.53	0.49
Superset	0.60	0.59	0.61	0.63	0.57	NA

Table 7.14: ASPH and ESPH Pearson correlation coefficients for cross protein C_α atom blind B factor prediction obtained by boosted gradient (GBT), convolutional neural network (CNN), and consensus (CON) for the small-sized protein set.

PDB ID	N	GBT	CNN	CON
1AIE	31	0.75	0.7	0.78
1AKG	16	0.27	0.32	0.29
1BX7	51	0.74	0.74	0.76
1ETL	12	0.37	0.82	0.55
1ETM	12	0.37	0.63	0.43
1ETN	12	0.07	0.48	0.13
1FF4	65	0.61	0.66	0.64
1GK7	39	0.77	0.9	0.82
1GVD	56	0.71	0.55	0.69
1HJE	13	0.84	0.75	0.9
1KYC	15	0.62	0.69	0.66
1NOT	13	0.69	0.96	0.8
1O06	22	0.94	0.93	0.95
1P9I	29	0.73	0.73	0.74
1PEF	18	0.79	0.82	0.82
1PEN	16	0.36	0.74	0.44
1Q9B	44	0.59	0.85	0.67
1RJU	36	0.6	0.46	0.58
1U06	55	0.44	0.4	0.45
1UOY	64	0.72	0.7	0.76
1USE	47	0.05	0.32	0.12
1VRZ	13	0.54	0.34	0.54
1XY2	8	0.79	0.82	0.81
1YJO	6	0.7	-0.06	0.57
1YZM	46	0.69	0.64	0.7
2DSX	52	0.34	0.34	0.36
2JKU	38	0.57	0.71	0.66
2NLS	36	0.23	0.47	0.29
2OL9	6	0.94	0.85	0.94
6RXN	45	0.59	0.6	0.61

Table 7.15: ASPH and ESPH Pearson correlation coefficients for cross protein C_α atom blind B factor prediction obtained by boosted gradient (GBT), convolutional neural network (CNN), and consensus (CON) for the medium-sized protein set.

PDB ID	N	GBT	CNN	CON
1ABA	87	0.73	0.71	0.74
1CYO	88	0.64	0.7	0.68
1FK5	93	0.59	0.6	0.61
1GXU	89	0.67	0.68	0.69
1I71	83	0.53	0.58	0.56
1LR7	73	0.62	0.61	0.64
1N7E	95	0.63	0.58	0.65
1NNX	93	0.78	0.79	0.8
1NOA	113	0.55	0.53	0.56
1OPD	85	0.42	0.34	0.41
1QAU	112	0.51	0.59	0.57
1R7J	90	0.71	0.77	0.75
1UHA	82	0.71	0.74	0.73
1ULR	87	0.54	0.53	0.56
1USM	77	0.73	0.72	0.75
1V05	96	0.6	0.64	0.63
1W2L	97	0.43	0.5	0.47
1X3O	80	0.41	0.43	0.44
1Z21	96	0.68	0.65	0.69
1ZVA	75	0.7	0.7	0.71
2BF9	35	0.48	0.79	0.58
2BRF	103	0.72	0.77	0.75
2CE0	109	0.6	0.66	0.64
2E3H	81	0.65	0.68	0.67
2EAQ	89	0.57	0.63	0.61
2EHS	75	0.62	0.67	0.65
2FQ3	85	0.77	0.82	0.81
2IP6	87	0.6	0.66	0.63
2MCM	112	0.71	0.77	0.75
2NUH	104	0.72	0.56	0.7
2PKT	93	0.01	-0.04	-0.01
2PLT	98	0.52	0.53	0.54
2QJL	107	0.54	0.57	0.56
2RB8	93	0.67	0.7	0.7
3BZQ	99	0.45	0.53	0.49
5CYT	103	0.39	0.34	0.39

Table 7.16: ASPH and ESPH Pearson correlation coefficients for cross protein C_α atom blind B factor prediction obtained boosted gradient (GBT), convolutional neural network (CNN), and consensus (CON) for the large-sized protein set.

PDB ID	N	GBT	CNN	CON
1AHO	66	0.66	0.66	0.7
1ATG	231	0.55	0.51	0.55
1BYI	238	0.61	0.5	0.6
1CCR	109	0.55	0.6	0.59
1E5K	188	0.74	0.72	0.74
1EW4	106	0.59	0.6	0.61
1IFR	113	0.7	0.64	0.7
1NLS	238	0.55	0.57	0.57
1O08	221	0.49	0.47	0.49
1PMY	123	0.59	0.7	0.65
1PZ4	113	0.72	0.8	0.77
1QTO	122	0.53	0.48	0.54
1RRO	108	0.4	0.45	0.43
1UKU	102	0.75	0.76	0.77
1V70	105	0.63	0.62	0.64
1WBE	206	0.6	0.56	0.6
1WHI	122	0.59	0.56	0.6
1WPA	107	0.65	0.65	0.67
2AGK	233	0.67	0.63	0.67
2C71	225	0.57	0.6	0.6
2CG7	110	0.3	0.32	0.32
2CWS	235	0.61	0.47	0.6
2HQK	232	0.77	0.77	0.78
2HYK	237	0.65	0.63	0.65
2I24	113	0.44	0.46	0.46
2IMF	203	0.53	0.58	0.56
2PPN	122	0.64	0.54	0.63
2R16	185	0.44	0.49	0.46
2V9V	149	0.53	0.52	0.54
2VIM	114	0.44	0.47	0.47
2VPA	217	0.66	0.75	0.71
2VYO	207	0.6	0.63	0.63
3SEB	238	0.63	0.6	0.63
3VUB	101	0.59	0.55	0.59

Table 7.17: ASPH and ESPH Pearson correlation coefficients of least squares fitting C_α B factor prediction of small proteins using 11Å cutoff. Two Bottleneck (B) and Wasserstein (W) metrics using various kernel choices are included.

PDB ID	N	B & W			B			W		
		Exp	Lor	Both	Exp	Lor	Both	Exp	Lor	Both
1AIE	31	0.97	0.88	0.99	0.78	0.64	0.90	0.90	0.77	0.96
1AKG	16	0.82	0.66	1.00	0.60	0.53	0.72	0.53	0.56	0.87
1BX7	51	0.86	0.74	0.89	0.79	0.68	0.82	0.81	0.69	0.82
1ETL	12	1.00	1.00	1.00	0.68	0.87	1.00	0.95	0.98	1.00
1ETM	12	1.00	1.00	1.00	0.45	0.74	0.86	0.70	0.83	1.00
1ETN	12	1.00	1.00	1.00	0.96	0.92	0.99	0.70	0.92	1.00
1FF4	65	0.77	0.72	0.80	0.70	0.65	0.75	0.68	0.68	0.76
1GK7	39	0.95	0.94	0.98	0.91	0.93	0.95	0.88	0.92	0.94
1GVD	56	0.75	0.68	0.84	0.67	0.63	0.69	0.61	0.62	0.66
1HJE	13	1.00	1.00	1.00	0.72	0.79	1.00	0.67	0.57	1.00
1KYC	15	0.96	0.99	1.00	0.92	0.93	0.99	0.88	0.88	1.00
1NOT	13	1.00	1.00	1.00	0.82	0.86	1.00	0.86	0.81	1.00
1O06	22	0.98	0.97	1.00	0.96	0.92	0.97	0.97	0.94	0.98
1P9I	29	0.89	0.88	0.98	0.87	0.82	0.92	0.87	0.84	0.89
1PEF	18	0.96	0.97	1.00	0.88	0.94	0.96	0.92	0.94	0.96
1PEN	16	0.96	0.90	1.00	0.60	0.67	0.83	0.47	0.73	0.94
1Q9B	44	0.79	0.76	0.94	0.58	0.59	0.69	0.69	0.57	0.71
1RJU	36	0.81	0.74	0.91	0.75	0.69	0.81	0.62	0.65	0.72
1U06	55	0.50	0.52	0.72	0.37	0.36	0.52	0.46	0.39	0.55
1UOY	64	0.73	0.72	0.83	0.65	0.66	0.69	0.65	0.69	0.73
1USE	47	0.66	0.75	0.91	0.50	0.52	0.72	0.46	0.53	0.64
1VRZ	13	1.00	1.00	1.00	0.92	0.92	1.00	0.77	0.85	1.00
1XY2	8	1.00	1.00	1.00	0.99	0.95	1.00	0.91	0.91	1.00
1YJO	6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1YZM	46	0.87	0.90	0.95	0.82	0.72	0.88	0.86	0.84	0.90
2DSX	52	0.54	0.50	0.78	0.37	0.30	0.56	0.41	0.36	0.55
2JKU	38	0.89	0.75	0.95	0.85	0.65	0.88	0.83	0.60	0.88
2NLS	36	0.75	0.66	0.88	0.61	0.32	0.76	0.49	0.47	0.69
2OL9	6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
6RXN	45	0.74	0.63	0.86	0.59	0.48	0.76	0.49	0.49	0.76

Table 7.18: ASPH and ESPH Pearson correlation coefficients of least squares fitting C_α B factor prediction of medium proteins using 11Å cutoff. Two Bottleneck (B) and Wasserstein (W) metrics using various kernel choices are included.

PDB ID	N	B & W			B			W		
		Exp	Lor	Both	Exp	Lor	Both	Exp	Lor	Both
1ABA	87	0.67	0.67	0.76	0.54	0.62	0.68	0.56	0.63	0.70
1CYO	88	0.71	0.69	0.78	0.66	0.58	0.68	0.65	0.59	0.67
1FK5	93	0.53	0.59	0.71	0.49	0.50	0.58	0.49	0.50	0.55
1GXU	89	0.75	0.78	0.82	0.72	0.61	0.75	0.69	0.72	0.77
1I71	83	0.44	0.66	0.76	0.41	0.46	0.56	0.38	0.58	0.59
1LR7	73	0.61	0.62	0.71	0.57	0.55	0.63	0.46	0.56	0.58
1N7E	95	0.67	0.71	0.80	0.54	0.68	0.72	0.54	0.63	0.73
1NNX	93	0.84	0.84	0.88	0.81	0.79	0.83	0.81	0.81	0.86
1NOA	113	0.63	0.65	0.72	0.60	0.57	0.63	0.53	0.57	0.59
1OPD	85	0.35	0.29	0.57	0.26	0.21	0.36	0.29	0.19	0.36
1QAU	112	0.59	0.61	0.66	0.57	0.55	0.58	0.55	0.57	0.58
1R7J	90	0.88	0.86	0.91	0.83	0.76	0.87	0.81	0.79	0.86
1UHA	82	0.70	0.75	0.82	0.69	0.68	0.74	0.67	0.69	0.73
1ULR	87	0.56	0.53	0.68	0.49	0.50	0.59	0.44	0.50	0.61
1USM	77	0.62	0.61	0.81	0.57	0.53	0.66	0.61	0.58	0.65
1V05	96	0.67	0.66	0.72	0.60	0.61	0.65	0.52	0.61	0.65
1W2L	97	0.72	0.72	0.79	0.60	0.63	0.69	0.56	0.61	0.69
1X3O	80	0.66	0.66	0.72	0.62	0.60	0.65	0.62	0.64	0.67
1Z21	96	0.70	0.73	0.82	0.61	0.63	0.64	0.64	0.69	0.72
1ZVA	75	0.85	0.85	0.94	0.84	0.78	0.92	0.83	0.81	0.86
2BF9	35	0.94	0.73	0.97	0.70	0.65	0.78	0.89	0.71	0.92
2BRF	103	0.74	0.73	0.76	0.74	0.71	0.74	0.72	0.72	0.75
2CE0	109	0.77	0.79	0.86	0.75	0.73	0.80	0.71	0.77	0.79
2E3H	81	0.66	0.71	0.82	0.62	0.69	0.76	0.56	0.69	0.78
2EAQ	89	0.81	0.77	0.86	0.79	0.72	0.81	0.77	0.76	0.82
2EHS	75	0.75	0.73	0.81	0.72	0.71	0.74	0.69	0.71	0.73
2FQ3	85	0.78	0.76	0.82	0.75	0.75	0.79	0.68	0.75	0.78
2IP6	87	0.72	0.66	0.82	0.67	0.58	0.73	0.64	0.64	0.78
2MCM	112	0.80	0.80	0.85	0.78	0.77	0.81	0.75	0.77	0.82
2NUH	104	0.77	0.74	0.85	0.73	0.63	0.81	0.75	0.66	0.80
2PKT	93	0.44	0.39	0.69	0.39	0.35	0.55	0.36	0.36	0.43
2PLT	98	0.66	0.63	0.72	0.57	0.59	0.67	0.52	0.59	0.66
2QJL	107	0.45	0.52	0.63	0.42	0.46	0.50	0.41	0.49	0.51
2RB8	93	0.81	0.78	0.84	0.78	0.75	0.80	0.74	0.76	0.81
3BZQ	99	0.57	0.62	0.69	0.50	0.55	0.61	0.47	0.55	0.59
5CYT	103	0.53	0.52	0.65	0.49	0.46	0.54	0.43	0.48	0.50

Table 7.19: ASPH and ESPH Pearson correlation coefficients of least squares fitting C_α B factor prediction of large proteins using 11Å cutoff. Two Bottleneck (B) and Wasserstein (W) metrics using various kernel choices are included.

PDB ID	N	B & W			B			W		
		Exp	Lor	Both	Exp	Lor	Both	Exp	Lor	Both
1AHO	66	0.75	0.78	0.88	0.72	0.73	0.79	0.53	0.65	0.75
1ATG	231	0.50	0.50	0.61	0.45	0.47	0.53	0.38	0.48	0.51
1BYI	238	0.50	0.51	0.58	0.41	0.46	0.49	0.44	0.48	0.54
1CCR	109	0.65	0.66	0.71	0.53	0.56	0.65	0.43	0.58	0.63
1E5K	188	0.67	0.68	0.74	0.66	0.67	0.68	0.63	0.67	0.69
1EW4	106	0.58	0.60	0.73	0.52	0.51	0.55	0.55	0.55	0.62
1IFR	113	0.65	0.59	0.73	0.56	0.54	0.65	0.47	0.53	0.62
1NLS	238	0.81	0.78	0.86	0.75	0.65	0.83	0.80	0.72	0.82
1O08	221	0.46	0.48	0.56	0.44	0.42	0.50	0.37	0.45	0.48
1PMY	123	0.71	0.70	0.76	0.62	0.59	0.67	0.68	0.69	0.71
1PZ4	113	0.88	0.82	0.93	0.86	0.74	0.89	0.85	0.76	0.88
1QTO	122	0.59	0.59	0.65	0.48	0.46	0.53	0.55	0.52	0.56
1RRO	108	0.39	0.35	0.56	0.31	0.23	0.45	0.33	0.19	0.45
1UKU	102	0.80	0.81	0.84	0.78	0.80	0.80	0.74	0.80	0.80
1V70	105	0.64	0.65	0.75	0.56	0.60	0.66	0.51	0.58	0.62
1WBE	206	0.53	0.47	0.63	0.43	0.38	0.55	0.36	0.42	0.48
1WHI	122	0.57	0.55	0.63	0.42	0.44	0.57	0.34	0.43	0.55
1WPA	107	0.70	0.69	0.79	0.61	0.52	0.71	0.66	0.56	0.70
2AGK	233	0.65	0.65	0.69	0.61	0.64	0.65	0.55	0.63	0.67
2C71	225	0.45	0.38	0.56	0.29	0.33	0.42	0.23	0.30	0.48
2CG7	110	0.32	0.44	0.63	0.29	0.31	0.36	0.30	0.33	0.41
2CWS	235	0.59	0.55	0.66	0.53	0.52	0.54	0.40	0.52	0.55
2HQK	232	0.80	0.79	0.83	0.70	0.74	0.80	0.68	0.76	0.81
2HYK	237	0.59	0.58	0.63	0.51	0.55	0.59	0.43	0.54	0.60
2I24	113	0.47	0.44	0.69	0.40	0.40	0.48	0.45	0.40	0.49
2IMF	203	0.61	0.65	0.71	0.59	0.56	0.60	0.59	0.59	0.64
2PPN	122	0.57	0.61	0.74	0.51	0.59	0.63	0.44	0.57	0.63
2R16	185	0.50	0.51	0.66	0.46	0.45	0.51	0.45	0.46	0.52
2V9V	149	0.60	0.51	0.66	0.53	0.48	0.56	0.55	0.50	0.62
2VIM	114	0.38	0.33	0.52	0.29	0.28	0.41	0.24	0.31	0.40
2VPA	217	0.73	0.75	0.78	0.72	0.71	0.73	0.68	0.73	0.74
2VYO	207	0.68	0.70	0.77	0.64	0.66	0.72	0.59	0.68	0.70
3SEB	238	0.63	0.66	0.77	0.62	0.61	0.68	0.61	0.62	0.67
3VUB	101	0.65	0.60	0.71	0.60	0.56	0.61	0.61	0.57	0.64

Table 7.20: ASPH and ESPH average Pearson correlation coefficients of least squares fitting C_α B factor prediction of small, medium, large, and superset using 11Å cutoff. Two Bottleneck (B) and Wasserstein (W) metrics using various kernel choices are included. Results for pFRI are taken from Opron et al[3]. GNM and NMA value are taken from the course grained C_α results reported in Park et al[4].

	B & W			B			W			pFRI	GNM	NMA
	Exp	Lor	Both	Exp	Lor	Both	Exp	Lor	Both			
Small	0.87	0.84	0.94	0.74	0.72	0.85	0.74	0.73	0.86	0.59	0.54	0.48
Medium	0.68	0.68	0.78	0.62	0.61	0.69	0.60	0.63	0.69	0.61	0.55	0.48
Large	0.61	0.60	0.70	0.54	0.54	0.61	0.51	0.55	0.62	0.59	0.53	0.49
Superset	0.65	0.64	0.73	0.58	0.58	0.65	0.55	0.59	0.66	0.63	0.57	NA

Chapter 8

Discussion

8.1 Element Specific Heat Maps

One useful application of the WCG method is the generation of element specific correlation heat maps. These maps provide a two dimensional visualization of important secondary and tertiary components of a given protein. Of course, maps of this kind are not new, for example see Opron *et al* for past use. However, the correlation maps provided here are the first of their kind. Previous correlation maps have only considered C_α interactions. The maps provided here in Figures 7.1, 7.2, and 7.3 illustrate that the more general frame work of the WCG method is a valid and useful approach to furthering our understanding of protein structure and flexibility. The results presented here generate correlation maps using PDB ID 3TYS, 1AIE, and 3PSM to demonstrate the applicability of this approach.

Protein PDB ID 1AIE consists of a random coil attached to a single alpha helix. The provided amine nitrogen and double bonded carboxyl correlation heat maps in Figure 7.1 clearly show the alpha helix as indicated by a thick band along the diagonal. This thick band corresponds to the rigidity imposed by the local interactions of nearby residues within the alpha helix.

Protein PDB ID 1KGM is made up of various random coils and two anti-parallel beta sheets. The provided amine nitrogen and double bonded carboxyl correlation heat maps in

Figure 7.2 illustrate the interaction between residues in the anti parallel beta sheet with thick bands perpendicular to the main diagonal. These perpendicular bands correspond physically to the rigidity imposed by the interactions between the anti parallel beta sheets.

Protein PDB ID 5IIV presents two parallel alpha helices. The provided amine nitrogen and double bonded carboxyl correlation heat maps in Figure 7.3 illustrate both the short range interactions within a single alpha helix and interactions between alpha helices. The two alpha helices are represented clearly as two distinct thick bands along the diagonal. Thick off diagonal bands illustrate interactions between alpha helices. The diagrams also illustrate the strength of each type of bonding. Bonding within an alpha helix is stronger and thus the main diagonal of the correlation heat maps is warmer than the off diagonal which corresponds to the weaker alpha helix to alpha helix interaction.

8.2 Hinge Detection

Figures 7.5-7.6 show the B factor prediction comparison of protein PDB ID 1CLL, 1WHI, and 2HQK. Figure 7.5 clearly indicates GNM misses the hinge region present in calmodulin (PDB ID 1CLL) around residue 75. The WCG method clearly agree with the experimental results as indicated in the provided results. The MWCG method is also included in this result to demonstrate the ability of the MWCG method to capture multiple scales and improve the overall B factor prediction. In the ribosomal protein L14 (PDB ID 1WHI) the results demonstrate that WCG provides a more reliable prediction compared to GNM as seen in Figure 7.4. In particular, GNM incorrectly predicts a large flexible region around the 75th residue that does not exist. Lastly, the engineered cyan fluorescent protein mTFP1 (PDB ID 2HQK) is also considered. Figure 7.6 shows that GNM predicts a highly flexible region

incorrectly around the 60th residue whereas the WCG method agrees with the experimental results of low flexibility in that region. The results presented in this work demonstrate that GNM consistently misses hinge regions and predicts hinge regions where none exist. Comparatively, the WCG method is more accurate than GNM, and MWCG the most accurate of all the hinge prediction techniques studied here.

8.3 Fitting Models

8.3.1 MWCG

The MWCG method is used to predict B factor of a large and diverse set of over 300 proteins. Results for C_α B factor prediction are provided in Tables 7.1-7.6, and 7.4. Results of protein subsets of small, medium, and large sizes are considered in 7.1-7.6 and their overall average Pearson correlation coefficients are provided in 7.5. In all cases of C_α B factor prediction, the MWCG method outperforms previously existing methods in terms of average Pearson correlation coefficient. The MWCG method is notable in that, averaged over the superset of proteins, it provides a 19% improvement over the best existing method opFRI and a 42% improvement over GNM. Table 7.6 provides results for B factor prediction of other heavy elements such as non C_α carbon, nitrogen, oxygen, and sulfur atoms. This is also notable because to date no other previous method has included B factor prediction of elements other than C_α . These predictions also have a similar average correlation coefficient to the C_α results indicating the robustness of the model.

8.3.2 ASPH & ESPH

ASPH and ESPH methods are used for C_α only B factor using the same protein dataset as MWCG. Results for C_α only B factor prediction are provided in Tables 7.17-7.20, and 7.21. Results of protein subsets of small, medium, and large sizes are considered in 7.17-7.19 and their overall average Pearson correlation coefficients are provided in 7.20. Overall fitting methods using the various ASPH and ESPH features performed similarly. The best results came from using features generated by both exponential and lorentz kernels and both Bottleneck and Wasserstein distances. Using both kernels, ASPH and ESPH distance metrics resulted in an average correlation coefficient of 0.73 for the superset.

8.4 Machine Learning Models

8.4.1 MWCG

Among the three methods considered for MWCG based B factor prediction, the convolutional neural network method outperforms the boosted gradient tree and random forest by 10% and 17%, respectively. As reported in Table 7.12, no machine learning method outperforms any other method for every protein.

Compared to the deep CNN, the ensemble methods do not require as much parameter tuning. The random forest method is the simplest and most robust with only one hyperparameter. Overall the boosted gradient tree method outperforms the random forest for MWCG based B factor prediction for all data sets. To balance cost, time, and quality, 500 trees were used for the random forest and 1000 trees were used for the boosted gradient method for the MWCG B factor prediction. It's possible that this may account for the perfor-

mance difference between the boosted gradient tree method compared to the random forest. Ensemble methods are generally robust to overfitting, and adding more features would likely improve their results [42]. Moreover, boosted gradient trees use several hyperparameters so the model could be improved by further tuning these hyperparameters. The image-like heat map data used in the deep CNN provides additional data compared to the dataset used for the ensemble methods. This very likely explains the improved performance as compared to the ensemble methods. Providing more refined images, and other novel image types, would undoubtedly improve the results further but would come at a computational cost.

Applying several dropout layers prevents the CNNs from overfitting the data. Much like the GBT the CNN contains several hyperparameters. Thus, the CNN model would benefit from more careful parameter tuning as well. Incorporating a larger dataset, more features, and higher resolutions images would also improve the CNN performance. In general the results of the machine learning methods generated in this work could be further improved by refining features, exploring new features, and further tuning the hyperparameters.

8.4.2 ASPH & ESPH

Machine learning results for ASPH and ESPH can be found in 7.14-7.16 and 7.22. Overall both GBT and CNN algorithms perform similarly. As expected, the CNN method outperforms the GBT with average correlation coefficients over the superset of 0.60 and 0.59 respectively. The consensus method improves upon both results with an average Pearson correlation coefficient of 0.61 over the superset. Table 7.13 shows that the blind prediction machine learning models perform better than fitting models GNM and NMA and similar to the pFRI fitting model.

Chapter 9

Conclusions and Future Directions

9.1 Conclusions

Protein flexibility and dynamics are important tools for understanding the function, conformational states, folding, binding, and molecular mechanisms of proteins. It is a well known paradigm that protein flexibility strongly correlates with protein function. Protein interactions span multiple interactions scales and their complexity and large number of degrees of freedom make quantitative understanding a great challenge. Molecular dynamics offers a useful tool but is limited in scope due to the computational cost involved with large biomolecules or long time scales. Several successful time-independent methods have been developed that provide good B factor analysis at low computational cost. Examples include NMA [12, 10, 13, 11], ENM [15], GNM [18, 19, 50], and FRI methods [51, 3, 24, 22]. However, none of these methods can blindly predict cross protein B factors of an unknown protein. The guiding principle of this work is that intrinsic physics lies in a low-dimensional space that is embedded in a high dimensional data space. Based on this, the results of this work introduces graph theory based MWCG, ASPH, and ESPH[52, 53]. Moreover these methods are combined with advanced machine learning techniques to provide models that provide efficient and accurate tools for protein flexibility analysis and prediction. This work also outlines methods to successfully blindly predict cross-protein B factors.

First, this work introduces WCGs that efficiently reduce the protein structural complexity while accurately predicting protein flexibility. This work shows that weighted colored graphs are a useful and novel tool for investigating flexibility and dynamics of proteins. In section 7.2 the WCG approach was compared to experimental and GNM predicted B factor results. Nitrogen-nitrogen and oxygen-oxygen element specific correlation heat maps were constructed for several proteins using the WCG technique described in this work. As seen in Figures 7.1-7.3 these maps provide a clear picture of the various secondary and tertiary structures presented in these proteins.

The correlation heat maps presented demonstrate a fresh approach to representing protein flexibility and interactions visually. Previous approaches only use data from C_{α} atoms whereas the WCG method allows previously unused protein PDB data to be utilized. This provides a viable alternative method and makes such heat maps more robust as multiple heat maps can be constructed for each residue using different elements. Using double bonded carboxyl oxygens and amine nitrogens the work presented here demonstrates generality of the WCG approach. The WCG method introduced a unique opportunity for alternative approaches and allows for redundancy since the method is able to make use of non C_{α} atoms. This method can also include hydrogen atoms without any modifications, which may prove useful in future work as empirical methods inevitably become more accurate and robust.

Several proteins are tested to demonstrate the efficacy of WCGs to predict hinge regions of proteins. In this study we use proteins with well known flexibility to compare the ability of the GNM and the WCG method to predict flexible residues. Figures 7.5-7.6 show the B factor prediction comparison of protein PDB ID 1CLL, 1WHI, and 2HQK. The examples provided in this work demonstrate that WCG is an improvement upon the commonly used method of

GNM for hinge prediction. In proteins calmodulin (PDB ID 1CLL) and ribosomal protein (PDB ID 1WHI) the results show that prediction using GNM completely misses highly flexible hinge regions. The results using the engineered cyan fluorescent protein (PDB ID 2HQK) show that GNM incorrectly predicts a highly flexible region where none exists. In all the cases tested in this work the WCG method was able to correctly capture all the hinge regions and did not identify any false positive hinge regions. For further comparison the MWCG flexibility prediction is included in the calmodulin protein (PDB ID 1CLL) results seen in Figure 7.5. This result highlights the predictive power of the multiscale information that the MWCG method captures as seen with the excellent agreement with experimental results. Overall these results demonstrate the WCG and MWCG methods are superior tools to the commonly used GNM method in terms of the accuracy of hinge prediction for the provided examples.

The WCG method is used to predict B factor of a large and diverse set of over 300 proteins. Results for C_α B factor prediction are provided in Tables 7.1-7.6, and 7.4. Results of protein subsets of small, medium, and large sizes are considered in 7.1-7.6 and their overall B factor averages are provided in 7.5. In all cases of C_α B factor prediction, the MWCG method outperforms previously existing methods. The MWCG method is notable in that, averaged over the superset of proteins, it provides a 19% improvement over the best existing method opFRI and a 42% improvement over GNM. Table 7.6 provides results for B factor prediction of other heavy elements such as non C_α carbon, nitrogen, oxygen, and sulfur atoms. This is also notable because to date no other previous method has included B factor prediction of elements other than C_α . These predictions also have a similar average correlation coefficient to the C_α only prediction results indicating the robustness of the model.

To capture the multiscale protein interactions that occur over several characteristic length scales multiscale weighted colored graphs are constructed. The MWCGs are successfully used to construct models by linear least square fitting and a variety of machine learning techniques.

Several machine learning approaches were considered in this work for blind cross protein B factor prediction. In particular this work considered random forest, gradient boosting, and a deep convolutional neural network machine learning models for MWCG based B factor prediction. By using MWCG based features along with several local and global features this work uses advanced machine learning approaches to blindly predict protein flexibility and B factors. Moreover, unlike previous methods, this approach is able to predict B factors of any element the user desires provided 3D spatial coordinates are available. MWCG based images were engineered for the deep convolutional neural network. Overall the MWCG feature based deep convolutional neural networks provide the strongest predictive power in terms of B factor prediction which is likely accounted for by the additional data provided by the MWCG based image-like heat map features.

Several local, semi-local, and global features were included as machine learning features. MWCGs capture local structural properties corresponding to the intrinsic flexibility of the given protein. X-ray crystallography resolution and R-value provide global structures that allow the algorithms the ability to compare B factor across proteins. Packing density is a semi-local feature that captures several protein interaction scales.

Ensemble methods include relative feature importance used in the model which is provided in Figures 7.8 and 7.9. As seen in the figures both local and global features play an important role in the model. Overall the most meaningful global features are protein resolution and surface accessible area. On average, the most meaningful local feature of the

random forest model was the set of 9 MWCG features with the carbon-carbon kernel having most significance. Machine learning models often suffer from the black box problem. That is, once the model has trained, the user is unable to explicitly see how the model is determining predictions. Feature importance provides important insight into the underlying mechanics of the machine learning model. The feature importance results are in good agreement with our expectations within the context of protein flexibility analysis.

Both MWCG based fitting and machine learning B factor prediction demonstrate that MWCG based B factor prediction is more accurate in terms of Pearson correlation coefficient than previous fitting based methods such as GNM and NMA. For B factor prediction of C_α only atoms, the fitting model provided a 20% improvement over the next best B factor prediction method, opFRI, with a Pearson correlation coefficient of 0.80 averaged over the superset. The MWCG based deep CNN also outperformed opFRI, with a Pearson correlation coefficient of 0.66 averaged over the superset.

The working hypothesis is explored further by creating a B factor predictor using tools from algebraic topology. To the author’s knowledge, this is the first time persistent homology has been used to predict the B factor of atoms in proteins. This is a novel approach because topology is a global property and on its own cannot be directly used to describe local atomic information. This unique approach allows a localized topological representation to be constructed using a global mathematical tool. This approach accounts for multiple spatial interaction scales and element specific interactions. These results demonstrate that this is an accurate and robust topological approach.

This work introduces atom-specific topology and atom-specific persistent homology to construct localized topological representations for individual atoms from global topological tools. This approach works by constructing two conjugated sets of atoms. The first set

of atoms is centered around the given atom of interest while the other set is identical but excludes the atom of interest. To embed biological information into atom-specific persistent homology, element-specific selections are implemented. The topological distance between topological invariants generated from these conjugated sets of atoms provides a local topological representation of the atom of interest. To estimate the topological distances between conjugated barcodes both Bottleneck and Wasserstein metrics are utilized. For topological barcode generation, the Vietoris-Rips complex is employed. Atom-specific persistent homology features are generated using several element-specific interactions, kernel choices, parametrizations, and barcode distance metrics.

In this work ASPH and ESPH B factor prediction results are validated in two ways. First, topological features are used to fit protein B factors using linear least squares. The fitting model outperformed previous fitting models with an average Pearson correlation coefficient of 0.73 over the superset of proteins. These results show that the method is comparable to existing commonly used methods such as GNM and NMA.

Secondly, ASPH and ESPH features are used in machine learning models to blindly predict protein B factors of C_α atoms. Two machine learning models are used, a gradient boosted tree (GBT) and deep convolutional neural network (CNN). Additionally the C_α predictions from the two models are averaged to generate a more robust consensus model. In addition to the generated topological features, a variety of local and global features were included. The blind prediction consensus model provided the best results, outperforming both GNM and NMA fitting models and produced results similar to those of the pFRI fitting model. These results demonstrate that this is a robust model that is more accurate than existing GNM and NMA predictions. There are many other machine learning approaches available and testing those approaches is certainly worth exploring. Moreover, these results

could easily be improved by including a larger dataset, fine tuning parameters, and exploring different machine learning algorithms.

The proposed methods are tested and validated using a set of over 300 diverse proteins, or more than 600,000 B factors. For all machine learning models, a leave-one-protein-out approach is used to blindly predict protein B factors of all heavy atoms as well as only C_α atoms.

The work presented in this study is a first step using the recent advances in MWCG, ASPH, and ESPH based machine learning techniques to blindly predict cross protein B factors. These approaches are particularly notable compared to previous methods because of their ability to blindly predict protein B factors across proteins.

The MWCG, ASPH, and ESPH based machine learning results provided in this work are efficient and accurate compared to previous methods. This work provides clear evidence that machine learning approaches are useful and efficient for protein flexibility analysis. Nonetheless, many new and compelling features can be implemented in future work. Without a doubt these results can be improved by experimenting with various advanced machine learning approaches, larger datasets, and designing better mathematical descriptions of intrinsic flexibility.

The methods presented here can be applied to a variety of interesting applications related to molecules and biomolecules. Examples include allosteric site detection, hinge detection, hot spot identification, chemical shift analysis, atomic spectroscopy interpretation, and prediction of protein folding stability changes upon mutation. More generally these methods may be amenable to problems outside chemistry and biology such as network dynamics and social network centrality measure.

9.2 Future Directions

This work provides a rich basis for further exploration of mathematical approaches to protein analysis and flexibility. The following sections provide several areas of potential future research based on this work.

9.2.1 Software Development

To provide awareness and accessibility of the methods provided here, an online tool could be developed consisting of PDB files from the PDB database or uploads of a compatible structural file. Ideally users would be able to do any of the following:

- Choose MWCG based or ASPH/ESPH based models.
- Choose the number of kernels, type of kernel, and kernel parameterization.
- Choose element specific pairs and element specific heat maps.
- Choose machine learning algorithm and training features.
- Predict hinge regions based on user defined B factor cutoff.
- Predict the B factor of any atoms in a protein.
- Provide an interactive B factor colored 3D representation of the protein with downloadable image or gif files.

To host the website and run the required computations a server or cloud resources would be required.

9.2.2 Inclusion of other datasets

The protein databank currently contains over 138,000 protein structures whereas the work presented here used around 350 protein structures. The machine learning models presented here would undoubtedly benefit by including a larger dataset. More data would provide better validation and a more general framework for protein B factor prediction. However, there are enough proteins available at this time that even restricting to only C_α atoms would result in a data set of roughly 116,610,000 data points. So care would need to be taken to balance the amount of data used with the computational cost of training such models.

In addition to using larger datasets for training, models could be trained using more specific data. For example datasets could be selected based on specific types of proteins such as enzymes, structural proteins, signaling proteins, regulatory proteins, transport proteins, sensory proteins, motor proteins, defense proteins, and storage proteins.

9.2.3 Specific applications in drug design and docking pose

The applications provided here demonstrate the validity of the proposed method. Future work could apply the method to a variety of interesting problems. Drug design is an important and open problem where accurate and robust prediction of protein flexibility and dynamics are essential. Docking pose is another area where reliable B factor prediction may improve existing methods. Molecular docking programs common modeling tools for predicting ligand binding modes and structure based virtual screening.

9.2.4 Other general approaches

These methods could easily be developed to predict anisotropic B factors of a protein. Pairing this method with a local or global Hessian would allow the Hessian matrix to be local or global and by definition adaptive depending on the physical problem. Moreover, the methods provided here could be used for the following related work.

- Integrate these methods to include genetic sequence information for a more comprehensive model.
- Predict protein flexibility and dynamics from mutations.
- Investigation these tools as a general centrality measure in areas outside of biology.
- Investigation related topological data analysis techniques to understand proteins and protein networks.
- Test other advanced learning approaches such as reinforcement learning and long short term memory algorithms.

BIBLIOGRAPHY

BIBLIOGRAPHY

- [1] D. Bramer and G.-W. Wei, “Multiscale weighted colored graphs for protein flexibility and rigidity analysis,” *The Journal of Chemical Physics*, vol. 148, no. 5, p. 054103, 2018.
- [2] D. Bramer and G.-W. Wei, “Blind prediction of protein b-factor and flexibility,” *The Journal of Chemical Physics*, vol. 149, no. 13, p. 134107, 2018.
- [3] K. Opron, K. L. Xia, and G. W. Wei, “Fast and anisotropic flexibility-rigidity index for protein flexibility and fluctuation analysis,” *Journal of Chemical Physics*, vol. 140, p. 234105, 2014.
- [4] J. K. Park, R. Jernigan, and Z. Wu, “Coarse grained normal mode analysis vs. refined gaussian network model for protein residue-level structural fluctuations,” *Bulletin of Mathematical Biology*, vol. 75, pp. 124–160, 2013.
- [5] U. Emekli, S. Dina, H. Wolfson, R. Nussinov, and T. Haliloglu, “HingeProt: automated prediction of hinges in protein structures.,” *Proteins*, vol. 70, no. 4, pp. 1219–1227, 2008.
- [6] S. Flores and M. Gerstein, “FlexOracle: predicting flexible hinges by identification of stable domains,” *BMC bioinformatics*, vol. 8, no. 1, 2007.
- [7] S. Flores, L. Lu, J. Yang, N. Carriero, and M. Gerstein, “Hinge atlas: relating protein sequence to sites of structural flexibility.,” *BMC bioinformatics*, vol. 8, 2007.
- [8] K. S. Keating, S. C. Flores, M. B. Gerstein, and L. A. Kuhn, “StoneHinge: hinge prediction by network analysis of individual protein structures,” *Protein Science*, vol. 18, no. 2, pp. 359–371, 2009.
- [9] M. Shatsky, R. Nussinov, and H. J. Wolfson, “FlexProt: alignment of flexible protein structures without a predefinition of hinge regions,” *Journal of Computational Biology*, vol. 11, no. 1, pp. 83–8106, 2004.
- [10] N. Go, T. Noguti, and T. Nishikawa, “Dynamics of a small globular protein in terms of low-frequency vibrational modes,” *Proc. Natl. Acad. Sci.*, vol. 80, pp. 3696 – 3700, 1983.
- [11] M. Tasumi, H. Takenchi, S. Ataka, A. M. Dwivedi, and S. Krimm, “Normal vibrations of proteins: Glucagon,” *Biopolymers*, vol. 21, pp. 711 – 714, 1982.
- [12] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. States, S. Swaminathan, and M. Karplus, “Charmm: A program for macromolecular energy, minimization, and dynamics calculations,” *J. Comput. Chem.*, vol. 4, pp. 187–217, 1983.

- [13] M. Levitt, C. Sander, and P. S. Stern, “Protein normal-mode dynamics: Trypsin inhibitor, crambin, ribonuclease and lysozyme.,” *J. Mol. Biol.*, vol. 181, no. 3, pp. 423 – 447, 1985.
- [14] J. P. Ma, “Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes.,” *Structure*, vol. 13, pp. 373 – 180, 2005.
- [15] M. M. Tirion, “Large amplitude elastic motions in proteins from a single-parameter, atomic analysis.,” *Phys. Rev. Lett.*, vol. 77, pp. 1905 – 1908, 1996.
- [16] H. Goldstein, *Classical Mechanics*. Cambridge: Addison-Wesley, 1953.
- [17] A. R. Atilgan, S. R. Durrell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar, “Anisotropy of fluctuation dynamics of proteins with an elastic network model.,” *Biophys. J.*, vol. 80, pp. 505 – 515, 2001.
- [18] I. Bahar, A. R. Atilgan, and B. Erman, “Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential.,” *Folding and Design*, vol. 2, pp. 173 – 181, 1997.
- [19] I. Bahar, A. R. Atilgan, M. C. Demirel, and B. Erman, “Vibrational dynamics of proteins: Significance of slow and fast modes in relation to function and stability.,” *Phys. Rev. Lett*, vol. 80, pp. 2733 – 2736, 1998.
- [20] L. W. Yang and C. P. Chng, “Coarse-grained models reveal functional dynamics—I. elastic network models—theories, comparisons and perspectives.,” *Bioinformatics and Biology Insights*, vol. 2, pp. 25 – 45, 2008.
- [21] K. L. Xia, K. Opron, and G. W. Wei, “Multiscale multiphysics and multidomain models — Flexibility and rigidity,” *Journal of Chemical Physics*, vol. 139, p. 194109, 2013.
- [22] K. Opron, K. L. Xia, Z. Burton, and G. W. Wei, “Flexibility-rigidity index for protein-nucleic acid flexibility and fluctuation analysis,” *Journal of Computational Chemistry*, vol. 37, pp. 1283–1295, 2016.
- [23] D. D. Nguyen, K. L. Xia, and G. W. Wei, “Generalized flexibility-rigidity index,” *Journal of Chemical Physics*, vol. 144, p. 234106, 2016.
- [24] K. Opron, K. L. Xia, and G. W. Wei, “Communication: Capturing protein multiscale thermal fluctuations,” *Journal of Chemical Physics*, vol. 142, no. 211101, 2015.
- [25] K. L. Xia, K. Opron, and G. W. Wei, “Multiscale Gaussian network model (mGNM) and multiscale anisotropic network model (mANM),” *Journal of Chemical Physics*, vol. 143, p. 204106, 2015.

- [26] J. A. McCammon, B. R. Gelin, and M. Karplus, “Dynamics of folded proteins,” *Nature*, vol. 267, pp. 585–590, 1977.
- [27] M. Newman, *Networks: an introduction*. Oxford university press, 2010.
- [28] C. Ambedkar, “Application of centrality measures in the identification of critical genes in diabetes mellitus,” *Bioinformatics*, vol. 11, no. 2, pp. 90–5, 2015.
- [29] W. Y. G. Y. . L. Y. Gao, S., “Understanding urban traffic-flow characteristics: A rethinking of betweenness centrality,” *Environment and Planning B: Planning and Design*, vol. 40, no. 1, p. 135153, 2013.
- [30] A. Bavelas, “Communication patterns in task-oriented groups,” *The Journal of the Acoustical Society of America*, vol. 22, no. 6, pp. 725–730, 1950.
- [31] A. Dekker, “Conceptual distance in social network analysis,” *Journal of Social Structure (JOSS)*, vol. 6, 2005.
- [32] A. Zomorodian and G. Carlsson, “Computing persistent homology,” *Discrete Comput. Geom.*, vol. 33, pp. 249–274, 2005.
- [33] T. Schlick and W. K. Olson, “Trefoil knotting revealed by molecular dynamics simulations of supercoiled DNA,” *Science*, vol. 257, no. 5073, pp. 1110–1115, 1992.
- [34] D. W. Sumners, “Knot theory and DNA,” in *Proceedings of Symposia in Applied Mathematics*, vol. 45, pp. 39–72, 1992.
- [35] I. K. Darcy and M. Vazquez, “Determining the topology of stable protein-DNA complexes,” *Biochemical Society Transactions*, vol. 41, pp. 601–605, 2013.
- [36] C. Heitsch and S. Poznanovic, “Combinatorial insights into rna secondary structure, in N. Jonoska and M. Saito, editors,” *Discrete and Topological Models in Molecular Biology*, vol. Chapter 7, pp. 145–166, 2014.
- [37] O. N. A. Demerdash, M. D. Daily, and J. C. Mitchell, “Structure-based predictive models for allosteric hot spots,” *PLOS Computational Biology*, vol. 5, p. e1000531, 2009.
- [38] B. DasGupta and J. Liang, *Models and Algorithms for Biomolecules and Molecular Networks*. John Wiley & Sons, 2016.
- [39] X. Shi and P. Koehl, “Geometry and topology for modeling biomolecular surfaces,” *Far East J. Applied Math.*, vol. 50, pp. 1–34, 2011.
- [40] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer, “Stability of persistence diagrams,” *Discrete & Computational Geometry*, vol. 37, no. 1, pp. 103–120, 2007.

- [41] D. Cohen-Steiner, H. Edelsbrunner, J. Harer, and Y. Mileyko, “Lipschitz functions have L_p -stable persistence,” *Foundations of computational mathematics*, vol. 10, no. 2, pp. 127–139, 2010.
- [42] Z. X. Cang and G. W. Wei, “Analysis and prediction of protein folding energy changes upon mutation by element specific persistent homology,” *Bioinformatics*, vol. 33, pp. 3549–3557, 2017.
- [43] Z. X. Cang, L. Mu, and G. W. Wei, “Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening,” *PLOS Computational Biology*, vol. 14(1), pp. e1005929, <https://doi.org/10.1371/journal.pcbi.1005929>, 2018.
- [44] K. L. Xia and G. W. Wei, “Persistent homology analysis of protein structure, flexibility and folding,” *International Journal for Numerical Methods in Biomedical Engineering*, vol. 30, pp. 814–844, 2014.
- [45] M. W. Wolpert, D.H., “No free lunch theorems for optimization,” *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 67, 1997.
- [46] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [47] T. K. Ho, “The random subspace method for constructing decision forests,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 20, no. 8, pp. 832–844, 1998.
- [48] J. H. Friedman, “Greedy function approximation: a gradient boosting machine,” *Annals of statistics*, pp. 1189–1232, 2001.
- [49] J. H. Friedman, “Stochastic gradient boosting,” *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367–378, 2002.
- [50] B. Brooks and M. Karplus, “Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor,” *Proceedings of the National Academy of Sciences*, vol. 80, no. 21, pp. 6571–6575, 1983.
- [51] K. L. Xia and G. W. Wei, “A stochastic model for protein flexibility analysis,” *Physical Review E*, vol. 88, p. 062709, 2013.
- [52] D. Bramer and G. W. Wei, “Weighted multiscale colored graphs for protein flexibility and rigidity analysis,” *Journal of Chemical Physics*, vol. 148, p. 054103, 2018.
- [53] D. Bramer and G. W. Wei, “Blind prediction of protein b-factor and flexibility,” *Journal of Chemical Physics*, vol. 149, p. 021837, 2018.