Summer 8-9-2017

Machine Learning based Protein Sequence to (un)Structure Mapping and Interaction Prediction

Sumaiya Iqbal
University of New Orleans, New Orleans, siqbal1@uno.edu

Machine Learning based Protein Sequence to (un)Structure Mapping and Interaction Prediction

A Dissertation

Submitted to the Graduate Faculty of the
University of New Orleans
in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy
in
Engineering and Applied Science
Computer Science

by

Sumaiya Iqbal

B.Sc. Bangladesh University of Engineering and Technology, 2009

M.Sc. Bangladesh University of Engineering and Technology, 2012

August, 2017

Dedicated to my mother.

Acknowledgements

First and foremost, I would like to thank my advisor Dr. Md Tamjidul Hoque for his great guidance and timely support over the past four years. I am greatly thankful that he would take me as a student, teach me the research methodology, guide me in choosing interesting and influential research topics, and encourage me when I am stuck. Besides, I thank Dr. Hoque for giving me the freedom to discover and explore new subjects in machine learning, scientific computing, and computational biology.

My other committee members, Christopher Summa, Shengru Tu, Wendy Schluchter and Huimin Chen, have also been very supportive. I thank all these amazingly helpful professors for their invaluable assistance, feedback, and patience at all stages of this thesis. Their criticisms, comments, and advice are critical in making this dissertation more accurate and complete. In particular, Christopher Summa has been a warm supporter on my endeavor in bioinformatics and computational biology, and a source of new problems and objective opinions from the non-machine-learning community.

I would also like to thank the University of New Orleans for providing me an excellent environment for research and the financial support. In addition, special thanks to my friends and colleagues at UNO with whom I have had the pleasure of working over the years. These include Md Nasrul Islam, Avdesh Mishra, Aisha Ali-Gombe, Denson Smith, Glenn Robert McLellan, and all the members of the Bioinformatics and Machine Learning group. Their friendship and their help have brought me incredible joy during my UNO days.

Last, but definitely not the least, I would like to express my gratitude to my family for their love and support. I must mention my husband, Md Nasrul Islam, for bearing with my countless weekends and late night studies, for listening to my wild ideas and endless details, and for the inspiration on my work with love, patience, and understanding. I thank my father, elder sister and younger brother giving me motivations to achieve higher. I want to hold my parents in high esteem for taking care of my newborn daughter for last three years. Finally, everything I have ever achieved has been possible because of my mother, Salma Iqbal, who taught me believing in myself and persevering my ambitions to the hardest. Her example is always with me.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

3DIGARS	3-Dimentional Ideal Gas Reference State based energy function
AA	Amino Acid
ACC	Balanced Accuracy
AR	Amyloidogenic Region
ASA	Accessible Surface Area
AUC/AUROC	Area Under the Curve
BG, MG	Bigram, Monogram
BLAST	Basic Local Alignment Search Tool
BRCT	BRCA1 C Terminus
CASP	Critical Assessment of Protein Structure Prediction
CC	Correlation Coefficient
Chromo	Chromatin organization modifier
CR	Contact Radius
CV	Cross Validation
DD	Disorder Dataset
DisPredict	Disorder Predictor
DisProt	Database of Protein Disorder
DNA/RNA	Deoxyribonucleic Acid/Ribonucleic Acid
DNN	Deep Neural Network
DSSP	Dictionary of Protein Secondary Structure
ELM	Eukaryotic Linear Motif
ET	Extremely randomized Tree
FASTA	Standard format of protein and nucleic acid sequence
FDT	Frequency Distribution Table
FHA	Fork Head Associated
GA	Genetic Algorithm
GBC	Gradient Boosting Classifier
HT	High Throughput

IDEAL	Intrinsically Disordered Proteins with Extensive Annotations and Literature
IDP/IDR	Intrinsically Disordered Protein/ Intrinsically Disordered Regions in protein
KNN	K Nearest Neighbor classifier
LogReg	Logistic Regression
MAE	Mean Absolute Error
MBT	Malignant Brain Tumor
MCC	Mathews Correlation Coefficient
MHC	Major Histocompatibility Complex
MxD	Mixed Disorder
NCBI	National Center for Biotechnology Information
NMR	Nuclear magnetic Resonance
NR Database	Non-Redundant database
PBRpredict	Peptide-Binding Residue predictor
PCC	Persons Correlation Coefficient
PDB	Protein Data Bank
PID	Protein Interaction Domain
PP	Physical or Physicochemical Properties
PPI	Protein-Protein Interaction
PPV	Precision
PR	Precision-Recall
PRD	Peptide-Recognition Domain
PSBE	Position Specific Binding Energy
PSEE	Position Specific Estimated Energy
PSI-BLAST	Position Specific Iterated BLAST
PSSM	Position Specific Scoring Matrix
PTB	Phospho-Tyrosine Binding
PTM	Post-Translational Modification
RBF	Radial basis Function
RDF	Random Forest/Random decision Forest
RefSeq	NCBI reference Sequence database

REGAd^3p	Regularized Exact regression with degree 3 polynomial as kernel and Genetic Algorithm
ROC	Receiver Operating Characteristics
RSA	Relative Accessible Surface Area
SH2, SH3	Src Homology 2, 3
SL	Short and Long
sM	Stacked Model
SS	Secondary Structure
SSD	Secondary Structure Dataset
STDEV/std	Standard Deviation
SVM	Support Vector Machine
UniProt	Universal Protein Resource
UniProtKB	Uniprot Knowledgebase

Abstract

Proteins are the fundamental macromolecules within a cell that carry out most of the biological functions. The computational study of protein structure and its functions, using machine learning and data analytics, is elemental in advancing the life-science research due to the fast-growing biological data and the extensive complexities involved in their analyses towards discovering meaningful insights. Mapping of protein's primary sequence is not only limited to its structure, we extend that to its disordered component known as Intrinsically Disordered Proteins or Regions in proteins (IDPs/IDRs), and hence the involved dynamics, which help us explain complex interaction within a cell that is otherwise obscured. The objective of this dissertation is to develop machine learning based effective tools to predict disordered protein, its properties and dynamics, and interaction paradigm by systematically mining and analyzing large-scale biological data.

In this dissertation, we propose a robust framework to predict disordered proteins given only sequence information, using an optimized SVM with RBF kernel. Through appropriate reasoning, we highlight the structure-like behavior of IDPs in disease-associated complexes. Further, we develop a fast and effective predictor of Accessible Surface Area (ASA) of protein residues, a useful structural property that defines protein's exposure to partners, using regularized regression with 3rd-degree polynomial kernel function and genetic algorithm. As a key outcome of this research, we then introduce a novel method to extract position specific energy (PSEE) of protein residues by modeling the pairwise thermodynamic interactions and hydrophobic effect. PSEE is found to be an effective feature in identifying the enthalpy-gain of the folded state of a protein and otherwise the neutral state of the unstructured proteins. Moreover, we study the peptide-protein transient interactions that involve the induced folding of short peptides through disorder-to-order conformational changes to bind to an appropriate partner. A suite of predictors is developed to identify the residue-patterns of Peptide-Recognition Domains from protein sequence that can recognize and bind to the peptide-motifs and phospho-peptides with post-translational-modifications (PTMs) of amino acid, responsible for critical human diseases, using the stacked generalization ensemble technique. The involved biologically relevant case-studies demonstrate possibilities of discovering new knowledge using the developed tools.

Machine Learning; Large-Scale Data Analysis; Bioinformatics; Intrinsically Disordered Protein; Predictor Framework; Protein-Protein Interaction

Chapter 1

Introduction

Proteins, made up of smaller units called amino acids, are the fundamental macromolecules within a cell that carry out most of the biological functions and regulations according to the information encoded in the genes. Proteins are responsible for nearly every task of cellular life, including cell shape and inner organization, product manufacturing, waste cleanup, and routine maintenance. Proteins also receive signals from outside the cell and mobilize an intracellular response. They are the workhorse molecules of the cell and perform diverse set of functionalities. Proteins act as enzymes that carry out almost all of the chemical reactions that take place in cells as well as assist with the formation of new molecules by reading the genetic information stored in deoxyribonucleic acid (DNA). Moreover, they may bind to specific foreign particles, such as viruses and bacteria to help protect the body as antibodies. Proteins may serve as hormones, which transmit signals to coordinate biological processes between different cells, tissues, and organs, and as transcription factors that guide the differentiation of the cell and its responses to signals, and participate in the formation of tissues and muscular fiber.

It is widely believed that the protein structures play key roles in determining their functions [1]. However, in the first place, the way to know the structure experimentally is extremely labor intensive, and sometimes it is even impossible to determine the structure of a protein. Besides, there exist protein sequences in nature that do not adopt well-defined stable three-dimensional (3D) structure under normal physiological environments *in vitro*, however actively participate in molecular recognition functions [2]. As the distinct sequence of amino acids encode all the information about the three-dimensional conformation to express its functions [3], it becomes essential, as an alternative way, to design effective computational methods that can map the protein sequence to its structural properties, which is one of the major focus of the thesis. The research carried out in this thesis, not only effectively maps protein's primary sequence to its structure, but also identifies the disordered (or unstructured) component, which helps explain complex interaction within a cell that was otherwise obscured.

Proteins are often described as the building block of smaller substructures in a hierarchical manner. Based on the various level of conformational complexities, the proteins are defined in four different levels:

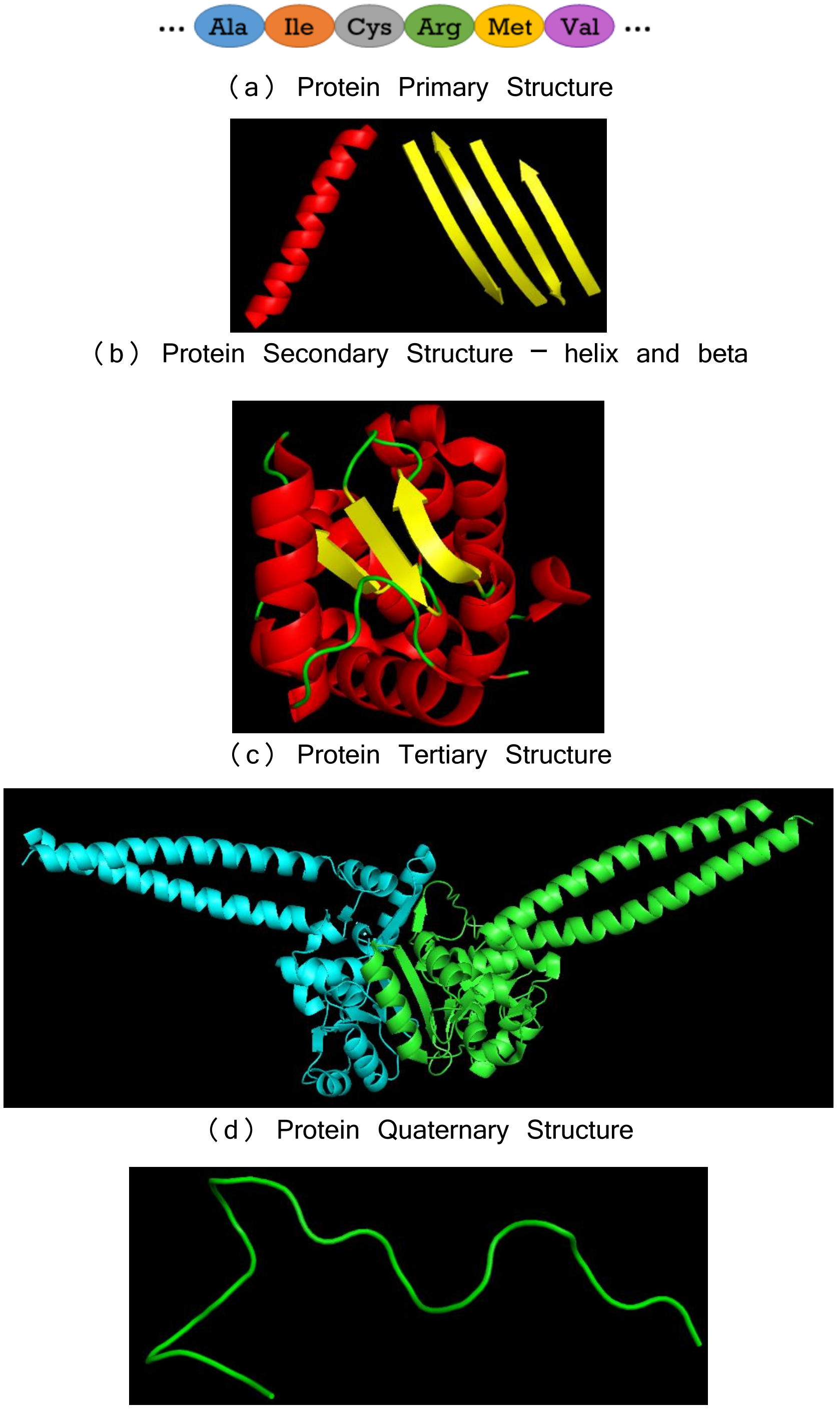
(1) *Primary structure*: It is the simplest level of protein structure and is the linear sequence of amino acid residues in a polypeptide chain. The sequence of a protein is encoded by the base-pair pattern of the gene of a DNA. A single nucleotide change in the DNA sequence may lead to a change in the amino acid sequence of the protein, and subsequently may alter the structure and function of the corresponding protein.

(2) *Secondary structure*: The local conformation of protein structures is determined by the pattern of hydrogen bonds in the biopolymer. There are three types of major secondary structures, known as alpha-helix, beta-pleated sheet sand coils (or loops), having different pattern of hydrogen bond between carbonyl and amino groups.

(3) *Tertiary structure*: This is the overall three-dimensional structure of a polypeptide. The formation is primarily due to the interactions between the side chains (known as R groups) of the amino acids that make up the protein. Side chain interactions that contribute to tertiary structure include hydrogen bonding, ionic bonding, dipole-dipole interactions, and van der Walls forces. Moreover, a net force that determines the core of the 3D structure is the hydrophobic effect by which hydrophobic R groups cluster together on the inside of the protein, leaving hydrophilic amino acids on the outside to interact with surrounding water (solvent) molecules.

(4) *Quaternary structure*: It is made up of multiple polypeptide chains, formed as a result of hydrogen bonds between multiple proteins as subunits. The interaction between multiple chains in a complex is the primary determinant of the signal transmission and reception within a cell.

However, there is another important class of proteins that do not adopt a stable structure in vitro, called Intrinsically Disordered Proteins. The critical property of disordered proteins or regions in proteins is that they can undergo conformational changes in the presence of an appropriate binding partner. Thus, in the bound state, disordered protein regions can transiently interaction with globular partner proteins and can participate in crucial functions related to pathogenesis. **Fig 1** shows a sample illustration of the hierarchy of protein structures as well as the unstructured or disordered state of proteins.



（a） Protein Primary Structure

（b） Protein Secondary Structure – helix and beta

（c） Protein Tertiary Structure

（d） Protein Quaternary Structure

（e） Protein Disorder State

Fig 1. Different levels of protein structures and unstructured state. （a） Primary structure, linear sequence of amino acid residues. （b） Two secondary structure types, helix （*red*） and beta sheets （*yellow*）, created by different hydrogen bond pattern. （c） Tertiary structure, the tree dimensional state of protein （d） The quaternary structure, complex of two protein chains （*cyan* and *green*）. （c） The disordered state （random coil-like） of a protein.

In this thesis, we focus on predicting different structural properties of protein residues from protein sequence alone via machine learning approaches, ranging from protein disorder prediction and protein accessible surface area prediction to identification of binding regions in proteins that interact with other proteins (specifically, short peptides) in a complex. Another major contribution of this thesis is to extract energy-like quantities from protein sequence anole that can characterize the structural stability of protein residue, hence can serve as critical feature for protein structure and interaction prediction.

1.1 Thesis Overview

With the exponential growth of proteomic data and the enormous complexities involved in their modeling, bioinformatics becomes essential for the management and mining of biological data in modern biology, medicine and drug discovery. Development of computational tools demands expertise from several core dimensions of computer science discipline, such as *i*) Data Science in data collection, mining and preparation, *ii*) Scientific Computing to extract useful knowledge from large sets of data and mathematically quantify the knowledge as characteristics features, *iii*) Machine Learning to develop of novel algorithms to model the data using features, and *iv*) Statistical and Probabilistic Analysis to empirically evaluate the model by comparative analysis and visualize the outputs. Over the course of this thesis, we have developed and implemented several tools to predict structural properties of proteins from its sequence using the above-listed areas of expertise.

1.2.1 Statement of Research Problem

The three-dimensional structures of proteins have the major association with their functional activities [1, 4]. Proteins may misfold when exposed to extreme conditions, however, many protein regions and some entire proteins do not adopt well-defined three-dimensional (3D) structures in an isolated state and under physiological condition [5-7]. These proteins or partial regions of proteins are called intrinsically disordered proteins (IDPs) or disordered regions in proteins (IDRs), respectively, also known as natively unstructured, denatured or unfolded. Intrinsically disordered proteins (or regions) undergo several conformational changes. The coordinates of their backbone atoms have no specific equilibrium states and can vary largely due to variable physiological conditions, and thus adopt dynamic structural ensembles. Recognition of these protein disordered regions is important for appropriate protein structure prediction, disease causing protein identification, proper annotation of function, induced folding and binding region prediction. However, due to highly flexible characteristics of the residues of IDRs or, IDPs [8]), experimentally verified annotation of intrinsic disorder is growing slowly. Thus, to keep pace with this faster growth of the protein database, effective computational methods for correct identification of disordered residues in IDPs or, IDRs become indispensable.

It is found that the primary protein sequence alone has the essential information needed to determine its corresponding secondary and tertiary structures [3]. Therefore, the three-dimensional structures for proteins can be determined by their one-dimensional sequence of amino acid residues, called *ab initio* protein structure prediction, which is challenging as it requires an efficient sampling algorithm to search in astronomically large conformational space and an accurate energy function to rank the protein structures and guide the conformational search. Thus, development of energy functions and search algorithms are highly demanding in the research area of proteomics.

The thermodynamic hypothesis of Anfinsen [3] explains that a protein in its natively stable structure, gains the lowest free energy. The structural stability of proteins requires large number of inter-residual interactions that contributes to gain in energy, required for protein folding. Therefore, a structured protein usually stay in a favorable (negative) energy state, while an unstable protein cannot gain favorable energy. While most of the available energy functions are based on the structural information, extraction of energy score from sequence only will have higher implication as it will be a useful feature in sequence-based structure prediction.

Both the protein structure prediction and structural state (order or, disorder) identification problems are highly complex but crucial, thus it is essential to use the outputs of many smaller sub-problems to solve the ultimate big problems. A feature that can map one dimensional sequence information into three-dimensional information and guide a machine learning algorithm to learn about the states to be predicted is crucial in developing predictive tools for this research problem. These smaller sub-problems include secondary (SS) prediction, Accessible Surface Area (ASA) prediction, backbone torsion angle (ϕ and ψ) prediction, and residue exposure. An accurate prediction of these structure properties from protein sequence alone has wide application in the field of bioinformatics and computational biology.

IDPs and IDRs have interesting characteristics of going through disorder-to-order transitions, and interacting with multiple partners to fold into different conformations when bound to different partners. Protein with peptide-recognition domains (PRDs) can recognize short peptide motifs that are usually present with IDPs/IDRs and can promote induced-folding of the peptides in disordered regions. It is crucial and challenging to computationally identify the peptide-binding regions in proteins that can promote coupled-binding with peptides as the tools can be utilized to assemble potential peptide-protein interactome. Additionally, identification of the residues of peptide motifs that primarily contribute in the enthalpy-gain necessary for the induced-binding and to investigate their complex biological functions in critical human disease and drug discovery, offer an essential research dimension in the study of recent bioinformatics and computational biology.

1.1.1 Contribution of The Thesis

Given a primary sequence of protein as input, prediction of structural descriptors or properties of protein, such as its secondary structure, tertiary structure, accessible surface area, torsion angles and flexibilities, thermal factor, contact map, and state of interaction-energy using computational methods, has further implications in the study of the proteins' functions. The computational tools can recognize patterns within inscrutable datasets, and can generate predictive-solutions fast for these challenging problems with reasonable accuracy, thus became an emerging research area in bioinformatics. Problems in computational and systems biology further vary from understanding sequence data to the analysis of protein shapes and protein classification to well-segregate and better understand their functions.

For some problems, the need of these computational efforts are essential. For instance, to understand the functions of proteins that are *Intrinsically Disordered Proteins* (IDPs) or have *Disordered Regions* (IDRs) a computational model can help capture the dynamics which would otherwise unmanageable to surmise. IDPs/IDRs do not adopt well-defined structure; however, they can change their states and fold through binding, and can perform important biological functions. Therefore, experimental investigation of IDPs/IDRs can reveal little information about their possible structures and functionalities. On the other hand, computational tools can provide a supplementary way for large-scale IDPs/IDRs analysis. Besides, the ultimate goal in the description of a protein is essentially to determine its structural properties as well as to determine the state of interactions with other proteins to perform function within a complex cellular network of living cell.

The objective of this dissertation is to develop effective *in silico* methods and tools to predict disordered protein and it dynamics, interaction paradigm by systematically mining and analyzing large-scale biological data. The involved case studies demonstrate possibilities of discovering new knowledge using the developed tools. The key outcome includes prediction of intrinsically disordered proteins (IDPs), prediction of protein accessible surface area (ASA), extraction of position specific energy (PSEE) of protein residues to score their stability, development of novel genetic algorithm variants for numerical optimization including protein conformational sampling, and identification of peptide-binding region of proteins that can recognize post translational modifications (PTMs) of amino acid, responsible for critical human diseases, within peptide motifs.

We started our work on the above-mentioned research problems from a hypothesis, which was then evolved through theoretical modeling and logical analysis, and eventually proven effective by empirical modeling and simulations. Therefore, the research methodologies described in this thesis is a fusion of theory, model and method development, evaluations and useful applications of the developed methods. Further, the tools developed under this dissertation-works are established as standalone software and has been published online to be utilized by broader scientific community.

1.2 Technical Results of the Thesis

1.2.1 An Optimized SVM for Protein Disorder Prediction

We have developed a framework to predict *intrinsically disordered proteins* (IDPs), **DisPredict** [9-11]. In this research, we performed large scale proteomic data collection, purification and analysis from multiple sources such as PDB, DisProt and IDEAL. To develop the predictors, we exercised machine learning algorithms, such as *Support Vector Machine* (SVM) with Radial Basis Function (RBF) kernel and *Deep Neural network* (DNN). The final SVM-based predictor was optimized, specifically the cost of SVM and the mode of RBF, using grid search. Such optimized parameter set made the predictor competitive. Moreover, we used three new features, Monogram and Bigram, giving high-dimensional evolutionary profile in DisPredict to predict disorder for the first time.

We developed a residue-wise potential score (PSEE) that can be computed from protein sequence alone and can be utilized for structure prediction from sequence. Towards an application of IDPs/IDRs prediction using the PSEE feature, we developed DisPredict2 [12] using RBF kernel SVM, in which we included PSEE in the feature space and optimized the threshold to segregate the disordered and ordered residues. **DisPredict2** performed very well in comparison to several other state-of-the-art predictors. Both DisPredict and DisPredict2 are implemented as a standalone tool in C, and are freely available from GitHub repositories[1,2] and Bioinformatics and Machine Learning (BML) lab website[3] of Computer Science department, UNO.

[1] DisPredict 1.0: https://github.com/tamjidul/DisPredict_v1.0
[2] https://github.com/tamjidul/DisPredict2_PSEE
[3] BML Lab: http://biomall.cs.uno.edu/software/

1.2.2 Extraction of Energy Scores from Protein Sequence

Computational tools for existing protein structure prediction problems require features that can capture the complexity of molecular level interactions. With a view to doing this, I proposed a novel approach to quantify *position specific estimated energy* (**PSEE**) [11, 13] of a residue using the pairwise thermodynamic interaction energy and solvent accessibility of the residue in local neighborhood. Here, the pairwise interaction captures the sequential environment, whereas the predicted solvent accessibility, which is eventually used to compute relative burial of a residue, includes the hydrophobic effect and captures the respective structural environment in PSEE. It has been verified empirically that PSEE can effectively classify disorder versus ordered residues, can segregate different secondary structure type residues by computing the constituent energies, and the PSEE value for each amino acid strongly correlates with the hydrophobicity value of the corresponding amino acid.

We have further utilized PSEE to compute residue-wise binding energy, *position specific binding energy* (**PSBE**) [14] from sequence alone. We performed alanine scanning on protein scanning to recomputed PSEE and the induced gap is formulated as PSBE. The PSBE was found effective in identifying the amino acid residues that gives higher contribution in binding energy.

1.2.3 A Reinforced Regression for Accessible Surface Area Prediction

We have developed a predictor of *accessible surface area* (ASA) of protein residues as real value. In this research work, we developed a new predictor paradigm, namely **REGAd^{3}p** [15], for real value prediction through Regularized Exact regression and Genetic Algorithm (GA). GA was used to optimize both Mean Absolute Error (MAE) and Pearson Correlation Coefficient (PCC). Further, the kernel of the exact regression was extended to degree 3 polynomial as this kernel was found to be the best to predict ASA while testing with large datasets collected from PDB. However, the framework is general for any real-value prediction work and the kernel can easily be tuned for a particular application. The predictor is developed in *C* programming language and it is available online[4].

Further, we have applied my tool in several other applications of bioinformatics. I modeled the error between actual and predicted ASA in terms of Energy to discriminate native proteins from their decoys. We combined this ASA based energy linearly with the components of an existing energy function, 3DIGARS [16], using Genetic Algorithm to develop two improved versions subsequently, **3DIGARS2.0**

[4] REGAd^{3}p: http://cs.uno.edu/~tamjid/Software/REGAd3p/REGAd3p.tar.gz

[17] and **3DIGARS3.0** [18]. I have further utilized my own ASA predictor to quantify the relative exposure (or burial) component in PSEE.

1.2.4 A Stacked Model for Peptide-Binding Residue Prediction

We have proposed a new computational tool to predict peptide-binding residues of receptor proteins in peptide-protein complex from sequence alone, named **PBRpredict**. A set of protein complexes with wide range of peptide-binding domains, such as MHC I and II, PDZ, SH2, SH3, WW, 14-3-3, Chromo and Bromo, Polo-Box, PTB, enzyme inhibitor, were collected from PDB and were annotated with interaction information based on atomic distances from peptide residues in the structure. Using a comprehensive set of sequence-based features including chemical and evolutionary profile, secondary structure, surface area and local backbone profile, flexibility and an energy based profile, we guide our predictor to learn about peptide-binding residues using model-stacking approach. To develop the model, we explored six different machine-learning algorithms as base learners, and those are support vector machine, gradient boosting, bagging, random forest, extra tree and k-nearest neighbor classifier. The outputs of the base learners were aggregated using a meta-learner, logistic regression (classifier).

We carried-out a rigorous performance evaluation using statistical metrics and case studies. After careful analysis of the prediction performance, we tuned the classification thresholds of the base-level and the meta-level learners of the stacking approach to trade-off between the true positive rate and false positive rate. Finally, we established three different PBRpredict models of a similar framework that apply different thresholds for segregating binding and non-binding residue under the name **PBRpredict-Suite**[5]. The results manifest that PBRpredict-Suite models provide well-balanced and biologically relevant outputs for proteins of different lengths and with a wide variety of PRDs. As an important outcome, one of the models recognized potential peptide-binding sites in the Gid4 subunit of the ubiquitin ligase GID in the yeast *Saccharomyces cerevisiae* for which no structure is available to date.

1.3 Thesis Organization

In this thesis, our primary goal is to develop *in silico* tools that can map protein sequence to its structural properties using machine learning algorithms. In addition, we target at extracting new biological knowledge

[5] PBRpredict-Suite: http://cs.uno.edu/~tamjid/Software/PBRpredict/pbrpredict.zip

that can serve as features to best capture the properties of protein structures so that we can contribute both computationally and biologically. Following this guideline, we organize the rest of the thesis as follows:

In Chapter 2, we describe the design and development of DisPredict, a predictor of intrinsically disordered protein. The framework is based on an optimized SVM with RBF kernel. In addition to the development of the predictor, we executed an analysis of the structural features of experimentally annotated disordered and ordered regions of proteins using feature correlation plot. We have also discussed that a post-processing of probabilities can further improve the prediction accuracy of the predictor.

In Chapter 3, we develop a general framework for real-value prediction using regularized exact regression with degree 3 polynomial as kernel, which is further optimized using genetic algorithm, named REGAd^{3}p and tuned specifically for protein accessible surface area (ASA) prediction. The framework is rigorously evaluated, compared and analyzed. Moreover, we modeled the actual and predicted ASA to develop an energy component, which is integrated in an existing energy function and found effective in improving the performance of the energy function.

In Chapter 4, we describe the formulation of a novel score, position specific estimated energy (PSEE), extracted from protein sequence information only. Computation of PSEE utilizes our previously developed tool, REGAd^{3}p to compute ASA from sequence. Further, we included contact energy in the computation of PSEE. PSEE is applied as an important feature application for disorder prediction and to develop an improved version of DisPredict, called DisPredcit2. DisPredcit2 is also described and evaluated in the same chapter.

In Chapter 5, we study the complex induced-binding between modular proteins and short peptide-motifs. Specifically, we develop a set of tools to predict the peptide-binding regions of proteins with peptide-recognition domains. The tools are found to be effective for a wide range of peptide-binding domains, which is evaluated through case studies. On the other hand, in this chapter, we have described the extraction of position specific binding energy (PSBE) from protein sequence that can approximate the higher binding energy contribution of peptide hot spots.

In Chapter 6, we conclude the thesis work, state its major contributions and provide brief future directions.

1.4 Related Publications

Parts of this thesis work have been published in journals, conferences and workshops of bioinformatics and computational biology, as well as under preparation for submission. Below is the list of the publications:

Related publications of Chapter 2:

- **Sumaiya Iqbal** and Md Tamjidul Hoque. DisPredict: A Predictor of Disordered Protein using Optimized RBF Kernel. *PLoS One*, 2015, 10(10), e0141551, DOI: 10. 1371/ journal. pone. 0141551.
- **Sumaiya Iqbal**, Denson Smith, Avdesh Mishra, Md Nasrul Islam, Md Tamjidul Hoque. Disordered Protein Prediction by Spiders. *CASP11 proceedings*, 2014, pp. 215–216.
- **Sumaiya Iqbal**, Md Nasrul Islam, Md Tamjidul Hoque. Improved Protein Disorder Predictor by Smoothing Output. *IEEE Conference on Computer & Information Technology (ICCIT)*, 2014, pp. 110–115, DOI: 10.1109/ICCITechn.2014.7073113, Dhaka, Bangladesh.

Related publications of Chapter 3:

- **Sumaiya Iqbal**, Avdesh Mishra and Md Tamjidul Hoque. Improved Prediction of Accessible Surface Area Results in Efficient Energy Function Application. *Journal of Theoretical Biology*, 2015, 380, pp. 380–391, DOI: 10.1016/j.jtbi.2015.06.012.
- Avdesh Mishra, **Sumaiya Iqbal** and Md Tamjidul Hoque. Discriminate Protein Decoys from Native by using a Scoring Function based on Ubiquitous Phi and Psi Angles Computed for All Atom. *Journal of Theoretical Biology*, 2016, 398, pp. 112–121, DOI: 10.1016/j.jtbi.2016.03.029.
- Avdesh Mishra, **Sumaiya Iqbal** and Md Tamjidul Hoque. An eclectic energy function to discriminate native from decoys. *The 4th Annual LA Conference on Computational Biology and Bioinformatics*, 2016, New Orleans, LA.

Related publications of Chapter 4:

- **Sumaiya Iqbal** and Md Tamjidul Hoque. Estimation of Position Specific Energy as a Feature of Protein Residues from Sequence alone for Structural Classification. PLoS One, **2016**, 11(9), pp. e0161452, DOI: 10.1371/journal.pone.0161452.
- **Sumaiya Iqbal** and Md Tamjidul Hoque. Estimation of free energy contribution of protein residues for structure prediction from sequence. *The Great Lakes Bioinformatics and the Canadian Computational Biology Conference (GLBIO/CCBC)*, **2016**, Toronto, Canada.

- **Sumaiya Iqbal**, Denson Smith and Md Tamjidul Hoque. Accurate identification of disordered protein residues using deep neural network. *The 4th Annual LA Conference on Computational Biology and Bioinformatics*, 2016, New Orleans, LA.

Related publications of Chapter 5:

- **Sumaiya Iqbal** and Md Tamjidul Hoque, A Study of Disorder-to-Order Transition by Characterizing the Binding Partners using a Statistical Potential, Biophysical Journal, vol. 112, p. 209a, **2017**, DOI: 10.1016/j.bpj.2016.11.1153.
- **Sumaiya Iqbal** and Md Tamjidul Hoque, Prediction of Peptide-Binding Residues of Receptor Proteins in a Complex, in *the 5th Annual Conference on Computational Biology and Bioinformatics*, New Orleans, LA, **2017**.
- **Sumaiya Iqbal**, Md Tamjidul Hoque, Modeling sequence Pattern of Peptide-Binding Domain Residue using Stacking (*submitted*), **2017**.
- **Sumaiya Iqbal,** Md Tamjidul Hoque, PBRpredict-Suite: Learning the Residue Pattern of Ppetide-Binding Domains from Sequence using Stacked Generalization (*submitted*), **2017**.

Miscellaneous:

- **Sumaiya Iqbal**, Tamjidul Hoque. hGRGA: A Scalable Genetic Algorithm with Homologous Gene Schema Replacement. *Swarm and Evolutionary Computation*, vol. 34, pp. 33 - 49, 2017.
- **Sumaiya Iqbal** and Md Tamjidul Hoque. A homologous gene replacement based genetic algorithm. *Genetic and Evolutionary Computation (GECCO)*, 2016, Denver, CO
- Tamjidul Hoque, **Sumaiya Iqbal**. Genetic Algorithm based Improved Sampling for Protein Structure Prediction. *International journal of Bio-inspired Computation*, vol. 9, pp. 129 - 140, 2017.

Chapter 2

DisPredict: A Predictor of Disordered Protein

— A Framework using optimized RBF kernel SVM

Intrinsically Disordered Proteins (or unstructured proteins) constitute a unique class of the protein kingdom, and have been recently recognized as a key player in the functional proteomics. Intrinsically disordered proteins or regions in proteins (IDPs/IDRs) lack rigid three-dimensional (3D) structure under physiological conditions *in vitro* [2]. However, IDPs, in full or in regions of the sequence, possess important biological functions despite their extremely flexible, essentially non-compact (or extended) structures. While the molecular recognition functions of IDPs/IDRs include pathways to carry out cell division, signaling, recognition and regulation [19], the structural heterogeneity of IDPs are highly linked to the amyloid aggregation that is involved in critical human diseases such as cancers, Parkinson's disease, Alzheimer's disease, type II diabetes and others [20]. Accurate identification of IDPs has significant implications in proper annotation of protein function and further understanding of drug design to combat disorder-associated diseases. Fast growing protein sequence repository [21] demands for high throughput computational techniques for identification of disordered residues from protein sequence, which is regarded as an imperative area of research in bioinformatics and computational biology.

In this chapter, we introduce our proposed disorder predictor framework, called *DisPredict* (**Dis**order **Predic**tor) [10] that classifies ordered and disordered residues from protein sequence alone. DisPredict employs a support vector machine with RBF kernel. With an optimal set of parameters for RBF kernel and a unique set of features including several novel features for reliable characterization of protein structure, DisPredict yields promising performance in both order versus disorder, i.e., binary classification as well as per-residue probability prediction, specifically in terms of Mathews Correlation Coefficient (MCC) and Area Under the receiver operating characteristics Curve (AUC).

DisPredict is evaluated using a 10-fold cross validation as well as tested with independent test datasets. The use of multiple data sources makes the predictor generic. Moreover, by comparison with other state-of-the-art approaches and case studies, DisPredict is found to be a useful tool with competitive performance. In addition to the development of the predictor, we performed an analysis of the structural features of experimentally annotated disordered and ordered regions of proteins using feature correlation plot. This experiment gave us insight of the collected overlapping annotation of the ordered and disordered segments of proteins in their feature space. The result of this experiment indicates the possible noise in the annotation of disordered and ordered residues in the available databases and instigates to formulate new characteristic feature to segregate disordered and ordered residues more clearly – in this direction, the outline of the rest if the chapter is given as follows.

- We start by giving the background information about intrinsically disordered proteins and their functions, and motivation behind developing a new predictor in Section 2.1.
- Next, we review existing disordered protein predictors in Section 2.2 along with our contribution.
- In Section 2.3, we describe the experimental materials, such as data sources, data collection and mining processes, input features used to train the predictor, and the criteria to evaluate and compare the predictor.
- Section 2.4 describes the first version of the predictor, DisPredict (version 1.0). In this thesis, by 'DisPredict1.0' or just by 'DisPredict', we refer to the first version of our disorder predictor.
- We described the performance evaluation related to optimal window size and parameter selection and for the comparison of the performance of DisPredict1.0 with existing predictors in Section 2.5.
- The analyses of the results and datasets as well as the feature correlation are presented in Section 2.6.
- In Section 2.7, we discuss an investigative strategy to make an improvement over DisPredict1.0. Keeping the similar framework and features to build the predictor, we included a post-processing of the output probabilities generated by DisPredci1.0 to develop DisPredic1.1, which improves the accuracy of prediction.
- Finally, we conclude in Section 2.8 with future research directions.

2.1 Background and Motivations

Proteins are the primary building block of the living cell. While proteins participate in almost all biological functions, abnormality in their functions can cause different pathological conditions. Therefore, to understand the mechanism of protein function is of utmost importance in the study of protein science. The well-known protein-structure function paradigm, i.e., "*sequence* → *structure* → *function*", states that the amino acid sequence of a protein specifies a unique (mostly) spatial structure, which represents a kinetically accessible and an energetically favorable state (local or, global minimum energy conformation of the protein). This conformation is usually refer to as the native state of a protein and is a precondition for a protein to be able to perform important biological functions [22]. Existing ample research-works support this view for more than 100 years since Emil Fischer proposed the lock-and-key hypothesis after performing the experimentation with enzyme and glucoside [1, 23], which essentially states that only a correctly shaped substrate (like a *key*) can fit into the key-hole (*active site*) of a particular enzyme (*lock*) to exert a chemical effect on each other. Later, Hsien Wu proposed the theory of protein denaturation [24, 25] that explains that proteins can lose its ordered state (i.e., structure), and then lose its ability to carry out functions due to their exposure to different non-physiological conditions, such as acid, urea or high temperature. In 1950s, Linus Pauling postulated the structural modeling of protein polypeptide chain [26, 27], which was followed by the first crystal structures of globular proteins (myglobin) [28, 29] and of an enzyme (lysozyme) [30]. Consequently, 3D structure was considered as an obligatory form for protein to function, until the early 1990.

The classic experiments of Anfinsen revealed that all the necessary information for the correct folding of protein is included in its amino acid sequence [3] and the kinetics behind the unfolding of proteins due to environmental perturbation and refolding after restoration of physiological state [31, 32]. Many proteins unfold in different non-physiochemical circumstances, such as extreme pH [33] whereas some proteins do not unfold in extreme conditions [34]. While the charge-charge repulsion is found to be the driving force behind the former phenomenon, later is governed by the strong hydrophobic interactions over charge repulsion [2]. Unlike these proteins, the *intrinsically disordered proteins* do not adopt ordered structure under physiological condition (neutral pH) *in vitro* or in the absence of a binding partner [2, 6, 7, 35, 36].

During the development of describing proteins or their regions that fail to form specific 3D structure and preliminarily to understand the functionalities of the flexible proteins, since 1940s [37-41], the disordered proteins were called using different terms, like floppy, pliable, rheomorphic, flexible, mobile, partially folded, vulnerable, chameleon, malleable, 4D, protein clouds, dancing proteins, proteins waiting for partners [42], and several names, such as combinations of "natively/naturally/inherently/intrinsically" with "unfolded/unstructured/disordered/denatured" [43]. However, it has been argued later that the native state of a protein is analogous to an active and functional form of a protein [42, 44]. On the other hand, disordered proteins (or regions) do not adopt well-defined structure in the normal physiological state, and therefore called "intrinsically" disordered proteins or IDPs [45] in recent years.

Although the major contribution of this thesis is to develop *in silico* framework for mining proteomic data to characterize disordered protein residues and segregate them from the ordered residues in a sequence, before digging into the details of prediction algorithms, in the following sections, we describe the basic characteristics of the IDPs, their abundance and functions, and the role of computational methods in the study of disordered protein.

2.1.1 Intrinsically Disordered Proteins – Types and Characteristics

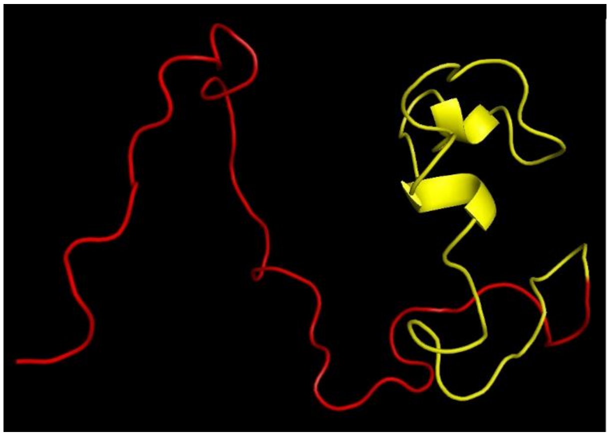
The intrinsically disordered regions (IDRs) of proteins exist as dynamic ensembles in which the coordinates of the atoms and the backbone Ramachandran angles vary largely over time. In the ordered regions, the coordinates of the atoms can fluctuate due to the random thermal change or conformational change of the local sequence neighborhood. However, this small-amplitude motions of the Ramachandran angles of the ordered residues can be characterized by the equilibrium positions defined by the time-averaged values. On the other hand, the atom positions and dihedral angles of the disordered residues cannot be characterized by an equilibrium state around which the residues stay most of the time, rather the IDRs undergo heterogeneous conformational changes that are random.

The residues in IDPs or IDRs possess several characteristics in terms of structural conformation and sequence composition [6, 8, 36, 46]. In [2], Uversky *et al*. discussed that the combination of low mean hydrophobicity and relatively large net charge is an important prerequisite for the absence of regular structure in proteins under physiologic conditions, which was further verified using charge-hydropathy (CH) plot [7, 36], showing a linear boundary line to separate ordered and disordered proteins based on their mean net charge and mean hydrophobicity, in early 2000s. The IDPs or IDRs are also found to have a compositional bias in their amino-acid residues; for example, they are depleted in Trp, (W) Tyr (Y), Phe (F), Ile (I), Leu (L), Val (V), Cys (C) and Asn (N), so-called *order-promoting* amino acids, whereas they are enriched in Ala, Arg (R), Gly (G), Gln (Q), Ser (S), Glu (E), Lys (K) and Pro (P), so-called as *disorder-promoting* residues. The general properties of IDPs/IDRs are listed as below.

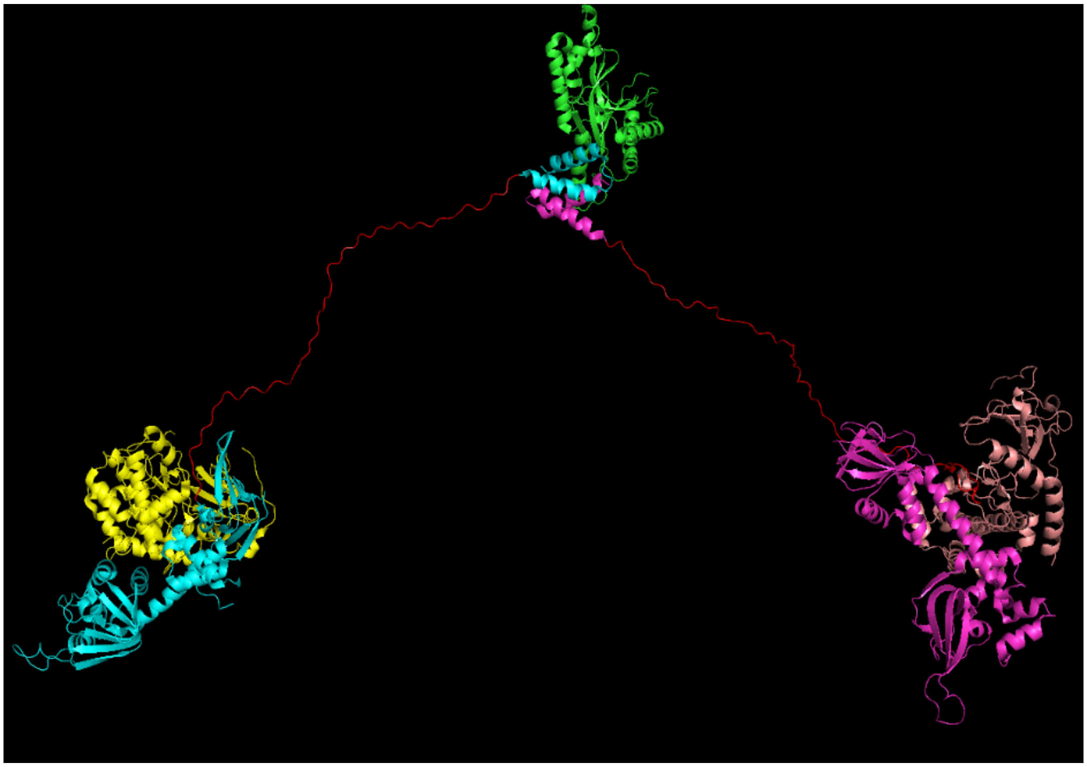
- **Sequence based properties**
    - Presence of charged amino acid residues (especially negative)
    - Low content of hydrophobic amino acids residues
    - Low sequence complexity (use of reduced alphabet out of 20 amino acid)
- **Structure based properties**
    - Low compactness
    - Absence of globularity

- Low content of secondary structure
- High amount of flexibility (no specific equilibrium state of the backbone atoms)

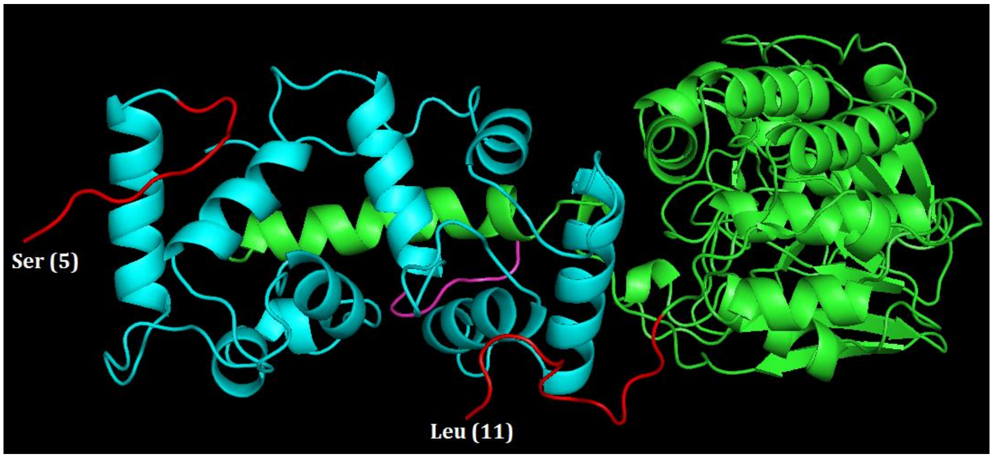
IDPs can adopt different conformations under various environmental conditions, such as effects of temperature change, pH change, and presence of ions and ligands [36]. Various degree of disorder has been observed in nature [44]. **Fig 2** portrays three samples of intrinsic disorder in monomer and complex structures.



（a） PDB ID：2JU4



（b） PDB ID：3J4Q



(c) PDB ID: 5SVE

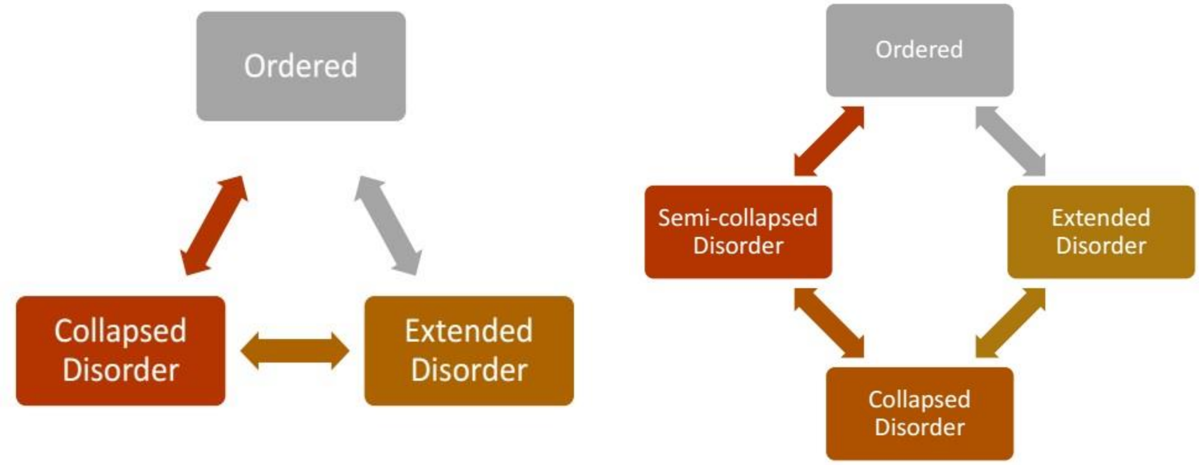
Fig 2. Intrinsically disordered proteins or regions in proteins. (a) PDB ID: 2JU4 [47], NMR structure of intrinsically disordered gamma-subunit (PDEgamma) of cGMP phosphodiesterase. It has extended-disordered N-terminal region (*red*), whereas residues 46 – 87 (*yellow*) shows loose structural features, bound to alpha(t) in the transition-state complex. (b) PDB ID: 3J4Q [48], intrinsic disorder within an AKAP-protein kinase A complex. The complex has disorder linker (*red*) between each PKA regulatory subunit. (b) PDB ID: 5SVE [49], disordered regions in human Calcineurin interaction network with LxVP short linear motif (*pink*). The complex has Serine/threonine-protein phosphatase 2B catalytic subunit as chain A (*green*) and Calcineurin subunit as chain B (*cyan*). Both chains have disordered N-terminals (*red*). For chain A (green), structure of residues 1 – 10 and that of residues 1 – 4 for chain B are missing in the electron density due to their flexibility. We used PyMOL [50] to view the structures and DSSP to assign the secondary structures [51].

IDPs challenge the classical structure-to-function relationship of protein [8, 20, 46, 52]. Disordered proteins, having no rigid structure can show larger plasticity, can interact with different targets, such as ligands, small molecules, substrates, cofactors, other proteins, peptides, membranes etc. and participate in most of the key biological and disease-associated processes [5, 46, 49]. Some disordered regions are not known to bind to any partner, but they still carry out important functions such as providing flexible linkers between structured domains (*see* **Fig 2 (b)**, red colored regions) or providing flexible tails that regulate the structured domains [35, 46].

Together with ordered state of the proteins, two different protein structure–function paradigms were proposed considering the disordered states and the transitions between these states, called Protein Trinity [53] by Dunker *et al*. (2001) and Protein Quartet [35] by Uversky (2002), shown in **Fig 3**. Various structural forms of IDPs, discussed in the literature, have been listed below.

- **Collapsed disorder** (*molten globules*) [6, 53, 54], which contain stable but highly dynamic side-chains and well-developed secondary structure elements [52, 53, 55-57].

- **Semi-collapsed disorder** (*pre-molten globules*) [56], like polyglutamine regions [58] and polar sequences [59, 60], which arise due to the presence of rapidly exchanging backbone side-chain hydrogen bonds that make the region fail to form specific secondary structure [35, 36, 44].
- **Extended disorder** (*random-coils*), like intrinsic coils [20, 53, 54], which are formed by the combination of low hydrophobicity and high net charge that result only marginal level of residual secondary structure [35, 36, 46].



（a） Protein Trinity

（b） Protein Quartet

Fig 3. Two protein structure–function paradigms, which emphasize the ordered state of a functional protein as well as consider three possible dynamic states occurred by intrinsic disorder phenomenon. （a） Protein trinity, includes two different states, fully extended （or random coil） and collapsed （or molten globule）, of protein disorder. （b） Protein quartet, includes an additional state of disorder, semi-collapsed or pre-molten globule like state.

2.1.2 Abundance of Protein Disorder

The disordered proteins or residues have been found abundant in nature. Approximately 70% of the structures released by Protein Data Bank (PDB) [61, 62] contain some disordered residues [63]. Significant proportion of some genomes (such as, Eukaryota) encode the proteins with regions of disordered residues [54]. In humans, roughly one-third of all proteins is intrinsically disordered, of which, approximately 50% of these proteins are more than 30 residues long, and 25% of them are fully disordered [64].

At proteome level, approximately 33% of eukaryotic proteins are found to have IDRs, having length greater than 30 residues and 19.6% of them have IDRs with greater than 50 residues long. For bacterial proteins, 4.2% and 1.6% of the them hold IDRs, having length greater than 30 and 50 residues, respectively [65-67]. Further, 6.8% of the archaea proteins have disordered residues [68].

2.1.3 Protein Disorderedness and Functions

Intrinsically disordered proteins participate in numerous biological functions by exhibiting a multitude of structural conformations and dynamics, such as cell cycle control and cellular signal transduction, transcriptional and translational regulation, membrane fusion and control pathways [5, 69, 70]. Disordered regions in proteins are found to be evolutionary conserved [36], which further confirms their intriguing role in biological processes. Moreover, IDPs are more frequent in eukaryotic genomes in comparison to bacteria and archaea, which supports the need of IDPs/IDRs recognition or prediction in signaling and regulation in nucleated cells [53, 69, 71, 72] as well as its potential involvement in human diseases.

Intrinsic disorder enables a number of capabilities of a protein [36, 44, 73, 74] which are crucial for molecular recognition, such as (*i*) separation of specificity and affinity due to the free energy penalty paid to fold disordered state; (*ii*) "***binding diversity***", by which a region can fold differently to be recognized by different shaped partners, (*iii*) "***binding commonality***", by which a sequence can fold differently but is recognized by a common surface of a partner; (*iv*) formation of large interaction surfaces; (*v*) faster rates of association and disassociation; (*vi*) reduced life-time *in vivo* and rapid turnover of regulatory molecules and so on. Besides molecular recognition, IDPs/IDRs participate in molecular assembly and protein modification [70, 75] via protein-protein, protein-nucleic acid and protein-ligand interactions [76-80].

In addition, being rich in binding sites for various partners, IDRs are found to be important loci for alternative splicing [81] and for enzyme-driven posttranslational modifications (PTM) such as phosphorylation, methylation, or acetylation [79]. Furthermore, intrinsic disorder plays a fundamental role in the functionality of proteins with PEST sequence, hub proteins, transcription factors, 14-3-3 protein and scaffold proteins [20, 82].

Disorder proteins are also associated with critical human diseases [82-85]. Structural disorder was confirmed and studied in great detail in many other important disease-associated proteins, such as p53, T protein, and cystic fibrosis transmembrane conductance regulator [83]. IDPs are involved in cancer, amyloidoses, cardiovascular diseases, neurodegenerative disorders (i.e. Alzheimer's diseases, Nieman-Pick disease type C, Down's syndrome, Parkinson's disease, Hallervorden-Spatz disease), genetic diseases, prion diseases, accelerated fibrillation, and protein deposition diseases as well as in drug development [19, 83, 86-88]. Thus, locating the disordered regions in a protein and identifying their dynamic conformations for better understanding of protein function was one of the most studied research area in protein science for the last two decades.

2.1.4 Role of *in silico* Disorder Prediction

Due to highly flexible and dynamic characteristics of the residues of IDRs or IDPs, experimentally verified annotation of intrinsic disorder becomes very complex. In the X-ray crystallography [89] experimentation, disordered residues are indicated by the lack coordinates in structure, often refer to as missing-residues. In NMR spectroscopy [90-93] experimentation, disordered residues exhibit high variability within the structural ensembles. Among other experimental approaches - near or far ultraviolet circular dichroism (CD) [94-97], Fourier transform infrared [2], various hydrodynamic techniques (small angle x-ray scattering (SAXS), small angle neutron scattering (SANS), sedimentation, and dynamic and static light scattering), electron microscopy or atomic force microscopy etc are examples of some techniques used [98, 99].

A curated database of disordered proteins, called DisProt [100] contains annotation for 694 protein sequences and 1,539 disordered regions in its version 6.02. Recently, DisProt 7.0 [101] is launched with 803 proteins and 2,167 regions with annotated disordered residues of which 69.3%, 19.4%, 9.3% and 1.9% proteins are from eukaryota, bacteria, viruses and archaea domain. The recently established IDEAL [102, 103] database also provides useful collection of IDPs including 838 protein entries in its current release by March, 2017. On the other hand, PDB [61] database provides access to find disordered regions in the solved secondary or tertiary structures, which incorporates 119,163 protein entries (accessed on March, 2017)[6]. To compare, the overall number of non-redundant collected protein sequences is 81,027,309 according to the most recent 81 release of RefSeq database [104].

To keep pace with this large-scale increase in protein database, effective computational methods for correct identification of the disordered residues in IDPs or, IDRs are necessary. The *in vitro* experiments to analyze protein structure are costly both in terms of money and time, and in addition reveals little information about the possible structures and functionalities of IDPs/IDRs. In contrast, *in silico* tools provide a rapid and supplementary way for large-scale proteome-wise IDPs/IDRs analysis and prediction. MobiDB [105, 106] is an exemplar database that collects consensus annotation of protein disorder from predictors.

[6] PDB statistics: http://www.rcsb.org/pdb/statistics/holdings.do

2.2 Review of Disorder Prediction

As the IDPs/IDRs differ dramatically from the ordered proteins in their amino acid sequences, possible development of successful predictors of protein disorder from its sequence made perfect sense. Thus, in this section, we provide a brief review the current development on disorder prediction techniques.

Based on a very small number of proteins, Williams [107] developed a predictor of intrinsic disorder based on the ratio of the number of charged residues and the number of hydrophobic residues in the protein sequence in 1979. This predictor was used to separate only two IDPs from a small set of ordered proteins and later found to be not effective in general [108]. However, this article can be considered as one of the pioneer works that attempts to identify IDPs based on amino acid compositions, was substantially different from those of proteins with 3-D structure.

Later, Dunker and Uversky and their coworkers independently developed predictors of IDPs [2, 109]. Since then, various prediction ideas and different computing techniques have been utilized to identify protein disorder. In the following section, we discuss these techniques by characterized them into three broad categories: (*i*) machine learning based methods, (*ii*) amino-acid composition and chemical property based methods, and (*iii*) methods that combine outputs of multiple predictors.

2.2.1 Predictors based on Machine Learning

Currently, machine learning is taking on a leading role in solving critical pattern classification tasks in an efficient manner, which is expanding rapidly in the field of biology with the vast amount of proteomics and genomics data being available. The available disorder prediction tools utilize pattern recognition methods, alone or in combination, such as Logistic Regression (LR), Discriminant Analysis (DA), Ordinary Least Squares (OLS), Artificial Neural Network (ANN) [110], Support Vector Machine (SVM) [111], Bayesian Classifier, Random Forest (RF) [112], K Nearest Neighbor technique etc. The underlying idea of these techniques is to train a machine learning algorithm that can capture patterns using a set of characterizing features so that the resulting model can predict those patterns in an independent test data set.

The predictors of PONDR® (Predictor of Natural Disordered Regions) series [62, 113-117], such as PONDR® VL-XT, VL3 (VL3-E, VL3-P, VL3-H), VSL (VSL1, VSL2) are some of the established software tools that use feed-forward neural networks to predict disordered region of different length or in different location of a sequence. Among other ANN based tools, DisEMBL [118] identifies three kinds of disorder, including loops/coils, hoot loops and residues with missing coordinates in X-ray crystallography (REMARK 465 of PDB structure file), RONN [119] uses bio-basis functional alignments and NORSnet

[120] was trained for long loop or extended disordered regions. DisPro [121] uses a one-dimensional recurrent neural network (1D-RNN), whereas ESpritz [122] employs a bi-directional recurrent neural network. SPINE-D employs a single neural-network to predict disorder regions [123], focusing the differences between long and short disorder regions. DNdisorder [124] employs boosted ensembles of deep network [125], which are similar to the regular neural network but contain more layers to predict disordered regions in proteins.

The DISOPRED [126] series includes three predictor of which, DISOPRED [127] applied neural network, DISOPRED2 [128] used support vector machine with linear kernel, and DISOPRED3 [129] adopted a combined approach using SVM, ANN and nearest neighbor classifier to predict disordered regions as well as peptide-binding sites in disordered regions. Spritz [130] is a server for predicting disordered IDRs in proteins using two different SVMs for short and long disordered regions.

The POODLE series include POODLE-L [131] that incorporates two-level SVMs for long disordered region prediction, POODLE-S [132] that uses seven SVMs to predict disorder in different region of the sequence, and POODLE-W [133] that applies Joachims' spectral graph transducer (SGT) [134], which is a semi-supervised learning technique unlike others and operates by constructing k-nearest neighbor (kNN) graph. DisPSSMP [135] is based on radial basis function network (RBFN) with inputs from position-specific scoring matrices (PSSM) and DisPSSMP2 classifier is a derivative of the former one using condensed PSSM, which was integrated in the web server called iPDA [136].

Weathers *et al*. [137] used SVM to analyze and propose that a reduced amino acid alphabet is sufficient to accurately identify IDPs; and RAPID is a support vector regression-based predictor [138]. A recursive maximum construct tree (RMCT) was used in IUP to recognize IDPs. Bayes [139] incorporates a Bayesian method to compute conditional probability for a sequence, given a disordered protein and OnD-CRFs [140] predicts the intrinsic disorder in proteins using *Conditional Random Fields* (CRFs), which is a discriminatively supervised machine-learning method.

There are some tools, which measure the disorder content of a protein sequence instead of binary classification of ordered and disordered residues. DisCon [141] quantifies the disorder content using ridge regression using weighted PSSM profile. SPA [142] incorporates a non-linear neural network classifier to predict disorder in short peptides in two-steps. PON-Diso [143] identifies disorder-related amino acid substitutions in sequence using random forest classifier.

2.2.2 Predictors based on Physicochemical Properties

There exist tools which are based on the relative composition and propensity of amino acids or, their physical, chemical and structural properties.

GlobPlot [144] and TopIDP [145] uses relative proportion of amino acid residues to predict IDP, whereas FoldUnfold [146] uses mean packing density as a characteristic measure for disorder. PreLink utilizes hydrophobic cluster content along with measure of compositional bias to identify IDRs in proteins [147]. The key idea of IUPred [148, 149] to identify disordered regions was that the inter-residual interactions are responsible for determining the structure of proteins.

FoldIndex [150] is used to compute the ratio of net charge with hydropathy locally using a sliding window to predict disorder. SEG [151] identifies low-complexity disordered segments using complexity measures (Wootton and Federhen equation). Ucon [152] uses predicted contact information to identify unstructured regions, while DISOclust [153] is based on the analysis of how disorder is related with protein folding and uses predicted three-dimensional structural characteristics. IsUnstruct [154] employs the Ising model to distinguish disordered from the ordered regions based on statistical physics.

2.2.3 Predictors based on Meta-approach

In meta-approach, several self-complementary disorder predictors are combined to generate a consensus prediction or weighted predictions.

PONDR-FIT [57] is a meta-predictor, which combines six individual predictors including PONDR® VLXT, VSL2, VL3, FoldIndex [150], IUPred [148, 149], and TopIDP [145]. The three predictors of PONDR series use artificial neural network. FoldIndex, IUPred, and TopIDP to form disorder or ordered regions based on relative propensity of amino acids. PrDos [71] consists of two predictors, one of which uses the alignment of homologs. PreDisOrder [155] is based on an ab initio prediction method (MULTICOM-CMFR) along with a consensus prediction method, MULTICOM [156]. POODLE-I [157] is a predictor based on the meta-approach that integrates the sub-components of the POODLE series.

MetaDisorder [158] incorpoartes 13 disorderd predictors, including DisEMBL [118], DISOPRED2 [126], DisPro [121], Globplot [144], iPDA [136], IUPred [148, 149], Pdisorder [158], Poodle-S [132], Poodle-L [131], PrDOS [71], Spritz [130], DisPSSMP [135], and RONN [119]. The results generated by these predictors are weighted by the accuracy of the methods to produce final prediction by MetaDisorder [158]. DisMeta [159] assembles eight primary sequence-based predictors including DISEMBL [118], DISOPRED2[126], DISpro [121], FoldIndex [150], IUPred [148, 149], RONN [119], and VSL2, and generate a consensus-based output.

CSpritz [160] is a combination of Spritz, Punch (an SVM-based predictor extending Spritz), and ESpritz [122]. The metaPrDOS [161] is composed of seven individual predictors which are as follows: PrDOS, DISOPRED2, DisEMBL, DISPROT, DISpro, IUPred, and POODLE-S, while MD [162] is another metapredictor composed of NORSnet, Ucon, PROFBval [163], DISOPRED2, IUPred, and FoldIndex.

MFDp [164] fuses three different methods that are complementary to each other, DISOPRED2, IUPred and DISOclust , and combines outputs from three SVMs with linear kernel. A further improved version, named MFDp2 [165] was proposed later, which is also a meta-server combining two methods including residues-level based MFDp and sequence-level based DISCon [141].

The computational methods provide fast supplementary knowledge about potential location of disorder in proteins to the proteomics research community involved in analyzing protein functions, and their possible involvement in human diseases and drug discovery processes. Due to its importance, the critical assessment of protein structure prediction, popularly known as CASP competitions [166-171] evaluate the performances of existing disorder predictors biennially starting since 2002. In these competitions, participants make predictions of disordered residues in proteins on their amino acid sequences, of which the structures are being determined, but before the structures are known. An independent group of researchers then compares the various predictions from many research groups with the observed structures.

2.2.4 Our Contributions

It is exciting that in many cases predictions of protein disorder have been used to guide *in vitro* experiments, which in turn have led to the discovery of increasing numbers of disordered proteins. This prediction-experiment-prediction loop is leading to further increases in the rates of discovery for IDPs [108]. With a view to add an advanced predictor with improved accuracy, we developed a new framework named DisPredict that incorporates two novel aspects; an optimized kernel support vector machine and a higher dimensional evolutionary profile as feature.

We have further developed a residue-wise potential score (PSEE) that can be computed from protein sequence alone to characterize the disordered state of protein residues. Towards an application of IDPs/IDRs prediction using the PSEE feature, we developed DisPredict2 (discussed in **Chapter 4**) under a similar framework of DisPredict which showed promising performance compared to the state-of-the-art predictors.

2.3 Experimental Materials

In this section, we describe the data-sources, collection of training and test datasets, and aggregation of input-features for DisPredict (version 1.0).

2.3.1 Disorder Data Sources

In the prior studies, DisProt [100] and PDB [61] are considered as the primary repositories of IDPs. Disorder regions are composed of residues with missing coordinates in structure solved by X-ray crystallography, whereas the residues show highly variable coordinates within ensemble solved by NMR. We selected two datasets which combine sequences from PDB having disordered residues without coordinates (recorded in REMARK 465) and sequences from DisProt, having curated annotations of disorder regions including properties such as short (≤ 30 residues) and long (> 30 residues) disordered regions, partial as well as fully ordered or disordered chains.

2.3.2 Datasets

We used two different datasets, MxD and SL, to train, test and cross-validate our proposed DisPredict. MxD and SL datasets were used to train two disorder predictors, MFDp [164] and SPINE-D [123], respectively. We collected and utilized these datasets to be able to consistently compare DisPredict with these two state-of-the-art predictors. We further used three independent test sets to evaluate the model. The datasets are available online[7].

2.3.2.1 *Training Sets*

Dataset SL477: SL477 dataset was prepared by the developers of SPINE-D predictor from the benchmark SL (Short Long) dataset [172]. The SL dataset encompasses short and long disordered regions as well as ordered regions. It was built by re-annotating the sequences extracted from DisProt to include reliable order and disorder contents. Among the annotated regions in the SL dataset, 50% of the regions are of the short-disordered category. The short regions in SL dataset are of length 20 residues or less [172]. It is important to incorporate this disorder annotation in a dataset since these short disordered regions are found functionally important as they obtain induced folding with the close proximity of appropriate partners. SL477 also includes very long disordered regions as well as completely disordered proteins, called intrinsically disordered proteins (IDPs).

[7] DisPredict dataset link: http://cs.uno.edu/~tamjid/Software/DisPredict/Training_and_Test_Data.zip

SL dataset is comprised of proteins with disorder regions annotated by NMR experimental method as well. To achieve a combination of sequences with low sequence identity, SL dataset's sequences were clustered and filtered using BLASTCLUST [173] which resulted in 477 chains with < 25% sequence identity between each pair. SL477 has total 215,343 residues, of which 56,887 (about 25%), 72,808 (about 34%) and 85,648 (about 40%) residues are annotated as disorder, order and unknown, respectively. Unknown residues are those which are marked unknown in the source datasets. We disregarded the residues with unknown annotation during both in training and in evaluating our proposed approach.

Dataset MxD444: The Mixed Disorder (MxD) dataset is a combination of protein sequences with disordered residues from both PDB and DisProt. Originally developed MxD dataset has 514 protein sequences including 205 chains from PDB and 309 chains from DisProt. We carried out further purification by removing sequences with unknown amino acid (X-tag) since they do not have specific physicochemical properties to get corresponding features in our methodology. This led to the MxD444 dataset, with 444 chains and 214,054 residues, that mixes 49,090 (about 23%) disordered residues and 164,964 (about 77%) ordered residues.

2.3.2.2 *Test Sets*

Dataset SL171: We executed one more round of filtration using BLASTCLUST [173] to generate an independent subset from SL477 which encompasses sequences with no more than 10% sequence identity with MxD444 dataset. It gave us an independent test set of 171 chains with 42,572 residues, named as SL171, which was used to evaluate the DisPredict model while trained on MxD444 dataset. Another distinction between our two test datasets is, MxD134 contains sequences with disordered regions defined by PDB. On the other hand, SL171 contains protein sequences with disorder annotation only from DisProt.

Dataset MxD134: We extracted an independent test dataset from MxD444 and named it MxD134. To generate MxD134, we combined the sequences of MxD444 with the second dataset utilized to train DisPredict, SL477. The MxD444 dataset was then filtered to remove sequences with sequence similarity greater than 10% to any sequence from SL477 dataset using BLASTCLUST [173], retrieving a set of 134 protein chains with 38,823 residues. MxD134 dataset was employed to evaluate our predictor while training was performed on SL477 dataset.

Dataset DD73: Further, we prepared a completely new dataset that is entirely independent of the training sets of DisPredict, SPINE-D [123] and MFDp [164]. We collected 48 new protein chains from DisProt [100] released after version 5.1 upto current version of 6.02. These protein sequences were combined with another 25 protein chains culled from PDB [61]. Protein chains were extracted from PDB x-ray structures with resolution ≤ 3.0 angstroms, length ≥ 50, and sequence identity cut-off of 30% and by choosing single chain proteins.

We randomly selected 25 chains from the output of this experiment so that no sequence is more than 25% similar with the training sequences. To have a proper combination of ordered and disordered proteins, we ensured that none of these 25 proteins can contain disordered residues except terminal regions. It provided us with 73 protein sequences which is a combination of 37 full disorder chains, 23 full ordered chains and 13 protein chains with disordered and ordered regions. We call this Disorder Dataset as DD73. DD73 dataset allows us to perform a robust comparison among DisPredict, SPINE-D [123] and MFDp [164], as it is independent of both SL and MxD dataset.

2.3.3 Input Features

Input features for DisPredict were carefully chosen to be able to include useful properties such as the sequence information, evolutionary information as well as the structural information (listed in **Table** 1).

Table 1. List of features used in DisPredict.

Feature Category	**Feature Count**
Amino Acid (AA)	1
Physicochemical Property (PP)	7
PSSM Profile (PSSM)	20
Secondary Structure Content (SS)	3
Accessible Surface Area (ASA)	1
Torsion Angle Fluctuation (ϕ, ψ)	2
Monogram (MG)	1
Bigram (BG)	20
Terminal Indicator (T)	1
Total	56

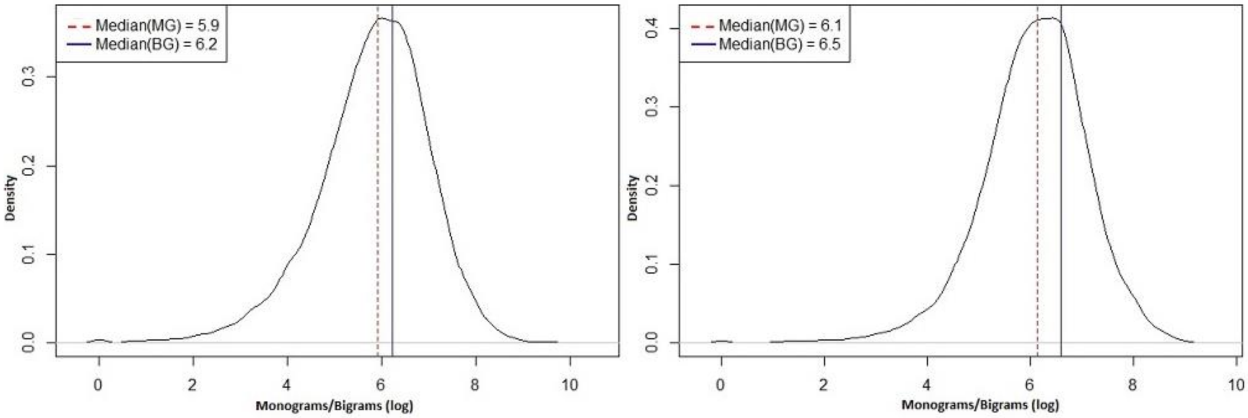
2.3.3.1 *Sequence Information*

Anfinsen’s dogma (also known as the *thermodynamic hypothesis*) of molecular biology suggests that all the necessary information for the correct folding of a protein is encoded in its primary amino acid sequence [3]. Further studies suggest that the misfolded regions or disordered regions of protein can also be predicted from its amino acid sequence [113, 127], as discussed in Section 2.1.1. The physicochemical properties of each amino acid, such as the steric parameter, polarizability, volume, isoelectric point and etc., are also correlated with the length of disordered regions, as the short disordered regions are mainly negatively charged while the long disordered regions are nearly neutral [8, 162]. These observations motivated us to use amino acid type (**AA** in **Table 1**, indicated by one numerical value out of twenty) and seven physical properties (**PP** in **Table 1**) [174] as features to identify disordered residues.

To distinguish the terminal residues for their position specific disorder like behavior, we included terminal indicator feature (**T** in **Table 1**) by encoding five residues of N-terminal as (-1.0, -0.8, -0.6, -0.4, -0.2) and C-terminal as (+1.0, +0.8, +0.6, +0.4, +0.2) respectively, whereas rest of the residues were labeled 0.0.

2.3.3.2 *Evolutionary Information*

Disordered regions and their related functions are conserved within the sequence during evolution [135], thus we considered position specific scoring matrix (**PSSM** in **Table 1**) as input features to capture evolutionary information. PSSM (size: sequence length × 20) was generated for each sequence by executing three iterations of PSI-BLAST against NCBI's non-redundant (NR) database. The PSSM values were normalized further using numeric value nine [175], which we call as PSSM normalizing factor.



（a） SL477 （b） MxD444

Fig 4. Density distribution curves of monograms and bigrams for (a) SL477 and (b) MxD444 dataset. The *x*–axis and *y*–axis show the monograms/bigrams in logarithmic scale and density index of the distribution, respectively. For each figure, the dotted （*red*） and solid （*blue*） vertical lines correspond to median values of the distribution for monograms （MG） and bigrams （BG）, respectively.

The literature suggests that the conserved evolutionary information given by PSSM can be transformed from primary structure (amino acid sequence) level to three dimensional structure level by computing monograms and bigrams from PSSM values [176]. The monogram-bigram probabilities characterize the subsequence of a protein sequence that can be conserved within a fold in terms of transition probabilities from one amino acid to another [177]. Thus, the monogram-bigram features are useful in identifying the evolutionary folded (ordered) or, unfolded (disordered) region of proteins, which motivated us to utilize them as features in disorder prediction. We computed monogram feature matrix (1 × 20) and bigram feature matrix (20 × 20) for each sequence from its PSSM. Monogram feature matrix consists of one monogram value (**MG** in **Table 1**) for each type of amino acid and bigram feature matrix consists of one bigram value (**BG** in **Table 1**) for each pair of 20 possible amino acids, respectively. Further, our analysis based on multiple datasets collected from PDB and DisProt shows that both monograms and bigrams follow a normal density distribution in their logarithmic space with approximately consistent median value equals to 6.0 within any dataset (**Fig. 4**). Therefore, we used $e^{6.0}$ to normalize these values and reduce the noise.

2.3.3.3 *Structural Information*

We employed sequence based predicted secondary structure (**SS** in **Table 1**) probabilities for helix, sheet and coil residues [175], predicted solvent accessibility (**ASA** in **Table 1**) [178] and predicted backbone dihedral torsion angles, phi and psi (ϕ and ψ in **Table 1**) fluctuations [179] as features. We included these six features since disordered residues can be characterized by the lack of stable secondary structure [35, 52] and also the unstructured regions are found to have large solvent accessible area [15].

Note that, we included the fundamental features to characterize disorder in proteins in our feature set which are well studied in the literature. Further, we enhanced the feature set by including new features, like MGs and BGs. Finally, we included the information of neighboring residues within the features of each residue by using a sliding window, keeping the target residue at the center of the window. The motivation was to incorporate the native interactions and contacts of neighboring residues which are found to play essential roles in determining protein structures and protein folding dynamics [180]. We determined the 10-fold cross-validation performance of DisPredict for 13 different window sizes (1, 3, 5, …, 23, 25) to find the optimal window size 21. Thus, there were 1176 (since, *window size* × *total feature count* = (21 × 56) = 1176) features used for each residue. The features were finally scaled within the range [-1, +1] before using.

2.3.4 Performance Evaluation Criteria

The performance of DisPredict [10] is evaluated using the criteria followed in the past Critical Assessment of protein Structure Prediction (CASP) competitions [166-168]. The measures and procedures used in CASP experiments are comprehensive. The predictions are done in two levels:

- Binary value, defining whether a residue is disorder or not (“$+1$” for disorder and “-1” for order) and

- Real value, quantifying the probability of a residue being disorder ("≥ 0.5" for disorder and "< 0.5" for order).

Binary Prediction Evaluation: In binary (two-class) prediction of disorder, TP (True Positive) = number of correctly predicted disordered residues, TN (True Negative) = number of correctly predicted ordered residues, FP (False Positive) = number of incorrectly predicted disordered residues and FN (False Negative) = number of incorrectly predicted ordered residues. To determine the total number of correct prediction (both ordered and disordered), $N_{correct} = TP + TN$ is calculated. Sensitivity ($SENS$) and specificity ($SPEC$) are two complementary statistical measures identifying the proportionate values of correct prediction of disordered (positive class) and ordered (negative class) residues, respectively.

$$SENS = \frac{TP}{TP + FN} = \frac{TP}{N_d} \tag{1}$$

$$SPEC = \frac{TN}{TN + FP} = \frac{TN}{N_o} \tag{2}$$

Here, N_d and N_o are the total number of disordered and ordered residues, respectively. As increment of one of these measures (SENS and SPEC) usually leads towards the decrement of another measure, neither of these two measures is a suitable indicator of performance for an imbalanced dataset. On the contrary, the balanced accuracy (ACC), weighted score (S_w) and Mathews correlation coefficient (MCC) are the measures that take all four components of prediction quality (TP, TN, FP and FN) into account and thus can be regarded as more important indicators.

$$ACC = \frac{1}{2}(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}) \tag{3}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4}$$

$$S_w = \frac{(w_d \times TP - w_o \times FP) + (w_o \times TN - w_d \times FN)}{(w_d \times N_d) + (w_o \times N_o)} \tag{5}$$

Here, $w_d = \frac{N_o}{N_o + N_d} =$ percentage of ordered residues, is the weight for N_d, and $w_o = \frac{N_o}{N_o + N_d} =$ percentage of disordered residues, is the weight for N_o [167]. The S_w measure includes weight to address the imbalance in the ratio of ordered and disordered residues and rewards correct disorder classification over correct classification of ordered residues, which is later found to have a linear relationship with ACC ($S_w = 2 \times Acc - 1$) [181]. Since both of these measures (ACC and S_w) have been used in CASP assessment, we have also included both of them in our paper instead of just one. MCC score, another measure that accounts for all four parameters of the prediction quality, is the most reasonable and consistent measure for disorder prediction assessment because of not being favorable to over prediction of any class (order/disorder). MCC and S_w scores vary from -1 to 1, where -1 and 1 represent perfect misclassification and classification, respectively with a random classification scoring by 0. More recently, precision ($PPV = \frac{TP}{TP+FP}$) has been appeared as a good measure for binary disorder prediction as it is totally insensitive to the prediction of the dominant class (*i.e.*, here the order state), is therefore computed to evaluate DisPredict. As the prediction becomes better, the values of these metrics also get higher.

We calculated Mean Absolute Error (MAE) $= \frac{\sum_{i=1}^{n}|c_d^a(i)-c_d^p(i)|}{n}$ to quantify the error of disorder prediction in content level. Here, n is the total number of protein chains, and c_d^a and c_d^p are the actual and predicted disorder content (fraction of disordered residues) for the i^{th} protein chain, respectively. The lower value of MAE corresponds to better prediction.

Evaluation of Predicted Probability: The SVM model of DisPredict generates a predicted probability value for each residue, which signifies the disorder confidence of that residue. This probability value is then used for binary classification and annotation by applying a threshold value 0.5. If the probability is greater than or equal to 0.5, the predicted class is considered ‘disorder’ and if the probability is less than 0.5, the predicted class is considered ‘order’. Assessment of the predicted probability by a DisPredict is performed by receiver operating characteristic (ROC) curve, which depicts the correlation between the true positive rate (TPR or, SENS) and false positive rate ($FPR = 1 - SPEC$) for a probability threshold. The area under the ROC curve (AUC) quantifies the predictive quality of a classifier, where the AUC value equal to 1 indicates a perfect prediction and 0.5 corresponds to a random prediction. Moreover, 95% confidence interval (CI) for the AUC score is evaluated using DeLong's [182] variance estimated by bootstrapping. The evaluation of AUC and CI are performed using the statistical R package with the pROC library [183].

2.4 DisPredict (version 1.0)

In this section, we describe the design and development of DisPredict (version 1.0).

2.4.1 SVM Design and Parameterization

DisPredict is a two-layer disorder predictor that integrates *optimization-layer* and *classification-layer*. The classification-layer is developed using a single support vector machine (SVM), namely LIBSVM [184]. Due to the working principle of SVM of simultaneously minimizing the empirical classification error (training error) and generalized error (test error) by maximizing the geometric margin of the separating hyperplane, it can be regarded as an effective technique in hard classification problems especially in bioinformatics and computational biology area. We used Gaussian or, radial basis function (RBF) kernel for the SVM to extend its capability to handle non-linearly separable classes. RBF transforms the input feature space into infinite dimension space (*i.e.* Hilbert space), which results in a linear separating affine-plane or a hyperplane.

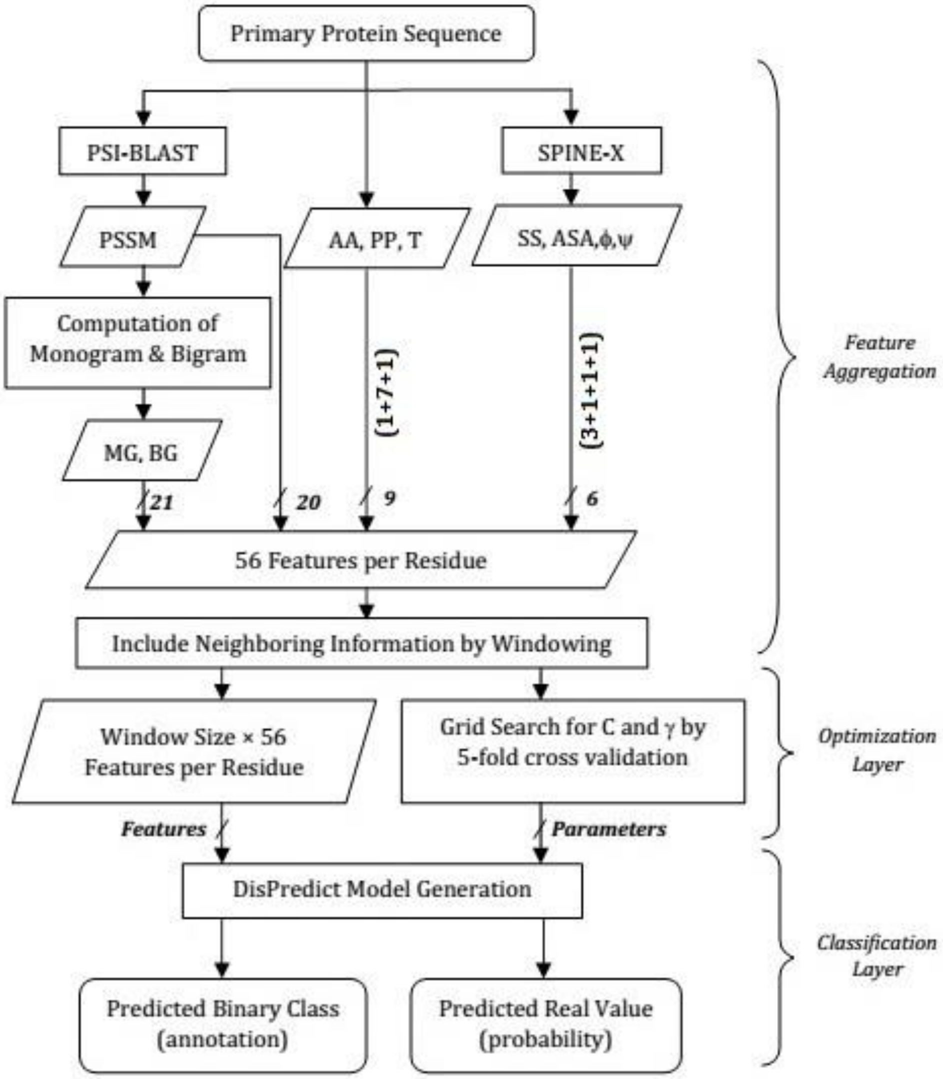

Primary Protein Sequence
PSI-BLAST
SPINE-X
PSSM
AA, PP, T
SS, ASA,ϕ,ψ
Computation of Monogram & Bigram
MG, BG
(1+7+1)
(3+1+1+1)
21
20
9
6
56 Features per Residue
Include Neighboring Information by Windowing
Window Size × 56 Features per Residue
Grid Search for C and γ by 5-fold cross validation
Features
Parameters
DisPredict Model Generation
Predicted Binary Class (annotation)
Predicted Real Value (probability)
Feature Aggregation
Optimization Layer
Classification Layer

Fig 5. DisPredict Framework: feature aggregation, optimization-layer and classification-layer. In the feature aggregation step, features are shown in their abbreviated form according to **Table 1** and the arrows are labeled by the number of features involved. The classification-layer receives final feature set from the feature aggregation step and optimal parameters from the optimization-layer. Then, it generates the predictor model and outputs both binary annotation and real-valued class probabilities.

On the other hand, in the optimization-layer of DisPredict, we selected two parameters, C and γ, where C is the cost of misclassification and γ is the parameter of fitting best mode of RBF. The optimal values for the parameters C and γ are determined by grid search using 5-fold cross-validation. However, in our case the grid search turned out to be computationally very intensive. Thus, we used 5% of the training dataset to determine the optimal parameters instead. The output of DisPredict, *i.e.*, the disordered or ordered residue probabilities, is optimized by a round of 5-fold cross validation. Using the threshold value 0.5, the probabilities are converted into binary decision variables, where probability ranges $0.5 \leq range_d \leq 1.0$ is considered as disordered and $0.0 \leq range_o < 0.5$ is considered as ordered. **Fig. 5** shows the detail paradigm of DisPredict.

2.4.2 Implementation and Availability

We implemented the DisPredict tool in C. The software is developed and tested on Linux platform. It is dependent on two external packages, namely PSI-BLAST[8] and NR database[9], which are publicly available. The software is available online[10] with a user manual.

2.5 Evaluation of DisPredict1.0

Here, we evaluate the performance of DisPredict through cross-validation and testing with independent datasets. We further compare the of performance DisPredict with two other existing predictors, and analyze its effectiveness through case-studies.

2.5.1 10-Fold Cross Validation

2.5.1.1 *Default parameters for SVM*

We evaluated the 10-fold cross-validation performance of DisPredict separately on SL477 and MxD444 dataset. Regarding the optimum selection of the window size, we ran cross-validation for 13 different windows, shown in **Table 2**, for both of the SL477 and MxD444 dataset with default parameters for SVM. The best result for window size 25 was found with ACC, MCC and AUC values equal to 0.82, 0.65 and 0.91, respectively for SL477 dataset, whereas for MxD444 dataset the values are 0.77, 0.48 and 0.85,

[8] PSI-BLAST link: ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/

[9] NR database link: ftp://ftp.ncbi.nlm.nih.gov/blast/db/

[10] DisPredict link: https://github.com/tamjidul/DisPredict_v1.0

respectively. The gradual increase in performance becomes a plateau as window goes higher above size 23 (**Fig 6**).

Table 2. 10-fold Cross Validation Performance of DisPredict (Default Parameter).

W_{size}	TP	TN	FP	FN	$N_{correct}$ (total)[1]	SENS	SPEC	ACC	S_w	PPV	MCC	AUC [95%CI][2]
SL477 Dataset												
1	4440	5804	1469	1260	10244 (12973)	0.779	0.798	0.788	0.577	0.751	0.574	0.869 [0.862 , 0.876]
3	4467	5954	1319	1233	10421 (12973)	0.784	0.819	0.801	0.602	0.772	0.601	0.884 [0.877 , 0.890]
5	4457	6020	1253	1243	10477 (12973)	0.782	0.828	0.805	0.609	0.781	0.609	0.889 [0.882 , 0.896]
7	4441	6076	1197	1259	10517 (12973)	0.779	0.835	0.807	0.614	0.787	0.615	0.893 [0.886 , 0.899]
9	4457	6086	1187	1243	10543 (12973)	0.782	0.837	0.809	0.618	0.789	0.619	0.895 [0.888 , 0.902]
11	4483	6113	1160	1217	10596 (12973)	0.786	0.841	0.813	0.627	0.794	0.628	0.898 [0.891 , 0.905]
13	4502	6114	1159	1198	10616 (12973)	0.790	0.841	0.815	0.63	0.795	0.631	0.899 [0.892 , 0.905]
15	4513	6150	1123	1187	10663 (12973)	0.792	0.845	0.819	0.637	0.801	0.638	0.902 [0.896 , 0.909]
17	4540	6133	1140	1160	10673 (12973)	0.796	0.843	0.82	0.64	0.799	0.64	0.902 [0.895 , 0.902]
19	4545	6148	1125	1155	10693 (12973)	0.797	0.845	0.821	0.643	0.802	0.643	0.903 [0.896 , 0.910]
21	4548	6148	1125	1152	10696 (12973)	0.798	0.845	0.822	0.643	0.802	0.643	0.903 [0.896 , 0.910]
23	4555	6167	1106	1145	10722 (12973)	0.800	0.847	0.823	0.647	0.804	0.647	0.904 [0.898 , 0.911]
25	4564	6164	1109	1136	**10728** (12973)	**0.801**	**0.847**	**0.824**	**0.648**	**0.804**	**0.648**	**0.905 [0.898 , 0.911]**
MxD444 Dataset												
1	3284	13093	3397	1632	16377 (21406)	0.668	0.793	0.731	0.462	0.491	0.419	0.817 [0.810 , 0.825]
3	3369	13241	3249	1547	16610 (21406)	0.685	0.803	0.744	0.488	0.509	0.444	0.832 [0.826 , 0.840]
5	3410	13302	3188	1506	16712 (21406)	0.694	0.807	0.75	0.5	0.516	0.456	0.839 [0.833 , 0.847]
7	3419	13275	3215	1497	16694 (21406)	0.695	0.804	0.75	0.501	0.515	0.455	0.840 [0.833 , 0.847]
9	3446	13253	3237	1470	16699 (21406)	0.7	0.805	0.752	0.505	0.516	0.458	0.842 [0.834 , 0.849]
11	3503	13232	3258	1413	16735 (21406)	0.712	0.802	0.757	0.515	0.517	0.466	0.846 [0.839 , 0.853]
13	3523	13188	3302	1393	16711 (21406)	0.717	0.8	0.758	0.516	0.516	0.466	0.847 [0.839 , 0.853]
15	3564	13145	3345	1352	16709 (21406)	0.725	0.797	0.761	0.522	0.515	0.469	0.848 [0.842 , 0.855]
17	3578	13097	3393	1338	16675 (21406)	0.728	0.794	0.761	0.522	0.513	0.469	0.848 [0.841 , 0.855]
19	3607	13068	3422	1309	16675 (21406)	0.734	0.792	0.763	0.526	0.513	0.471	0.849 [0.842 , 0.856]
21	3613	13078	3412	1303	16691 (21406)	0.735	0.793	0.764	0.528	0.514	0.473	0.850 [0.843 , 0.857]
23	3640	13059	3431	1276	16699 (21406)	0.74	0.792	0.766	0.532	0.515	0.476	0.851 [0.845 , 0.859]
25	3658	13064	3426	1258	**16722** (21406)	**0.744**	**0.792**	**0.768**	**0.536**	**0.517**	**0.479**	**0.852 [0.847 , 0.861]**

W_{size} indicates the size of sliding window.

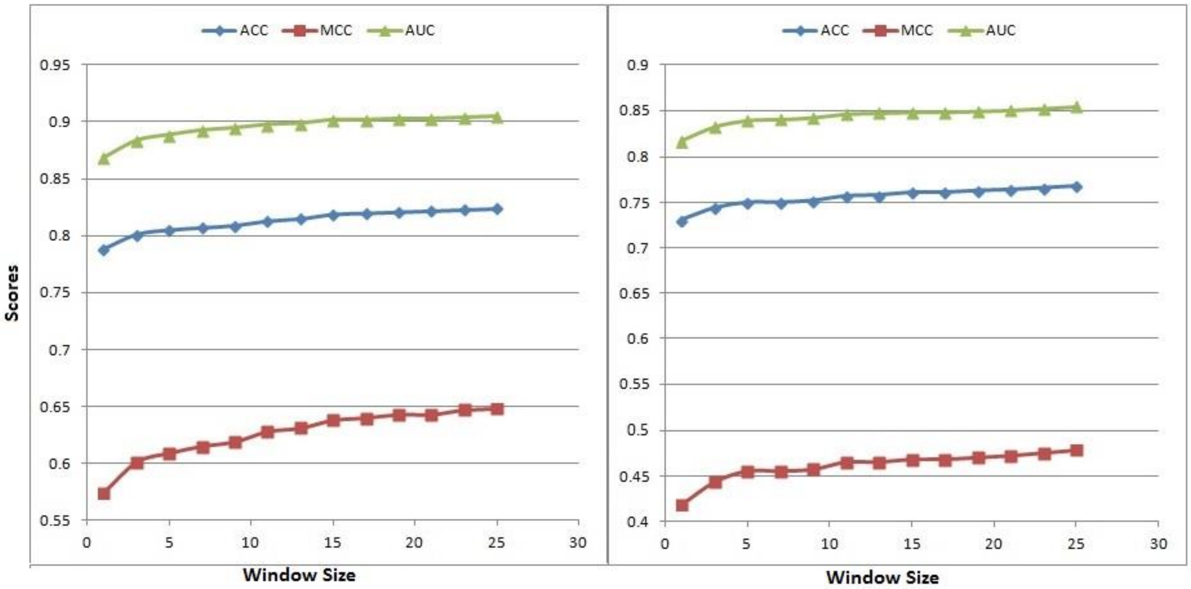
Best values of each metric are marked in bold for each dataset separately.

[1] $N_{correct}$ is reported with total number of residues (Residue$_{total}$) to be predicted in parentheses. Both the counts correspond to one subset (fold) of the full dataset, which is generated for performing cross validation.

[2] For AUC, the values within bracket indicate 95% confidence interval with 2000 stratified bootstrap replicas.

As the window size continues to increase, the rate of increase in scores becomes slow. Increase of scores is ≤0.001, as the windows size grows from 23 to 25 for SL477 dataset and ≤0.004 for MxD444 dataset, respectively.

Table 2 also depicts the inverse relationship between SENS and SPEC scores with increasing window size for MxD444 dataset. The best SENS (0.74) is achieved by window size 25 while the best SPEC (0.81) is achieved at window size 5. Overall, the consistent increment in balanced accuracy (ACC) and PPV prove our methodology to be well balanced.



(a) SL477 (b) MxD444

Fig 6. 10-fold cross-validation (default parameter) performance of DisPredict with different window sizes. Results are shown in terms of ACC, MCC and AUC scores on (a) SL477 and (b) MxD444 dataset. The x-axis and y-axis represent the window sizes and scores, respectively.

2.5.1.2 *Optimized Parameters for SVM*

The preliminary extensive analysis of performance with multiple window sizes is done without selection of optimal parameters for SVM. For a specific window size (W_{size}) and total number of residues ($Residue_{total}$) in a dataset, we have a feature matrix of dimension, $Residue_{total}\times(W_{size}\times56)$. Therefore, the increase in window size leads towards the increase in the dimensions of the feature space, which in turn makes the time expensive grid search for parameters slower. To tradeoff between performance with optimization and time complexity of parameter selection along with model generation, we determined the optimal values of parameters with a 5% randomly selected subset of residues from training dataset for 3 window sizes (15, 21 and 25). The optimal parameters (C and γ) found from grid search are reported in **Table 3**.

Furthermore, we inserted repeated disordered residue information only in case of training to balance the dataset as the support vector points for the less dominant class may not be sufficient to determine the optimal SVM margin. Specifically, duplicates (2 times for SL477 dataset and 3 times for MxD444 dataset) of disorder samples were provided during generation of predictor model. However, in case of testing, no repeated information was inserted. **Table 4** illustrates the detail of the cross-validation results with optimized parameters for 3 different window sizes.

Table 3. Optimized Parameters used to build final DisPredict Model.

	SL477 dataset		MxD444 Dataset	
W_{size}	C	γ	C	γ
15	8.0	0.0019531	8.0	0.0312500
21	2.0	0.0078125	2.0	0.0078125
25	0.5	0.0078125	0.5	0.0078125

C is the soft penalty parameter to handle overlapped class.
γ is the parameter for radial basis kernel for SVM.

The improvement of performance with optimized parameters over non-optimized parameters one was significant. To compare, for SL477 dataset (window size 21), FP and FN values are reduced to 1,002 and 1,083 from 1,125 and 1,152 due to optimization. In case of MxD dataset (window size 21), the FN value is increased by 133 residues. However, the FP value is also decreased by 1,812 residues which maintains the overall increase in the total number of correctly predicted residues from 16,691 to 18,370. The improvement of prediction, both in terms of increased correct classification and decreased misclassification, is also visible from both the sensitivity and specificity scores. For window size 21, the values of S_w, precision and MCC are improved by 4.5%, 2.5% and 4.5% respectively due to optimized training on SL477 dataset. At the same time, for MxD444 dataset, these progresses are 15.7%, 33.3% and 26.8% respectively. Note that, this significant improvement in MCC strongly supports our method's capability in handling the imbalance ratio of ordered and disordered residues. Further, the AUC score is also increased by 4.4% and 0.4% as the result of optimization for SL477 and MxD444 dataset, respectively.

A comparative analysis of **Table 2** and **Table 4** also shows that optimized DisPredict model with window size 21 outperforms all the other models of its own kind. Thus, we select 21 as the optimal window size for our proposed DisPredict. Furthermore, to understand the relevance of the new features (MGs and BGs) with protein disorder, we separately evaluated optimized DisPredict's performance without monograms and bigrams. We performed 10-fold cross validation on SL477 dataset with the optimal window size 21 and optimal parameters of SVM as reported in **Table 3** for SL477 dataset with window size 21. The result of this experiment in terms of ACC, MCC and S_w score are 0.810, 0.651 and 0.621, respectively. The comparison of these scores excluding MGs and BGs with those of including MGs and BGs (reported in

Table 4 for SL477 dataset) shows that involvement of MGs and BGs along with PSSM leads to a further increase in binary prediction accuracy in terms of 3.2% improved ACC (0.810 to 0.836), 3.8% improved MCC (0.651 to 0.673) and 8.2% improved S_w score (0.621 to 0.672).

Table 4. 10-fold Cross Validation Performance of DisPredict (Optimized Parameter).

W_{size}	TP	TN	FP	FN	$N_{correct}$ ($Residue_{total}$)[1]	SENS	SPEC	ACC	S_w	PPV	MCC	AUC [95%CI][2]
SL477 Dataset												
15	4655	6056	1217	1045	10711 (12973)	0.817	0.833	0.825	0.649	0.793	0.647	0.898 [0.890 , 0.904]
21*	4617	6271	1002	1083	10888 (12973)	0.810	0.862	0.836	0.672	0.822	0.673	0.956 [0.950 , 0.963]
25	4624	6234	1039	1076	10858 (12973)	0.810	0.857	0.834	0.668	0.816	0.669	0.911 [0.904 , 0.917]
MxD444 Dataset												
15	2590	15590	900	2326	18180 (21406)	0.527	0.945	0.736	0.472	0.742	0.538	0.838 [0.831 , 0.845]
21	3480	14890	1600	1436	18370 (21406)	0.708	0.903	0.805	0.611	0.685	0.600	0.853 [0.847 , 0.859]
25	3367	3367	1635	1549	18222 (21406)	0.685	0.901	0.793	0.586	0.673	0.582	0.850 [0.843 , 0.858]

W_{size} indicates the size of sliding window.
Best values of each metric are marked in bold for each dataset separately.
[1] $N_{correct}$ is reported with total number of residues ($Residue_{total}$) to be predicted in parentheses. Both of the counts correspond to one subset (fold) of the full dataset which is generated for performing cross validation.
[2] For AUC, the values within bracket indicate 95% confidence interval with 2000 stratified bootstrap replicas.

2.5.1.3 *Probability and Performance Analysis of Residual Overlap for Residue and Chain Level Splitting of Dataset*

To uniformly distribute the residues into ten subsets for cross validation, we applied modular arithmetic operation to split the dataset in residue level. As the residues are already included within the neighboring information based on the window, they are detachable from their original sequence. However, this inclusion of residue information within window may yield overlap of information between training and test sets in case of residue level splitting of dataset for cross validation. We analyzed the probability of this residual overlap between training and test sets.

Let, there are N sequences in the dataset and the expected length of the sequence is $\mathcal{L}$. Then, the possibility of picking two residues for training and test subsets of 10-fold cross-validation which belongs to same sequence is $\left(\frac{1}{\frac{9N}{10}} \times \frac{1}{\frac{N}{10}}\right) = \frac{100}{9N^2}$. Since the expected length of a sequence is $\mathcal{L}$, the chance of training and test overlap for a specific window size (W_{size}) is $\frac{W_{size}-1}{\mathcal{L}}$. Altogether, the probability of a train and test residue overlap from the same sequence is $\left(\frac{100}{9N^2} \times \frac{W_{size}-1}{\mathcal{L}}\right) = \left(\frac{100}{9}\right)\frac{W_{size}-1}{N^2\mathcal{L}}$. For SL477 dataset with $N =$

477, approximate $\mathcal{L} = 400$ and $W_{size} = 21$, the probability of the overlap is 2.44 × 10^{-06}, which is significantly low and thus can be safely ignored. Further, we reevaluated DisPredict's 10-fold cross validation performance with sequence level sampling by modular operation for SL477 dataset to generate training and test subsets. **Table 5** quantifies the difference in performance between residue level and state-of-the-art practice of sequence level splitting of dataset for cross validation with window size 21 and default parameters for SVM. It showed that DisPredict's performance remains consistent without any significant over-prediction in terms of all the metrics.

Table 5. Cross-validation performance of DisPredict with residue level and sequence level splitting of SL477 dataset.

Splitting Method	SENS	SPEC	ACC	S_w	PPV	MCC	AUC [95% CI]
Residue Level	0.798	0.845	0.822	0.643	0.802	0.643	0.903 [0.896 , 0.910]
Sequence Level	0.784	0.844	0.814	0.628	0.793	0.627	0.892 [0.886 , 0.898]

Default values of C and γ are applied for SVM.
Window size 21 is used.

2.5.2 Independent Training and Testing

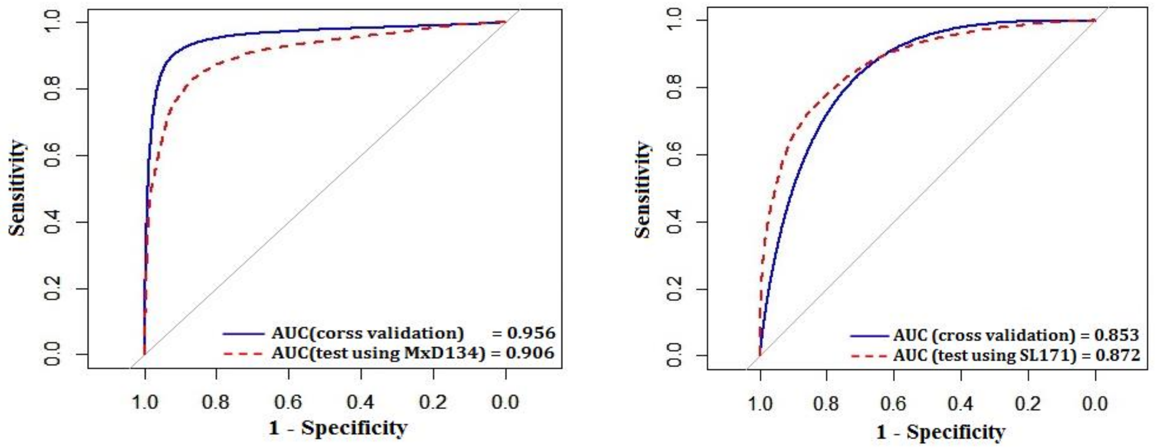
With optimized parameters and balanced dataset, we carried out independent training on SL477 and MxD444 datasets followed by testing the resulting predictor model with MxD134 and SL171 dataset, respectively. Note that, these independent test datasets (MxD134 and SL171) were generated at low sequence identity (10%) with the corresponding training datasets (SL477 and MxD444). The consistent results of these two tests done through cross validation and independent test confirm the usage of robust technique and effective feature set in DisPredict as well as training efficacy avoiding possible over-fittings.

Table 6 further illustrates the results of these tests, where we reported the average of the scores computed for equally divided 10 subsets of the full dataset along with the corresponding standard deviation (STDEV). **Table 6** reveals that training by SL477 dataset gives consistent performance regardless of test datasets and test procedures (cross validation or independent test) in terms of ACC: 0.836, 0.833 and S_w: 0.672, 0.667. These consistencies are also evident in case of training with MxD444 dataset while tested by different datasets and the evaluations are, ACC: 0.805, 0.789 and S_w: 0.611, 0.577. We calculated the Mean Absolute Error (MAE) which is also reported along with its corresponding STDEV from mean. The score indicates that the error does not increase from cross validation to independent test as the test-results were robust.

Table 6. Performance Comparison of Cross Validation and Independents Tests.

Model Evaluation	SENS	SPEC	ACC	S_w	PPV	MCC	AUC	MAE
Procedure[1]	(STDEV)	(STDEV)	(STDEV)	(STDEV)	(STDEV)	(STDEV)	(STDEV)	(STDEV)
10-fold cross validation on SL477	0.810	0.862	0.836	0.672	0.822	0.673	0.956	0.032
	(0.004)	(0.001)	(0.002)	(0.005)	(0.002)	(0.004)	(0.007)	(0.002)
Train by SL477, Test on MxD134	0.744	0.923	0.833	0.667	0.574	0.598	0.906	0.023
	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.004)	(0.001)	(0.001)
10-fold cross validation on MxD444	0.708	0.903	0.805	0.611	0.685	0.600	0.853	0.208
	(0.006)	(0.001)	(0.003)	(0.006)	(0.002)	(0.004)	(0.007)	(0.001)
Train by MxD444, Test on SL171	0.718	0.860	0.789	0.577	0.748	0.583	0.872	0.151
	(0.003)	(0.001)	(0.001)	(0.003)	(0.001)	(0.002)	(0.007)	(0.001)

[1] All the evaluations are carried out using a sliding window of length 21 and optimized parameters for SVM.



（a） SL477　　　（b） MxD444

Fig 7. ROC curves given by DisPredict on the training dataset (a) SL477 and (b) MxD444 dataset. In each figure, the solid （*blue*） curve corresponds to the cross-validation test on the same dataset and the dotted （*red*） curve corresponds to the independent test. The AUC values given in each figure correspond to the values in **Table 6**. The x-axis and y-axis show the Specificity and Sensitivity, respectively.

To analyze the quality of the predicted probability, the ROC curves given by DisPredict are plotted in **Fig 7** in continuous scale between 0.0 and 1.0. In each figure, two ROCs are plotted keeping the training dataset same with varying test datasets and evaluation procedure. Finally, we reported the AUC values which are found consistent for cross validation and independent test indicating our predictor's capability to avoid over-fitting.

2.5.3 Comparison with Other Predictors

The performance of DisPredict1.0 (or DisPredict [10]) is compared with the state-of-the-art disorder predictors, MFDp [164] and SPINE-D [123]. To remain fair while comparing DisPredict with each of the above two predictors, we train DisPredict separately with respective datasets and compare with each of them separately. Thus, DisPredict is compared with MFDp based on dataset MxD444, while dataset SL477 is used to compare DisPredict with SPINE-D (**Table 7**).

Table 7. Comparison of DisPredict with MFDp and SPINE-D respectively on MxD444 and SL477 dataset.

Method	SENS	SPEC	ACC	S_w	MCC	AUC
DisPredict[1]	0.71	**0.90**	**0.80**	**0.61**	**0.60**	**0.85**
MFDp[2]	**0.76**	0.75	0.75	0.51	0.44	0.81
DisPredict[3]	**0.81**	**0.86**	**0.84**	**0.67**	**0.67**	**0.96**
SPINE-D[4]	0.77	0.85	0.81	0.62	0.63	0.87

[1] 10-fold cross validation performance of DisPredict on MxD444 which is a subset of 444 chains out of 514 chains with no X-tag.
[2] 5-fold cross validation performance of MFDp on MxD dataset of 514 protein chains [164].
[3] 10-fold cross validation performance of DisPredict on SL477.
[4] 10-fold cross validation performance of SPINE-D [123] on SL477.

In particular, MFDp [164] is a meta predictor that combines the predictions from three disorder predictors (DISOPRED2 [126], DISOclust [185] and IUPred [186]). Further, MFDp combines the outputs from three SVMs with linear kernel using a threshold of 0.37, used to output binary prediction. In contrast, we utilized single SVM with RBF kernel and optimized parameters combined with a comprehensive set of features to develop the standalone predictor. However, the performance of MFDp in **Table 7** is of 5-fold cross validation whereas DisPredict is evaluated by 10-fold cross validation and hence to be considered reliable rather than over-fitted by chance. In terms of MCC, DisPredict improved significantly, which is 36.36% better than MFDp. The improvement in S_w score is also 19.6%. DisPredict showed lower sensitivity (7%) than MFDp while at the same time improved specificity by 20%, which in turn improved the balanced accuracy by 6.67%. Moreover, DisPredict outperformed MFDp in AUC score by 1.29% which is used to assess the probability based prediction.

The other state-of-the-art predictor, SPINE-D [123] utilizes ANN technique which was first developed to output three state prediction and later reduced into two state predictor of ordered and disordered residues.

SPINE-D employs a disorder probability threshold of 0.06 that was optimized to achieve maximum S_w score. On the contrary, DisPredict is a SVM-based two-state disorder predictor using a more meaningful threshold for two-class classification of value 0.5. DisPredict outperformed SPINE-D in terms of sensitivity as well as specificity by 5.19% and 1.18% respectively which leads to 3.7% improvement in overall accuracy. DisPredict also outperformed SPINE-D in terms of S_w, MCC and AUC by 8.06%, 6.34% and 10.34% respectively.

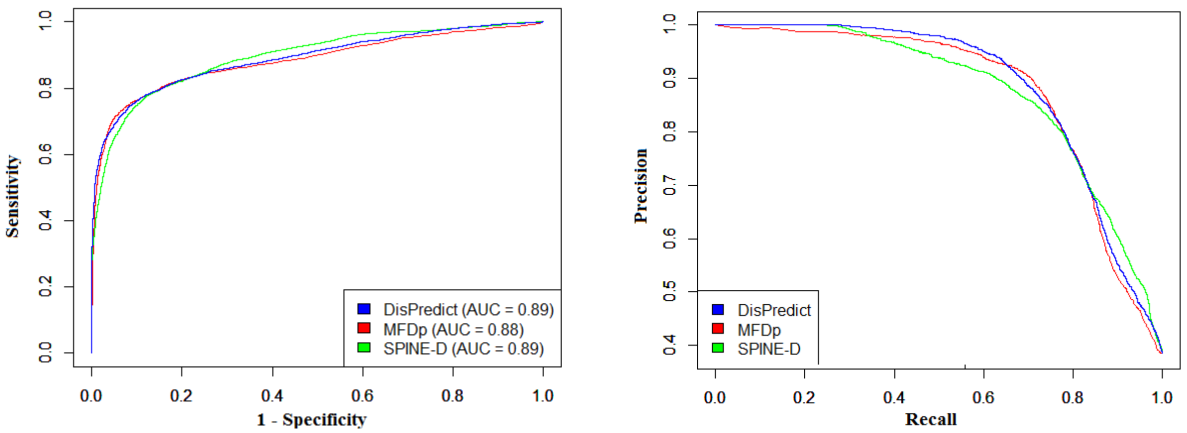
Table 8. Performance comparison among DisPredict, SPINE-D and MFDp on independent DD73 dataset.

Method	**SENS**	**SPEC**	**ACC**	**S_w**	**PPV**	**MCC**	**AUC [95% CI]**
DisPredict*	0.775	**0.883**	**0.829**	**0.658**	**0.806**	**0.663**	**0.89 [0.88, 0.90]**
SPINE-D	**0.769**	0.847	0.822	0.644	0.765	0.639	**0.89 [0.88, 0.90]**
MFDp	0.780	0.875	0.828	0.656	0.796	0.658	0.88 [0.87, 0.89]

* Window size = 21, C = 2.0 and γ = 0.0078125.
Best results are marked by bold.

In addition to the comparison on cross validation test, we evaluated DisPredict, SPINE-D [123] and MFDp [164] on independent DD73 dataset. The comparison among these three methods is illustrated in **Table 8**. It shows that DisPredict gives better performance among three predictors except in case of sensitivity. DisPredict yielded 2.63% lower sensitivity than that of SPINE-D[123], whereas DisPredict gave 4.25% higher specificity than that of SPINE-D [123]. **Table 8** also shows that DisPredict outperformed SPINE-D [123] and MFDp [164] in terms of MCC by 3.76% and 0.76%, respectively. At the same time, DisPredict gave 1.26% and 5.36% improved precision (PPV) than MFDp [164] and SPINE-D [123], respectively. However, DisPredict resulted slightly lower sensitivity than those of SPINE-D [123] and MFDp [164]. At the same time, both SPINE-D [123] and MFDp [164] gave lower specificity than that of DisPredict.

Fig 8 compares the ROC curves and precision-recall curves, respectively, given by DisPredict, SPINE-D [123] and MFDp [164]. **Fig 8(a)** shows that the ROC curves given by the three predictors are comparative. At the same time, the precision-recall curves (**Fig 8(b)**) depicts that DisPredict achieves consistently higher precision upto less than 65% sensitivity (recall).



（a） ROC （DD73）　　　（b） Precision-recall （DD73）

Fig 8. (a) ROC and (b) precision-recall curves given by DisPredict (*blue*), SPINE-D (*green*) and MFDp (*red*) while predicting disorder on DD73 dataset. The AUC values shown in the figure correspond to the values in Table 8. For （a）， the x-axis and y-axis show the Specificity and Sensitivity， respectively， and for （b） the x-axis and y-axis show the Recall （Sensitivity） and Precision （PPV）， respectively.

MFDp and SPINE-D have been established as the best disorder predictor among 8 and 11 existing disorder predictors [123, 164], respectively, covering different approaches in their relevant publication. In this article, our predictor is shown to be comparable with both of these methods. Therefore, DisPredict can be considered to be one of the finest disorder predictor and can be utilized to produce more reliable annotation of disorder versus order residues.

2.5.4 Case Studies: Characteristic Region and Protein Function

Proteins with disordered regions are found to contain several regions of interest, such as self-stabilizing folded regions, DNA or, nucleotide binding regions, short (up to 20 amino acids) conserved regions of biological significance (known as motif), mediating regions for protein interaction with different partners etc. These characteristic regions undergo various conformational changes, gain structure and affect many important biological functions. We selected three proteins as cases (UniProt IDs: P41212, P01116 and P04637) with experimentally verified regions of interest to analyze per residue disorder confidence score assigned by DisPredict, SPINE-D and MFDp. **Fig 9** illustrates the disorder probability of each residue with respect to residue index.

2.5.4.1 *UniProtKB – P41212 (ETV6_HUMAN)*

P41212 is a human ETV6 protein[11] for transcriptional repressor function, which is also involved in several kinds of leukemia and syndrome. **Fig 9(a)** indicates that for this protein, DisPredict and SPINE-D showed comparable performance in detecting the highly conserved region of PNT (pointed) domain (residues 40 – 124) [187] and ETS (E26 transformation-specific) DNA binding region (residues 339 - 420) [188] respectively, while MFDp outperformed both of them with relatively less noise.

2.5.4.2 *UniProtKB – P01116 (RASK_HUMAN)*

P01116 is a human KRAS protein[12] with intrinsic GTPase activity (binds GDP/GTP) [189] and related to several diseases, such as gastric cancer (GASC), acute myelogenous leukemia (AML), cardiofaciocutaneous syndrome 2 (CFC2) etc. **Fig 9(b)** shows that DisPredict could identify its GTP (guanosine triphosphate) binding region (residues 10 – 17) and effector region (residues 32 – 40) respectively, with close to cut-off (0.5) probabilities. Note that, these two regions are experimentally verified unstructured regions, which are strongly suggested as structured by both SPINE-D and MFDp. However, the C-terminal hypervariable region (residues 166 – 185) is consistently detected by all three of these predictors.

2.5.4.3 *UniProtKB – P04637 (P53_HUMAN)*

P04637 corresponds to human p53 protein[13] which acts as a tumor suppressor. **Fig 9(c)** illustrates that DisPredict and MFDp outperformed SPINE-D with relatively sharp detection of N-terminal TADI (transcriptional repression domain-I) motif (residues 17 – 25) [190]. On the other hand, DisPredict and SPINE-D outperformed MFDp in determining oligomerization domain [191] of residues 325 - 356. **Fig 9(c)** also shows that both SPINE-D and MFDp missed the very short, 3-residues (370 - 372) long [KR]-[STA]-K binding motif at C-terminal, while DisPredict detected it correctly. The overall comparison depicts that DisPredict's performance is more biologically relevant with correct identification of these short regions. Therefore, it would be interesting to utilize DisPredict in a broader scope in near future.

[11] UniProtKB – P41212 link: http://www.uniprot.org/uniprot/P41212
[12] UniProtKB – P01116 link: http://www.uniprot.org/uniprot/P01116
[13] UniProtKB – P04637 link: http://www.uniprot.org/uniprot/P04637

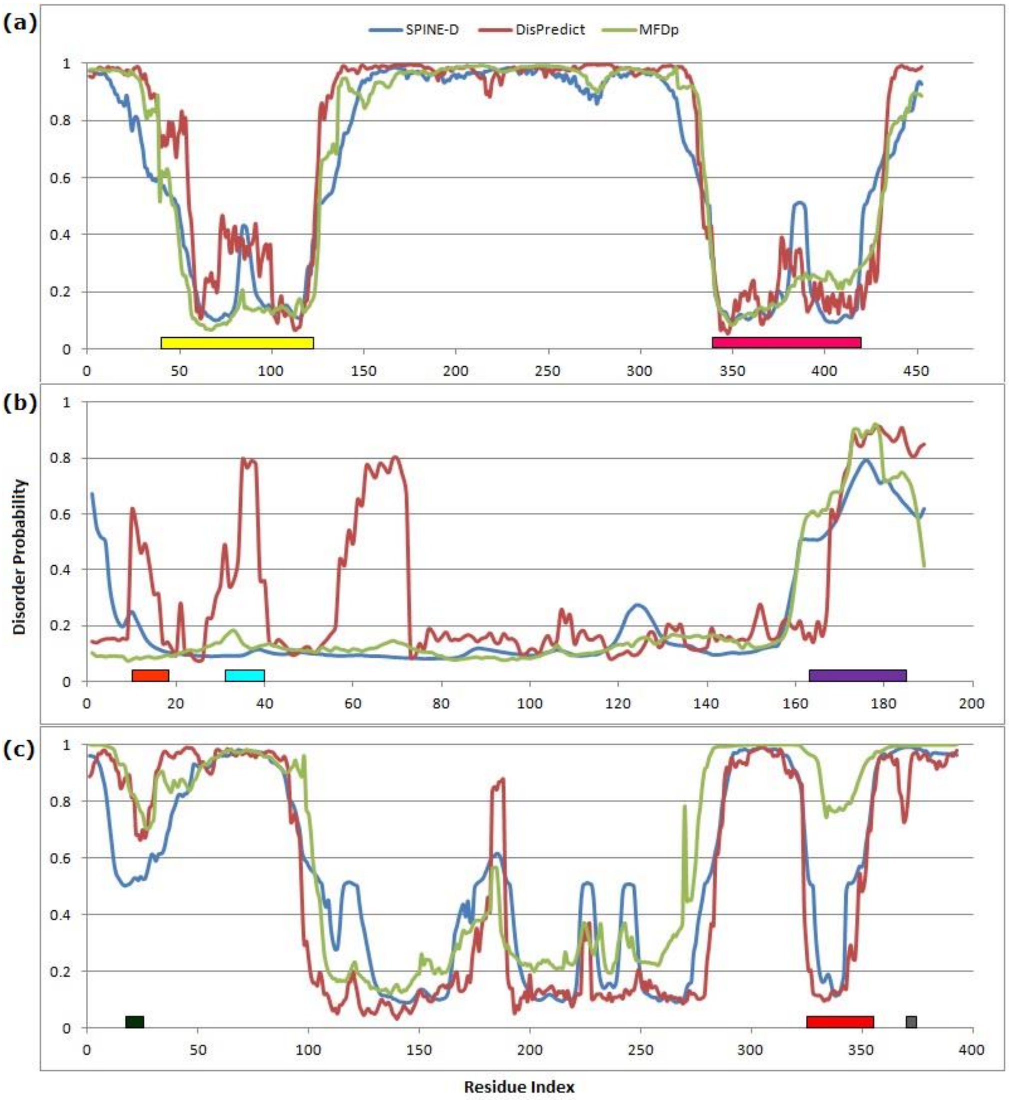


Fig 9. Disorder probability plot for (a) human ETV6 (P41212), (b) human KRAS (P01116) and (c) human p53 (P04637) proteins, given by DisPredict(*red*), SPINE-D (*blue*) and MFDp (*green*). In (P41212, A), the yellow (40 - 124 residues) and pink bar (339 - 420 residues) represent to the PNT domain [187] and ETS DNA binding region [188], respectively. In (P01116, B), the orange (10 - 17 residues), cyan (32 - 40 residues) and purple bar (166 - 185 residues) correspond to the GTP binding region [189], effector region and hypervariable region, respectively. In (P04637, C), the dark green (17 - 25 residues), red (325 - 356 residues) and gray bar (370 - 372 residues) highlight to the TADI motif [190], oligomer region and [KR]-[STA]-K binding motif, respectively.

2.6 Result Analysis and Discussions with DisPredict1.0

In this section, we discuss the different length of disordered regions, structural properties of ordered versus disordered regions along and their correspondence, possible overlap between annotation of ordered and disordered residues in datasets, sequence-based amino acid composition of disordered regions and their effect on prediction.

2.6.1 Distribution of Length Disordered Segments in the Datasets

The performance of DisPredict is also justified by training and testing the predictor with multiple datasets: SL477, SL171 and MxD444, MxD134. The datasets used to train DisPredict (SL477 and MxD444) encompass disorder annotation from several complementary sources (X-ray and NMR defined disorder from PDB and DisProt) as well as disorder region of various lengths.

The SL dataset comprises of 81 full disordered proteins (IDPs) while the rest of the chains contain 928 disordered regions (IDRs). On the other hand, the MxD dataset is composed of 55 full disordered chains, 4 full ordered chains and 385 chains, sharing both structured and disordered regions, which include 730 disordered regions (IDRs). Furthermore, 70% of the IDRs included within partially disordered proteins are short (≤ 30 residues) and 30% of them are long (> 30 residues). This combination of several length disordered regions (**Fig 10**) included within training confirms the consistent performance of DisPredict for disordered regions of all sizes as well as different types of disordered residues.

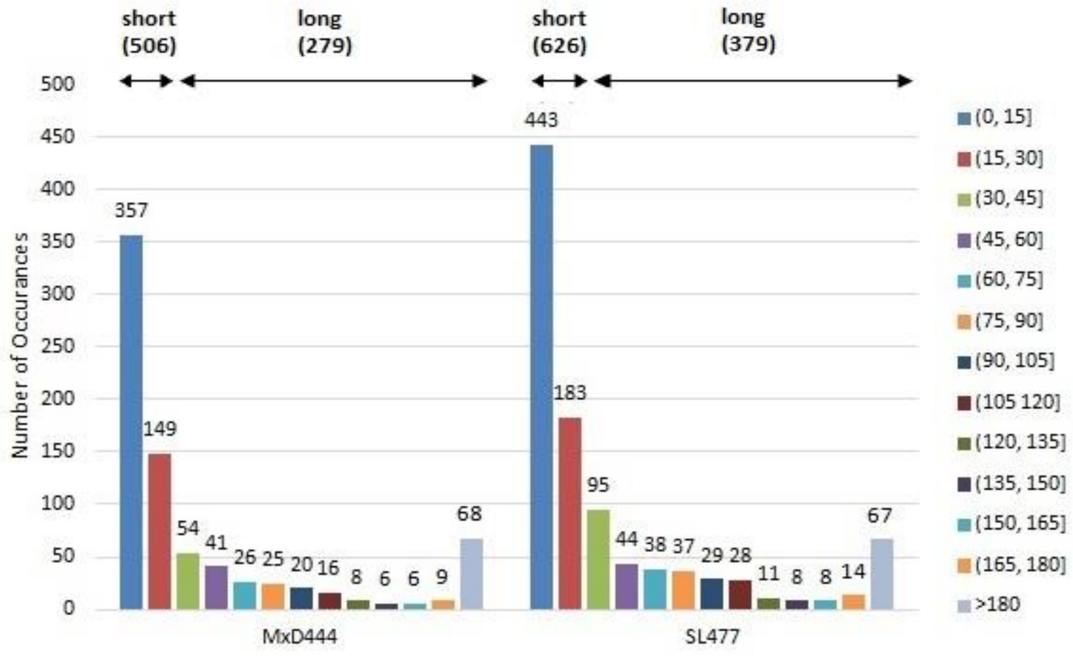
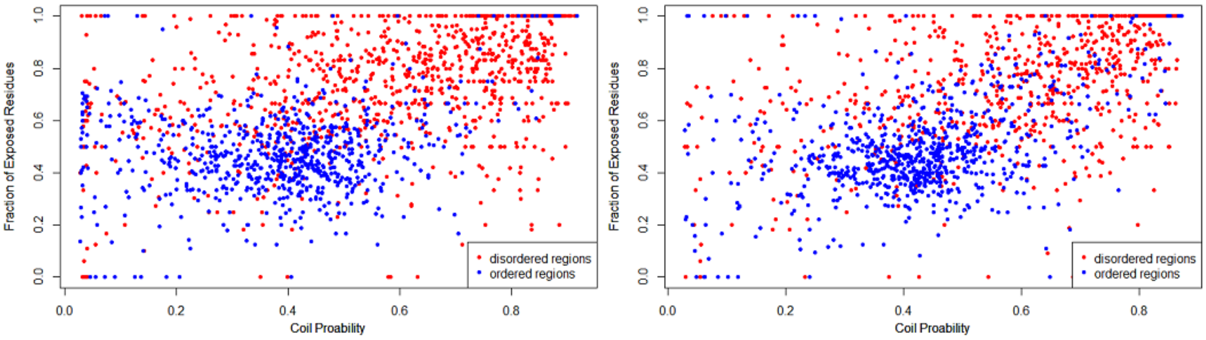


Fig 10. Distribution of disordered regions of different lengths in MxD444 (*left*) and SL477 (*right*) dataset. Legends are shown for different range of lengths （with interval size 15） and each bar is labeled with total number of occurrence of a disordered region of this specific length.

2.6.2 Feature Correlation Plots and Insights into Possible Noise in the Dataset

We observed that regardless of cross validation or independent test, DisPredict's performance is relatively better while it is trained on SL477 dataset than that of MxD444 (**Table 6**). To further insight into this discrepancy, we investigated the correlation of true annotation provided in the dataset with the actual structural characterization of disordered and ordered residues.

Disordered residues are distinguished from ordered residues by low content of secondary structure [8, 20], therefore high probability of coil residues than helical or beta strand residues and disordered regions are likely to have large solvent accessible (exposed) area [162]. We represented the correlation of the fraction of secondary structure content and fraction of exposed residues for disordered and ordered regions of all length in **Fig 11**. We employed the predicted probability of each residue to be coil and predicted per residue solvent accessibility provided by SPINE-X [175] since all residues do not have defined coordinates (structure) to compute secondary structure and solvent accessibility.



（a） SL477

（b） MxD444

Fig 11. Correlation plot between structural characterizations of ordered (*blue*) and disordered (*red*) regions within (a) SL477 and (b) MxD444 dataset. The x−axis and y−axis correspond to the probability of having well defined secondary structure （in terms of probability being coil） and fraction of exposed residues of that region，respectively.

We calculated the average coil probability (P_{coil}) for each ordered or disordered region and computed the fraction of exposed residues with greater than 25% solvent accessibility ($F_{exposed}$) of that region. In this analysis, we discarded 5 residues from the N and C-terminal regions of each protein sequences as they are mostly found on the surface of a protein chain (not buried in the core) and more likely to be affected by the interaction with nearby structured protein, yielding to a highly flexible and dynamic conformation. The plots for both datasets show that the ordered regions are mostly concentrated in the portion with relatively low coil probability, $0.3 \leq P_{coil} < 0.5$ (high content of well-defined helical or strand secondary structured residues) and low exposure, $0.2 \leq F_{exposed} < 0.5$. While on the contrary, the disorder regions are abundant in the area of high coil probability, $0.5 \leq P_{coil} \leq 0.9$ (low content of helical or strand secondary structured residues) and high exposure, $0.5 \leq F_{exposed} \leq 1.0$.

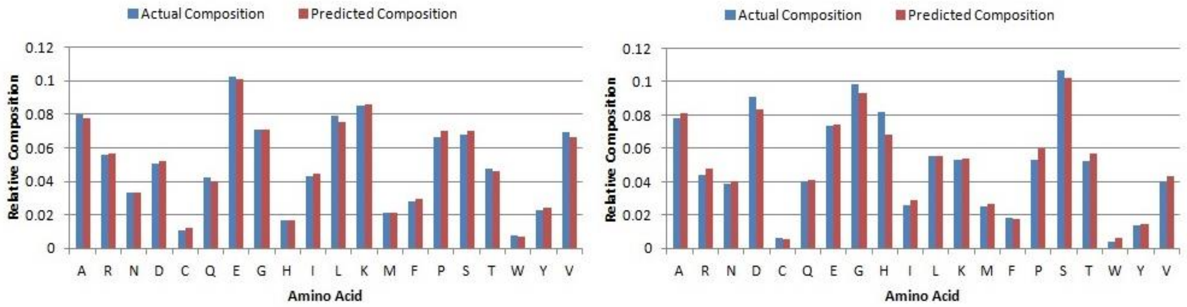
However, we found the intrinsic difference between these two datasets according to their annotation of residues as order and disorder. This difference is evident from the top right location of the correlation plot, $0.6 \leq P_{coil} \leq 0.8$ and $0.4 \leq P_{coil} \leq 0.9$, designated for disordered regions. For SL477 dataset (**Fig 11(a)**), the number disordered regions are predominant over the number of ordered regions in this top right location of disordered regions in the plot. In contrast, the same location of the plot in **Fig 11(b)** is overlapped by both ordered and disordered regions for that of MxD444.

We further quantified the difference as 13% of the data in MxD444's ordered set are more likely to be coil as well as highly exposed while 6% of the data in SL477's ordered set are exposed as well as coil. This higher proportion of misleading annotation in MxD444 dataset contributes relatively lower signal to noise ratio (SNR) of 87/13 compared to 94/6 for SL477 which is the most compelling reason of the better performance of DisPredict in case of training dataset SL477 over MxD444. As the prediction produced by DisPredict is well capable of detecting such discrepancies in the native annotation of the datasets, it can be utilized as a reliable source of correct annotation of the ordered and disordered residues. We should also focus that, a similar proportion of 11% and 13% of the disordered data are also mixed with the ordered residues in the low coil probability region of the plot for both MxD444 and SL477 dataset, respectively.

2.6.3 Residual Composition of IDPs/IDRs and their Effect on Prediction

Here, we investigate that the amino acid residue compositions in IDPs/IDRs that may vary in different datasets as well as within short (≤ 30 residues) and long (> 30 residues) disordered regions [8, 192]. Specifically, short disordered regions are enriched with aspartic acid (D), glycine (G) and serine (S). On the contrary, glutamic acid (E), lysine (K) and proline (P) are likely to be abundant in long disordered regions.

To give further insight into this residue composition and confirm the ability of DisPredict to detect the residue preferences of short and long disordered regions, we determined the residual composition profile for our two test datasets, SL171 (**Fig 10(a)**) and MxD134 (**Fig 10(b)**). It is to be noted that, these two datasets contain experimentally annotated disorder from two different sources. SL171 contains sequences with disorder annotation from DisProt while MxD134 contains that from PDB. The composition profile consists of the actual ratio (r_a) and predicted ratio (r_p) of each amino acid type out of total annotated and predicted disordered residues.



（a） SL171 （b） MxD134

Fig 12. Percentage of amino acid type residues in actual composition (*blue* or *left adjacent bar*) and predicted composition (*red* or *right adjacent bar*) of (a) SL171 and (b) MxD134 dataset. The *x*-axis and *y*-axis represent the 20 different amino acids and their relative proportions in the composition.

The composition profile in **Fig 12(a)** demonstrates that SL171's disordered residue set accommodates relatively higher ratio of amino acid type E (10%) and K (9%), which are long disorder prone residues. In contrast, MxD134's disordered residue set, shown in **Fig 12(b)** is enriched with high ratio of amino acid type S (11%), G (10%) and D (9%), known as short-disorder-prone residues. Another significant difference between the intrinsic compositions of these two datasets is in the proportion of histidine (H). Disorder annotation from PDB includes higher ratio of H-tag (8% in MxD134, compared to 2% in SL171), which is sometimes used for protein purification [123]. The predicted proportion of all these amino acids given by DisPredict ensures its capability of detecting residues in disordered region of all length accurately with no significant over prediction. Moreover, DisPredict could also accurately predict methionine (M) at highly flexible N-terminal region.

To further quantify DisPredict's performance in detecting residue composition, we evaluated the Root Mean Square Difference (RMSE) and Pearson Correlation Coefficient (PCC) between actual and predicted ratio (r_a and r_p) for each amino acid type. For MxD134 test dataset, we found RMSE of 0.0046, which was comparatively higher than the RMSE value computed for SL171 which equals to 0.0018. However, the correspondence between actual composition and predicted composition by DisPredict measured with PCC (P-Value < 10^{-5}) was found equally positive, 0.9976 and 0.9897 for SL171 and MxD134 dataset, respectively. It is important to note that, this consistent result is corresponding to the independent test where the dataset used to train DisPredict shared significantly low sequence identity (at most 10%) with test dataset, which once again implicates the strength of the classification methodology of DisPredict.

2.7 DisPredict (version 1.1)

Here, we propose an improved version of DisPredict [10], named DisPredict1.1 [193], which includes a post-processing of probability outputs given by initial DisPredict1.0 [10] to generate more accurate annotation of disordered protein residues. DisPredict1.1 applies window based averaging of per residue probability, generated DisPredict1.0 (described in **Section 2.4**) to reduce possible less-than-ideal noisy prediction output by DisPredict1.0.

In DisPredict1.0, the input feature set and model development process are kept similar to those of DisPredict1.0 to consistently quantify the improvement gained only due to smoothing the outputs. However, we evaluated DisPredict1.1 on a new independent test set. DisPredict1.1 was found better than DisPredict1.0 as well as competitive with 20 other existing predictors.

2.7.1 Datasets

DisPredict1.1 was separately trained using SL477 and MxD444 (*see* **Section 2.3.2**) and similar to DisPredict1.0, was tested using SL171 and MxD134 (*see* **Section 2.3.2**). However, we downloaded an additional dataset [165] to evaluate DisPredict1.1.

Dataset DP_NEW: This dataset was originally developed as part of MFDp2 [165]. DP_NEW dataset encompasses disorder annotation from PDB REMARK 465 as well as curated annotation from DisProt. It combines 43 protein chains with curated annotation of DisProt and 62 chains annotated by PDB. Moreover, this dataset contains 115 short disordered regions (less than 30 residues) and 28 long disordered regions (greater than or equal to 30 residues) combined with 17 full ordered and disordered proteins. BLATCLUST was used to filter the resulting dataset so that no sequence is more than 25% similar to MxD dataset which resulted another independent test dataset of 105 protein chains. DP_NEW dataset comprises of 31,511 residues that combines 4640 (about 14.7%) disordered residues, 17,798 ordered residues (about 56.4%) and 9,073 unknown residues (about 28.7%).

2.7.2 Predictor Framework

DisPredict1.1 follows our initially designed SVM based classifier model of DisPredict [10] (*see* **Section 2.4**) for prediction of per residue binary annotation (order or disorder) and assigning two real values as the probability score of being order or disorder. SVM with RBF kernel is used to develop the predictor model. The predictor consists of two layers. The selection of parameter values (C and γ) is done with optimization on accuracy (fraction of correctly predicted residues) by grid search, which is guided by 5-fold cross validation. The real values are binarized using a natural threshold equal to 0.5, 0.5 ≤ range ≤ 1.0 is considered as disordered probability and 0.0 ≤ range < 0.5 is considered as ordered probability. We utilized LIBSVM [184] for SVM parameterization and model generation.

In DisPredict1.1, we processed the probabilities by taking the average of the resulting probabilities with a sliding window of 29 residues (14 residues on either side of the target residue) and converted the scores into binary annotation using the same threshold of 0.5. We selected the window size which provided us the highest MCC scores in performance evaluation. With this post processing step, DisPredict1.1 applies a smoothing on the probabilities to take the impact of relative type (order or disorder) of the neighboring residues while assigning the score for a target residue which improves both MCC and AUC scores achieved by DisPredict1.0. However, we have not applied this smoothing of probability for the N and C terminal region due to their highly flexible and dynamic conformation.

2.7.3 Evaluation of DisPredict1.1

2.7.3.1 *Comparison with DisPredict1.0*

DisPredcit1.1 is evaluated using a similar set of criteria described in **Section 2.3.4**. With the additional correction of predicted probabilities by sliding window based averaging and transforming the resulting probabilities into binary annotation, DisPredict1.1 outperforms DisPredict1.0 [10] both in binary annotation and probability prediction.

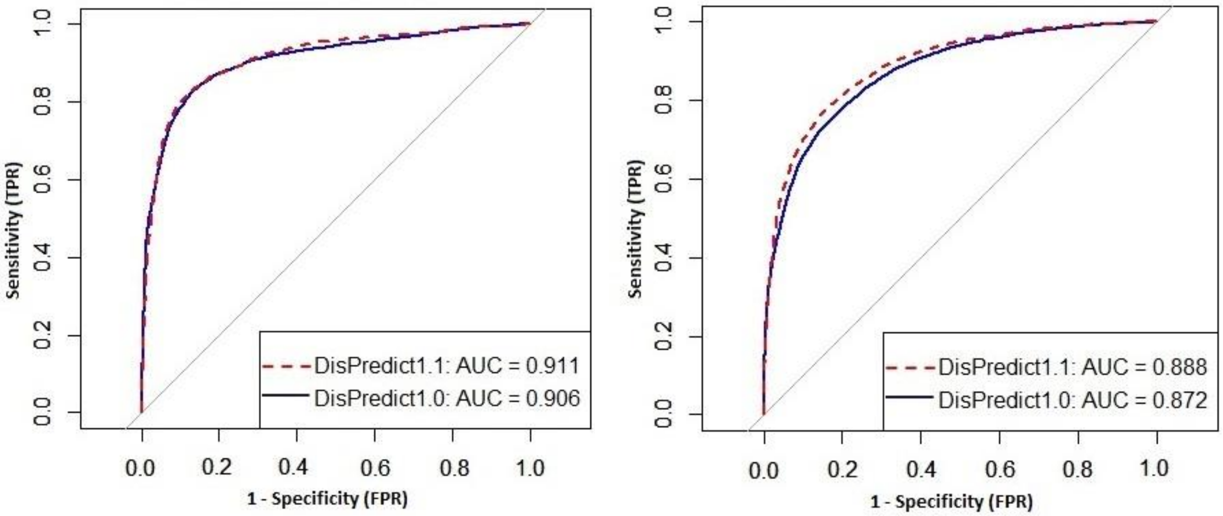
Table 9. Performance comparison between DisPredict1.0 and DisPredict1.1.

Predictor[1]	Test Set	SENS	SPEC	ACC	S_w	PPV	MCC	AUC	MAE
DisPredict1.1 (SL477)[2]	MxD134	**0.745**	**0.928**	**0.837**	**0.673**	**0.591**	**0.611**	**0.911**	0.083
DisPredict1.0 (SL477)[3]	MxD134	0.744	0.923	0.833	0.667	0.574	0.598	0.906	**0.023**
DisPredict1.1 (MxD444)[2]	SL171	0.644	0.926	0.785	0.57	0.834	0.61	0.888	0.032
DisPredict1.0 (MxD444)[3]	SL171	0.718	0.86	0.789	0.577	0.748	0.583	0.872	0.151

[1] The predictor name is specified with the corresponding training dataset in parenthesis. The training was done with window size twenty one and optimal SVM parameters.
[2] Probabilities smoothed with a sliding window size 29.
[3] No probability smoothing.

Table 9 further illustrates this comparison of results in the case of independent tests of the predictor two (MxD134 and SL171) datasets. DisPredict1.1 improved the performance for binary disorder or order prediction by 0.48%, 0.89%, 2.96%, 2.17% in terms of accuracy, S_w, precision and MCC, respectively during the test by MxD134 dataset. On the other hand, while testing with SL171 dataset, there are significant increase by 6.38% and 4.63% in precision and MCC, respectively. However, the accuracy decreased slightly which caused by the decrease of SENS along with significant increase in SPEC.

DisPredict1.1 also provided consistent improvement in assigning per residue confidence score with 0.55% and 1.83% increase in AUC score for MxD134 and SL171 datasets. This improvement is further analyzed with the ROC curves in **Fig 13** which depicts better correlation between sensitivity and specificity with smoothing. Overall, the consistent performance for two different test sets justifies rigorous training and precise methodology.



（a） Train： SL477， Test： MxD134

（b） Train： MxD444， Test： SL171

Fig 13. Comparison of ROC curves given by DisPredict1.0 and DisPredict1.1. （a） Train by SL477， test by MxD134 and （b） Train by MxD444， test by SL171. In each figure, the solid （*blue*） and dotted （*red*） curve corresponds to the performance of DisPredict1.0 and DisPredict1.1, respectively. The AUC values are given in the legend according to the respective ROC.

2.7.3.2 *Comparison with Other Existing Predictors*

Here, we compare the performance of DisPredict1.0 and DisPredict1.1 against twenty existing methods (including sub versions of some tools for different types of disorder) which cover various categories of disorder prediction methods using different machine learning algorithms. These methods include DISOPRED [126], 3 versions of ESpritz (X, N and D) [122], PROFbval [194], PrDOS [195], NORSnet [120], PreDisOrder [155], 2 versions of IUPred (short and long) [186], Ucon [152], DISOclust [185], 2 versions of CSpritz (short and long) [160], MD [162], SPINE-D [123], MFDp [164], PONRD-FIT [63] and very recent 2 versions of MFDp2 (with and without BLAST) [165].

Table 10. Performance comparison of DisPredict1.0 and DisPredict1.1 with 20 existing predictors when residues without actual annotation are assumed as ordered.

Method[a]	**SENS**	**SPEC**	**MCC**	**AUC**	**MAE**
DisPredict1.1 (SL477)	77.3	83.8	**0.499**	0.857	0.081
DisPredict1.1 (MxD444)	66.2	**88.1**	0.482	**0.862**	**0.04**
MFDp2	75.9	83.2	0.479	0.862	0.153
DisPredict1.0 (SL477)	**77.4**	82.2	0.478	0.85	0.092
MFDp2 (no blast)	75.4	83.2	0.475	0.86	0.153
MFDp	80.9	79.3	0.466	0.85	0.174
DisPredict1.0 (MxD444)	67.8	86.3	0.466	0.845	0.054
Cspritz L	83.5	77.5	0.463	0.87	0.242
MD	72.6	79.9	0.414	0.829	0.235
Espritz X	53.8	88.7	0.394	0.801	0.139
Cspritz S	73.5	77.2	0.39	0.823	0.209
PrDos*	55.8	86.8	0.388	0.818	0.137
PONDR-FIT	66.3	81.5	0.387	0.8	0.162
SPINE-D	78.4	72.9	0.381	0.823	0.204
IUPreD L	60.4	84.4	0.38	0.788	0.13
PreDisorder*	74.5	74.1	0.374	0.797	0.234
DISOPRED2	65.6	80.5	0.37	0.797	0.153
IUPreD S	54.5	86.7	0.368	0.782	0.133
Espritz D	40.9	92.0	0.349	0.827	0.186
DISOCLUST	75.3	71.3	0.343	0.803	0.19
Espritz N	60.2	80.5	0.329	0.785	0.168
NORSnet	47.3	87.6	0.323	0.761	0.172
UCON	60.5	76.6	0.289	0.732	0.179
PROFBVaL	52.8	65.1	0.13	0.631	0.307

[a] The methods are sorted according to MCC.
For each metric, our best result is marked in bold and previously found best result is underlined.
[*] According to MFDp2 [165], PrDos and PreDisorder failed for one chain and were evaluated on 104 chains.

To compare consistently, we collected the performances of these methods on DP_NEW benchmark dataset from MFDp2 article [165] and evaluated the performance of DisPredict1.0 and DisPredict1.1 on same dataset. Note that, DP_NEW dataset contains about 28.7% residues annotated as unknown. To remain consistent, we also evaluated our predictors assuming the unknown residues as order at first and then discarding the unknown residues. Comparisons among different predictors at both level are presented quantitatively in **Table 10** and **Table 11** in terms of SENS, SPEC, MCC, AUC, MAE and PCC. Here, SENS, SPEC, MCC and AUC are used to determine the performance in binary annotation prediction and probability prediction at residue level, while MAE indicates the performance of disorder prediction in content level.

Table 11. Performance comparison of DisPredict1.0 and DisPredict1.1 with 20 existing predictors when residues without actual annotation are discarded.

Method[a]	**SENS**	**SPEC**	**MCC**	**AUC**
DisPredict1.1 (SL477)	75.9	95.3	0.729	0.94
DisPredict1.1 (MxD444)	75.4	95.3	0.725	0.938
MFDp2	77.3	94	**0.711**	**0.925**
DisPredict1.0 (SL477)	80.9	92.2	0.704	0.925
MFDp2 (no blast)	66.2	**96.4**	0.683	0.912
MFDp	**77.4**	92.2	0.677	0.914
DisPredict1.0 (MxD444)	67.9	94	0.642	0.89
Cspritz L	83.5	85.9	0.621	0.909
MD	65.6	93.6	0.614	0.88
Espritz X	60.4	94.3	0.588	0.851
Cspritz S	75.3	87.4	0.581	0.904
PrDos*	72.6	88.4	0.576	0.873
PONDR-FIT	55.8	95.4	0.576	0.883
SPINE-D	78.4	85.4	0.575	0.893
IUPreD L	66.3	90.3	0.558	0.85
PreDisorder*	47.3	96.7	0.54	0.834
DISOPRED2	53.8	94.5	0.54	0.845
IUPreD S	54.5	93.6	0.525	0.83
Espritz D	73.5	83.6	0.512	0.857
DISOCLUST	74.5	82.4	0.503	0.85
Espritz N	60.2	89.4	0.492	0.844
NORSnet	40.9	94.4	0.426	0.866
UCON	60.5	84.4	0.42	0.78
PROFBVaL	52.8	67.2	0.167	0.647

[a] The methods are sorted according to MCC.
For each metric, our best result is marked in bold and previously found best result is underlined.
* According to MFDp2 [165], PrDos and PreDisorder failed for one chain and were evaluated on 104 chains.

Table 10 shows that DisPredict1.1 results highest MCC among all the other methods and outperforms the previous best result given by MFDp2 [165] by 4.18% when trained on SL477 dataset and by 0.63%

when trained on MxD444 dataset. The AUC score of DisPredict1.1 was also competitive. The best score of specificity was given by Espritz D at the cost of very low sensitivity. However, both sensitivity and specificity given by DisPredict1.1 are comparable.

Table 11 shows that all the scores provided by DisPredict are competitive and outperform 18 existing predictors in terms of MCC and AUC except MFDp2. However, MFDp2 does not consider relatively short disordered regions (less than 4 residues) in the evaluation, while DisPredict is evaluated for all types and length of disordered regions. We consider the short disordered regions since they are biologically significant and our result provides us with evidence that the methodology of our predictor gives promising performance for all types of disorder.

2.8 Summary and Conclusions

In this chapter, we described a disordered protein prediction framework, which utilizes a canonical support vector machine with RBF kernel and includes useful and advanced features for predicting disordered residues, called DisPredict. DisPredict not only generates the binary class annotation for ordered and disordered residues but also provides order-disorder probabilities that can be treated as the confidence level of the prediction as well. DisPredict is implemented in C and the code is publicly available in open source form at https://github.com/tamjidul/DisPredict_v1.0.

The DisPredict outperformed other existing top performing predictors both in predicting binary annotation and probability. The competitive performance of DisPredict is mainly due to the use of a novel methodology that incorporates firstly, radial basis kernel function (RBF) that can implicitly map the feature space in infinite dimension, secondly and most importantly the optimization of the parameters and thirdly, the novel features that assisted in determining an optimal as well as effective class separating hyperplane.

DisPredict was guided by a comprehensive set of features that captured the sequential (amino acid composition) and structural characterization of ordered and disordered residues or, proteins. We used SPINE X [175] to generate the secondary structure related fine features. The distinguishing property of our feature set in comparison with existing predictors is the inclusion of monogram (MG) and bigram (BG), computed from PSSM. When a region of a protein is evolutionarily conserved in a fold, then all the proteins within that fold are likely to have a conserved group of MGs and BGs. As some intrinsic disordered regions are conserved, addition of these features provides important structural evolutionary characteristics. By determining the appropriate window size, we have also included the effect of optimal interactions due to the contacts among neighboring residues.

While DisPredict1.0 was found comparable with two other existing predictors, DisPredict1.1 (a new version) was found competitive with 20 other state-of-the-art predictors. In DisPredict1.1, an additional post processing of probabilities with window based averaging was performed to correct the binary order or disorder annotation accordingly, which is found effective to reduce the noise in prediction as such averaging captures the impact of the relative structured or unstructured status of neighboring residues. In addition to that, our case studies ensure biologically relevant performances of DisPredict.

Finally, accurate prediction of disorder has useful implications in proteomic studies due to its direct involvement in the function of a protein. Successful detection of disordered region(s) of a protein is considered to be the first step in drug design to combat critical diseases. We have built DisPredict using the canonical SVM classifier with RBF kernel and established it as a successful predictor of disorder by utilizing the benchmark datasets, which we believe will be a useful tool in the study of proteomes and their functions.

One interesting observation is that predicted disorder probability can serve as a useful feature for sequence-based prediction of other structural properties of protein. For example, we have developed an improved accessible surface area predictor [15] and balanced secondary structure predictor [196] of protein residues from sequence alone, where the output of DisPredict was one of the major feature. The accessible surface area predictor is discussed under this thesis in **Chapter 3**. With an aim to contribute further in this field of study, we have extracted novel features to protein disorder and extended our framework, which is discussed in **Chapter 4**.

Chapter 3

REGAd^{3}p: A Predictor of Protein Accessible Surface Area

— A Framework to Predict ASA using Polynomial Kernel and Regularized Regression with Reinforcement Learning and its Application

Proteins consist of a linear chain of amino acid residues connected by peptide bonds to adjacent amino acid residues. Accessible Surface Area (ASA) is a one-dimensional structural property of amino acid residues of protein that measure their level of exposure to solvent (like, water) in a structure. Proteins perform a vast array of functions within the living organisms which are governed by their amino acid residue sequence and the 3-dimensional structures defined by the sequence. Proteins interact with appropriate partners to perform specific functions. Surface area of amino acid residues determines the interaction pane, which eventually play an important role in binding mechanisms and structures and functions of proteins. ASA has been helpful in understanding the 3-dimensional structure and function of a protein, acting as high impact feature in secondary structure prediction, disorder prediction, binding region identification and fold recognition applications. Thus, accurate prediction of accessible surface area (ASA) in real-value from protein sequence alone has wide application in the field of bioinformatics and computational biology.

To enhance and support broad applications of ASA, we have made an attempt to improve the prediction accuracy of absolute accessible surface area of protein residues by developing a new predictor paradigm, namely *REGAd^{3}p* [15], for real value prediction, discussed in this chapter. REGAd^{3}p exerts the **R**egularized **E**xact regression, which is reinforced by **G**enetic **A**lgorithm and incorporates **d**egree **3** **p**olynomial kernel function. While the higher degree polynomial kernel was applied to properly fit the high-dimensional data, regularization was incorporated to resist over-fitting. Furthermore, we applied genetic algorithm (GA) to optimize the weights computed by regularized regression. The kernel that we applied, was selected based on optimum values of *Mean Absolute Error* (MAE) and *Pearson Correlation Coefficient* (PCC).

The ASA prediction paradigm was trained and tested using a new benchmark dataset, mined under this work. We achieved maximum Pearson Correlation Coefficient (PCC) of 0.76 and 1.45% improved PCC when compared with existing state-of-the-art predictor in ASA prediction on independent test set. Further, we presented a rigorous analysis of the quality of the predicted ASA by REGAd^{3}p in terms of different amino acids and their physical properties, secondary structure components and range of ASA values. Another major contribution of this work is that we modeled the error between actual and predicted ASA in terms of energy and combined this energy linearly with a knowledge-based energy function 3DIGARS [16] which resulted in an effective energy function, namely 3DIGARS2.0 [15]. The outline of this chapter is as follows.

- We start by giving the background information about ASA and its implications, motivation behind developing a predictor and review of existing ASA prediction techniques in Section 3.1.
- In Section 3.2, we describe the experimental materials, such as data collection and mining process, input features used to train the ASA predictor, and the criteria to evaluate and compare the predictor.
- Section 3.3 describes the design and development of the new real-value predictor framework, REGAd^{3}p using several machine learning techniques.
- We report the results of feature selection, kernel selection, performance of the final ASA prediction model, and its comparison with existing predictor in Section 3.4.
- In Section 3.5, we analyzed the quality of predicted ASA in terms of physical and structural properties of protein residues.
- In Section 3.6, we discuss about the application of the predicted ASA to improve an energy function.
- Finally, we conclude in Section 3.7 including brief future research directions.

3.1 Background and Motivation

Most protein molecules have a hydrophobic core, which is not accessible to solvent and a polar surface in contact with the environment. Proteins perform their functions mostly through interactions using their solvent exposed surface with their partners for transmission or reception of signals. A protein residue, in its three-dimensional conformation, can be surrounded by other residues in the chain. In contrast, a residue can have a part of it accessible to the residues of the same chain or to the residues of other chains for interaction. The parameter to measure the level of interaction of a residue is thus can be determined by the accessible surface area (ASA), usually described in units of square Angstroms.

Lee and Richards [197] first described the ASA, often called as *Lee-Richards Molecular Surface* which can be calculated using the "rolling ball" algorithm of Shrake and Rupley [198], fast and analytical Power Diagram [199] technique or, can be approximated using LCPO method [200]. The van der Waals surface as defined by the atomic radii [201], whereas the solvent ASA is the surface area of a biomolecule (protein or protein residues) that is accessible to a spherical solvent while probing the surface of that molecule [197, 202]. **Fig 14** illustrates the ASA of a protein molecule.

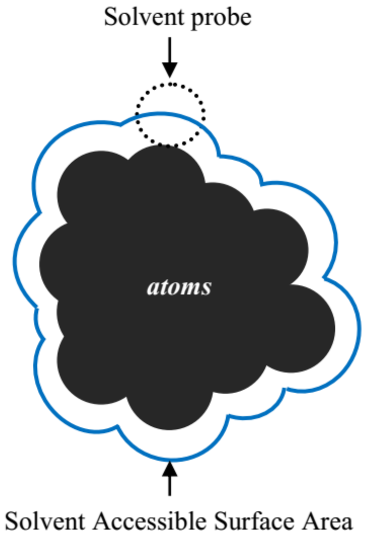
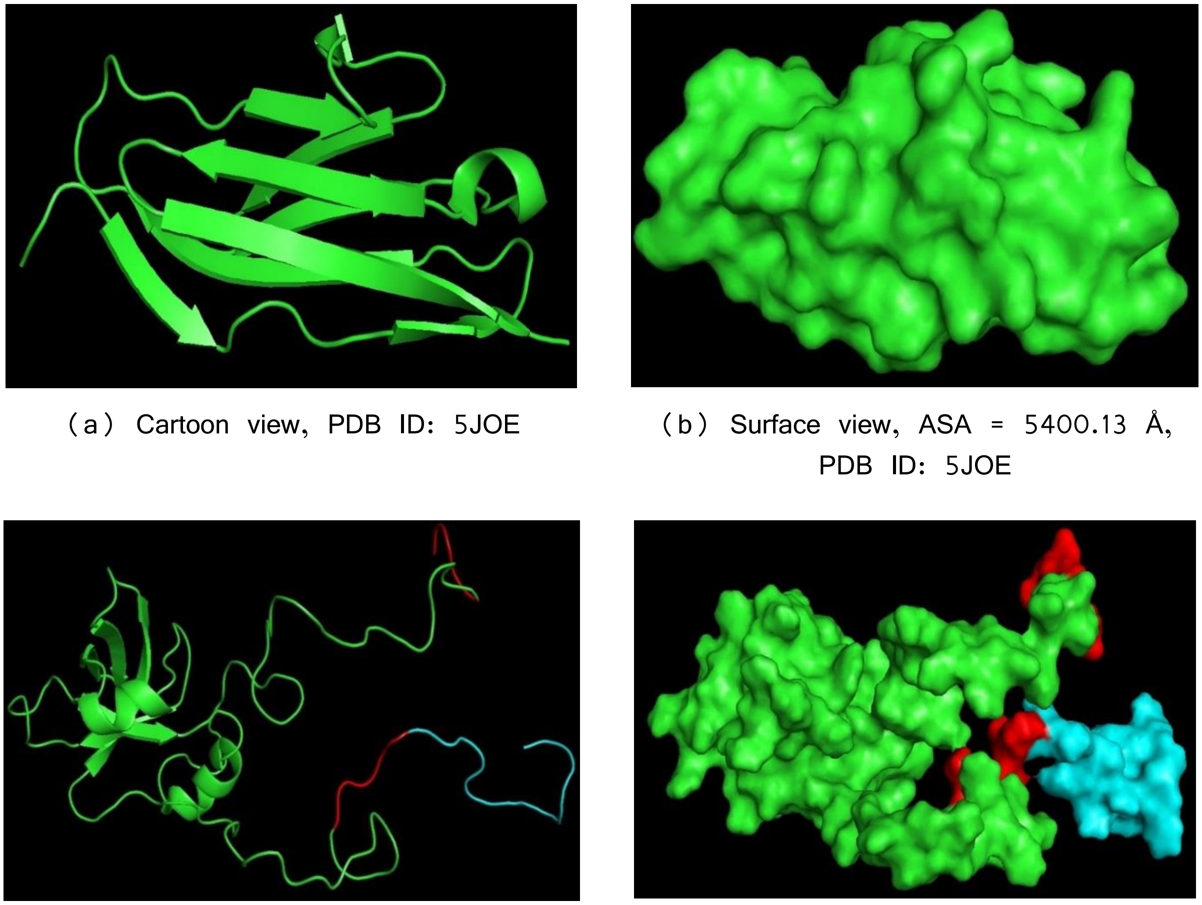


Fig 14. Accessible Surface Area (ASA) of Protein. The dark central area, composed of atoms, can be thought of a 3D protein and the circumference of the area is the van der Waals surface area. The outline （blue） around that area is the accessible surface area of the protein.

Function of a protein is found to be closely coupled with the ASA of its residues as it defines the interaction pane. The wide conformational dynamics of proteins, which is often exemplified by intrinsically disordered regions and thermal fluctuations (B-factor) of a protein, is crucial for their diverse functionalities and is found to be strongly correlated with the ASA of each of the residue of a protein [203, 204]. Surface areas, often in the form of exposed residues, are directly involved in the protein-protein interaction [205, 206]. ASA is also found to play an important role in the binding mechanism of proteins in the literature [207]. Thus, the measurement of the ASA is essential in understanding the 3-dimensional structure and function of a protein [208, 209]. It is well known that the hydrophobic effect is the major factor that drives a protein to collapse and fold, which is directly related to ASA [210, 211].

Fig 15 illustrates the protein ASA-structure relations with two collected samples. **Fig 15 (a)** and **(b)** shows the structure of human I81 domain from titin [212] (PDB ID: 5JOE) with secondary structure and surface views, respectively. The protein has a stable structure with helical and beta components, and has accessible surface area of 5400.13 square Angstroms, calculated using GETAREA[14] web server [213]. On the other hand, **Fig 15 (c)** and **(d)** portray the structure of elF1a (*green*) [214], which has an extended coil-like N- and C-terminals (*red*) where N-terminal is connected with histidine tag (*cyan*). The protein has a large binding surface area with ASA equal to 14400.45 square Angstroms. This illustration shows that ASA of a protein is directly related to its structure.



（a） Cartoon view, PDB ID: 5JOE

（b） Surface view, ASA = 5400.13 Å, PDB ID: 5JOE

（c） Cartoon view, PDB ID: 1D7Q

（d） Surface view, ASA = 14400.45 Å, PDB ID: 1D7Q

Fig 15. ASA-structure relationship of protein. （a） – （b） PDB ID: 5JOE [212], crystal structure of human I81 domain from titin with stable components and ASA of 5400.13 square Angstroms. （c） – （d） PDBID: 1D7Q [214], solution structure of elF1a with coil-like components and large RNA binding surface with ASA of 14400.45 square Angstroms.

Moreover, ASA has been found to be an important feature for secondary structure prediction, intrinsic disorder prediction, binding region identification, hot-spot prediction [215], domain boundary prediction [216], fold recognition and protein function identification [217-221]. Importantly, accurate prediction of

[14] GETAREA: http://curie.utmb.edu/getarea.html

surface area of protein residues elevates the success in *ab initio* protein structure prediction [222] and accurate energy function development for correct discrimination of native conformation from the decoys [223, 224]. The prediction of real valued accessible surface area from primary protein sequences alone is challenging, yet rewarding in the field of structural biology. We responded to this challenge by developing tools to find accurate ASA from a protein sequence alone and validated the outcome with test dataset as well as by significantly improving an energy function application.

Effective energy function is an essential component of protein's structure prediction for which homologous templates are absent. The major theme of the energy function developed to date are based on the fact that protein in their native state gains the lowest free energy compared to its other possible states. The developed Energy functions can be categorized into two different types [225-229]: first, physical-based potential, based on empirical molecular mechanics force fields [230, 231] and second, knowledge-based potentials or empirical potential energy function, based on statistical analysis of known proteins [232-237]. Knowledge-based potentials can be more successful over physical-based potential [238] as it uses growing number of experimental (known) protein structures, can capture unrecognized forces and the execution is much faster compared to the molecular mechanics based tools. Under this work, we compute predicted accessible surface area based energy component and integrate it with hydrophobic-hydrophilic model (HP model) based 3-Dimensional Ideal Gas Reference State (3DIGARS) potential [239] towards a better energy function application.

3.1.1 Role of *in silico* ASA Prediction

There exist some tools that can computationally assign secondary structure to proteins given its three-dimensional structure such as in PDB file format. From the coordinates of the atoms, these secondary structure assignment methods can produce coarse-grained descriptions of the local backbone structure, such as helical, beta or coil conformation as well as fine-grained descriptions, like ASA, dihedral angles (phi and psi).

One of the widely used assignment tools is DSSP [51], designed by Kabsch and Sander, which assigns secondary structure descriptions including ASA according the pattern of hydrogen bonds. STRIDE [240] is a knowledge-based assignment method that considers hydrogen bond as well as backbone torsion angles. KAKSI [241] determines secondary structure related descriptors using Cα distances and ϕ/Ψ dihedral angles. A surface area routine, GETAREA [213] calculates ASA, solvent energy and their gradients of a macromolecule using an analytical method from the atomic coordinates. The other tools in this study includes, POPS [242], a parameter optimized approach for atomic and reside-wise ASA calculation and FreeASA [243], which is an improved technique over NACCESS [244].

While these computational methods give us prior knowledge about ASA for a protein from its structure model, they cannot serve the purpose when the structure is not known, such as for intrinsically disordered protein. The NCBI RefSeq [245] database contains approximately 677 times higher number of protein sequence than the available protein structures solved at current date of April 03, 2017, deposited in PDB [61]. Thus, to analyze the local structure and function of a protein given only the sequence information, an *in silico* predictor becomes essential to predict structural properties such as ASA. With an effective pattern recognition algorithm and appropriate feature-set, a predictor can be trained using available proteins-structure relational datasets to generate protein structure and structural properties given only the sequence information.

3.1.2 Review of ASA Prediction

The solvent accessibility prediction has been studied in two forms: firstly, binary or, multiclass classification problem and, secondly, real-value prediction problem. Machine learning based methods, such as neural network, liner or polynomial regression methods, k-nearest neighbor, support vector machine, random forest etc., are some of the successful methods for ASA or relative accessibility prediction given the sequence information only.

The neural network based classifier of protein residues depending on their ASA into multiple states includes, 10-state classifier by Rost and Sander [246], 2-state model of JPred [247], bidirectional recurrent network model of [248], binary (buried/exposed) and ternary (buried/intermediate/exposed) models by Holbrook *et al* [249], NETASA [250] that classifies residues into three-states using multiple thresholds to categorize buried and exposed residues. Li and Pan [251] developed a two-state defining solvent accessibility using multiple linear regression. Using a cut-off value of 15%, a support vector machine was used in [252] to predict the exposed and buried state of protein residues, whereas threshold values of 25%, 16%, 5%, and 0% were adopted in [253]. The Bayesian method based framework was developed to take into account local interactions among amino acid residues, by extracting the information from single sequences or multiple sequence alignments to obtain posterior probabilities for RSA prediction in [254]. A fuzzy *k*-nearest neighbor was also used for 2-state and 3-state prediction of ASA [255].

The first real value prediction of ASA was conducted by Ahmad *et al*. [256] in 2003 using a neural network. Other state-of-the-art work for real value prediction of accessible surface area using artificial neural network includes Real-Spine [178, 257], SPINE-X [175], ASAquick [258], and recently developed

SPIDER package uses deep neural network [259]. SABLE [260] is another neural network based regression technique for predicting ASA, secondary structure and transmembrane domain from protein sequence. Wang *et al*. [261] developed a multiple linear regression model to predict ASA from protein sequence and evolutionary information. The support vector machine based regression technique was also used in two predictors [262, 263].

However, the real-value prediction approach is preferred over the former since the residue's solvent accessible surface area tends to vary largely due to their free movement in 3-dimensional space [204, 256]. Direct prediction of a continuously varying ASA as a real value reduces the inherent error introduced within the approaches, like binary state classification of the residues (exposed or, buried) or, multi-class classification using different choice of thresholds.

3.1.3 Our Contributions

It is common in the literature to express and predict ASA in the form of relative accessible surface area (RSA) which is calculated by normalizing the absolute ASA by residue specific maximum values of ASA found in the dataset or, ASA of the extended tripeptide conformation, such as, Ala-X-Ala or, Gly-X-Gly. However, depending on different normalizing factors, RSA values vary for same amino acid which makes the comparison of performance with existing predictors inconsistent. To overcome such inconsistencies, we avoided normalizing the ASA values. Instead, we directly predicted the absolute accessible area of the protein residues.

We introduced a new benchmark dataset in this work collected from Protein Data Bank (PDB) consisting of 1299 protein sequence, called as Secondary Structure Dataset (SSD1299), with 25% sequence identity cut-off. We tested our predictor (REGAd^{3}p) with three blind, harder test datasets and compared our predictor's performance on ASA prediction with SPINE-X [175]. The improved performance of our REGAd^{3}p in all cases suggests that integrating GA optimization with regression resulted a robust real value predictor. Furthermore, we developed a secondary structure predictor model for generating three dimensional secondary structure profile (helix, beta and coil probabilities) which is used as features for the ASA prediction using support vector machine package, the LIBSVM [184]. Finally, we applied the predicted ASA values to improve the accuracy of the energy function, 3DIGARS, which actually resulted in outperforming all the state-of-the-art energy functions significantly.

3.2 Materials for ASA Prediction

In this section, we describe the datasets and the feature set for building the ASA predictor and evaluation-steps to measure the effectiveness of our approach.

2.3.5 Datasets

We prepared a new dataset from Protein Data Bank (PDB) [61] which is referred to as the Secondary Structure Dataset (SSD1299), consisting of 1,299 protein sequences. Initially, we collected 2,793 protein chains (both single and multiple chain) from PDB with following specifications:

(*a*) Solved by X-ray crystallography

(*b*) Resolution ≤ 1.5 Å

(*c*) Chain length ≥ 40 residues and

(*d*) 30% sequence identity cut-off.

We further carried out three step refinement of this dataset: (*i*) we filtered the dataset so that the pair-wise sequence similarity is no more than 25% using BLASTCLUST; (*ii*) we discarded the protein sequences that contain unknown amino acids labelled as 'X' as the physical properties of this amino acid is unknown and (*iii*) we removed the sequences containing amino acids of unknown coordinates. This resulted a dataset of 1299 sequences (SSD1299) and 272,800 residues.

We separated randomly selected 298 sequences from this dataset as the test dataset (SSD_TS298), and the remaining 1,001 sequences are used as the training dataset (SSD_TR1001). SSD_TR1001 contains 211,048 residues which combines 69,333 helix (32.8%), 51,859 beta (24.5%) and 89,856 coil (42.5%) residues and SSD_TS298 comprised of 61,752 residues which combines 20,470 helix (33.1%), 16,052 beta (25.9%) and 25,230 coil (40.8%) residues. We determined the real or the actual annotation of secondary structural and surface area by the DSSP program [51].

2.3.6 Feature Set

We computed a comprehensive set of residue level features for predicting the secondary structures as well as the accessible surface area. The residue level information includes:

(*a*) One amino acid (AA) indicator

(*b*) Seven physical properties (PP)

(*c*) Twenty Position Specific Scoring Matrix (PSSM) values

(*d*) One monogram (MG) and twenty bigram (BG) values

(*e*) Two predicted disorder probabilities (short and long) (IUS and IUL)

(*f*) Three predicted secondary structure (SS) probabilities (helix, beta and coil)

(*g*) One terminal tag (T) to indicate five residues from N and C terminal as (-1.0, -0.8, -0.6, -0.4, -0.2) and (+0.2, +0.4, +0.6, +0.8, +1.0) while others as 0.0.

Feature AA is one numerical value for indices, ranges from 1 to 20, which correspond to the twenty different amino acids and PP are the seven physical properties of amino acids described in [174]. ASA is found to vary largely with different amino acids, and is correlated with properties, such as hydrophobicity and isoelectric point [204]. Thus, we used these features to predict real ASA values. PSSM values accommodate evolutionary information of protein residues and generated by executing 3 iterations of PSI-BLAST [264] against NCBI’s non-redundant database.

A feature extraction technique by [265] suggests that MG and BG values contain useful 3-dimentional evolutionary information of protein residues. Therefore, we calculated BG and MG values from PSSMs as described in [265] and used as features in our proposed ASA predictor. Disorder residues are often characterized to have large ASA values [162]. To incorporate this correlation into feature set, we computed disorder probabilities (IUS and IUL) using IUPRED [148] and incorporated into our feature set. Protein secondary structure is also closely coupled with ASA of the protein residues [266]. Thus, we developed our SVM model to generate predicted SS probabilities and used as features to predict ASA values.

We separately reported our predictor’s performance using two feature sets to depict the importance of 3-dimensional structural features in ASA prediction. Feature Plan #1 contains sequence based one-dimensional features (*a* – *c*, *g*). In addition, with the features in Feature-Plan #1, Feature-Plan #2 includes 3-dimensional feature (*d*) and predicted structural features (*e* – *f*).

Table 12 illustrates an overview of different feature plans used for secondary structure (SS) and ASA prediction along with the total counts of features in a given features plan. Finally, we included the neighboring residue information within the residue specific features set by applying a window size of 21 to incorporate the effect of residue contacts.

Table 12: List of features used in secondary structure and ASA prediction according to different feature plans.

Feature description (abbreviation)	Feature Count	Feature-Plan #1		Feature-Plan #2	
		SS	ASA	SS	ASA
Amino acid (AA)	1	√	√	√	√
Physical properties (PP)	7	√	√	√	√
Position specific scores matrix (PSSM)	20	√	√	√	√
Monogram (MG)	1	-	-	√	√
Bigram (BG)	20	-	-	√	√
Short and long probabilities (IUS/IUL)	2	-	-	√	√
Secondary structural probabilities (SS)	3	-	√	-	√
Terminal tag (T)	1	√	√	√	√
Total	55	29	32	52	55

'√' and '-' imply that the corresponding feature-set is included and excluded, respectively in the feature-plan.

2.3.7 Evaluation Measure

In our approach, we predicted ASA as real value, which was evaluated using Pearson's Correlation Coefficient (PCC) and Mean Absolute Error (MAE).

$$PCC = \frac{\sum_{i=1}^{N}(ASAr_i - \overline{ASAr})(ASAp_i - \overline{ASAp})}{\sqrt{\left[\sum_{i=1}^{N}(ASAr_i - \overline{ASAr})^2\right]\left[\sum_{i=1}^{N}(ASAp_i - \overline{ASAp})^2\right]}} \tag{6}$$

$$MAE = \frac{\sum_{i=1}^{N} |ASAr_i - ASAp_i|}{N} \tag{7}$$

Here, N is the total number of residue in the dataset, $ASAr$ and $ASAp$ are the real and predicted ASA. As the prediction obtains higher accuracy, the PCC value increases and the MAE value decreases.

3.3 REGAd^{3}p: Regularized Regression using Degree 3 Polynomial Kernel and Genetic Algorithm

REGAd^{3}p is a real value predictor framework that combines the exact regularized regression with optimization of weights predicted by genetic algorithm. The equation of basic exact regression [267] is:

$$\beta = (X^T X)^{-1} X^T Y \tag{8}$$

Here, X = input feature matrix having dimensions: number of residues ($N_{residue}$) × number of features ($N_{feature}$), X^T = transpose of the feature matrix X, Y = actual value of ASA (*ASAr*) and β= weights. **Equation (8)** analytically determines the best coefficients (weights) for the regression predicting the ASA value. After having the weights, **Equation (9)** is followed to predict the ASA. Here, $\hat{Y}$ = predicted values of ASA (*ASAp*).

$$\hat{Y} = X\beta \tag{9}$$

However, this basic equation is for linear regression model which can give a poor fit to the data. We extended the kernel of this regression method to 3rd degree polynomial function within the feature matrix using basis expansion by inserting two extra column vectors for each features which are the squares and cubes of the original feature values. This extension is expressed by the following equation, where p is the number of features given:

$$X = [1 \; x_1 \; x_2 \; x_3 \ldots \; x_p] \tag{10}$$

$$X^3 = [1 \; x_1 \; {x_1}^2 \; {x_1}^3 \; \ldots \; x_p \; {x_p}^2 \; {x_p}^3] \tag{11}$$

Here, X^3 is the extended feature matrix which is used in place of X is the basic Equation (8) and (9) to determine weights and calculate predicted ASA values, respectively. Extension of the kernel gave us the flexibility of model selection with higher order polynomial to select the best-fit model.

However, increasing the degree of polynomial can cause overfitting, resulted from highly fluctuating weights. An overfitted model towards training data can give poor performance on test dataset. To overcome this overfitting problem, we implemented regularization, which involves adding a penalty term to the error to shrink the value of weights. Therefore, the modified **Equation (12)** that includes regularization, where, λ = regularization parameter to control weight values and M_λ is the identity matrix of dimension (p+1) by (p+1) with the first diagonal element is assigned 0 to avoid affecting the bias term directly. We performed a search for the best value of λ within the range [-100.0 to 100.0] with an interval size equal to 2.0 and reported the best result (*see* **Section 3.5**) to compare the best result from regularized exact regression with and without GA optimization.

$$\beta = (X^T X + \lambda M_{\lambda_{(p+1).(p+1)}})^{-1} X^T \, Y \tag{12}$$

This search for λ gave us 100 sets of weights, which is then used as seeds for genetic algorithm. The parameter values of our genetic algorithm implementation are:

(i) Population size = 200

(ii) Number of generations = 2000

(iii) Chromosome length = number of weights ($N_{feature}$) × number of bits for each weight (19 in our implementation)

(iv) Elite rate = 10%

(v) Crossover rate = 80% and

(vi) Mutation rate = 10%

While generating initial population, 100 individuals are taken from the output of regularization and the rests were generated randomly. 10% ((200 × 0.1) or, 20) best performing weight sets are always forwarded towards next generation population from current one. To select the candidates for crossover, we implemented roulette wheel selecting algorithm to sample highly fitted individuals to be utilized for next generation population. We performed crossover on (200 × 0.8)/2 or, 80 pairs of chromosomes and filled up next 160 positions of next generation with 160 chromosomes resulted from crossover. Finally, we filled up last 20 positions of next generation population with 20 least fitted chromosomes of current generation. Then, we randomly selected a mutation candidate from these 20 chromosomes with repetition for (200 × 0.7) or, 140 times. In this process, a single chromosome can be selected multiples times (at most 70% of the total number of chromosomes in a population), and can get mutated at multiple positions.

Literature suggests that the best value of mutation rate is problem specific [268, 269]. The problem space of real valued accessible surface area is large and complicated. Moreover, we had highly diversified population of 200 chromosomes and the length of each chromosome is very high. We used 55 unique features per residue. Including the features of neighborhood residues within the window of size 21, we had (55 × 21) or, 1,155 features per residue. The coefficient of each feature is encoded by 19 bits within a chromosome. Therefore, a chromosome of GA is (1,155 × 19) or, 24,255 bits long. In a single mutation process, we flipped only one randomly chosen bit within the chromosome. Therefore, Mutation at multiple position was desirable to significantly change a long chromosome which aided in finding new improves solution within the large and complex search space of real values ASA. **Fig 16** illustrates an overview of the REGAd^{3}p real ASA prediction framework.

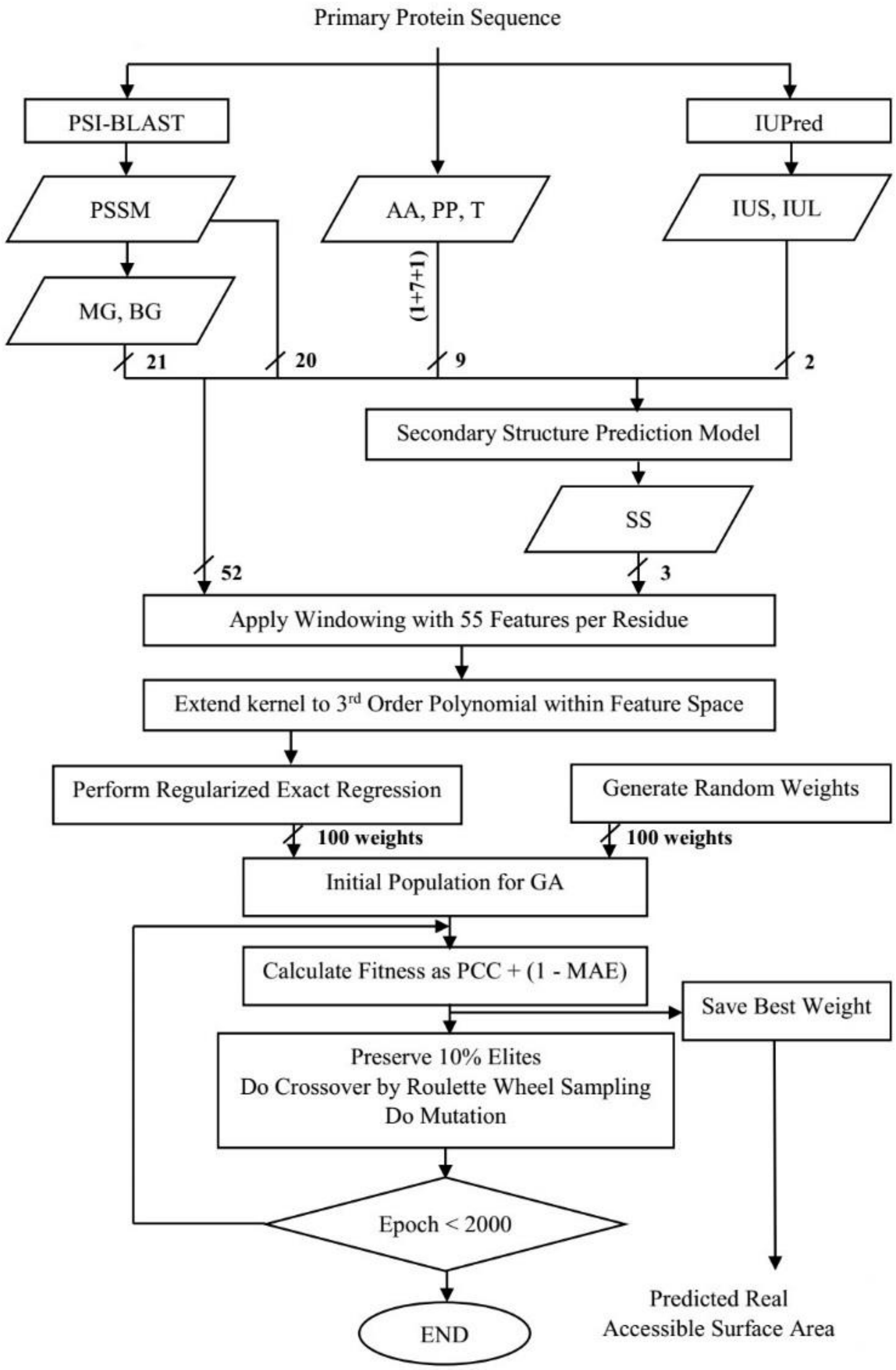

Primary Protein Sequence
PSI-BLAST
IUPred
PSSM
AA, PP, T
IUS, IUL
MG, BG
(1+7+1)
21
20
9
2
Secondary Structure Prediction Model
SS
52
3
Apply Windowing with 55 Features per Residue
Extend kernel to 3rd Order Polynomial within Feature Space
Perform Regularized Exact Regression
Generate Random Weights
100 weights
100 weights
Initial Population for GA
Calculate Fitness as PCC + (1 - MAE)
Save Best Weight
Preserve 10% Elites
Do Crossover by Roulette Wheel Sampling
Do Mutation
Epoch < 2000
END
Predicted Real
Accessible Surface Area

Fig 16. Overview of REGAd^{3}p real value accessible surface area prediction framework. It shows the feature aggregation, secondary structure prediction, and regularized exact regression and GA optimization. The features are represented by their abbreviations introduced in **Table 12.**

In contrast to the practice in the state-of-the-art predictors, we did not guide our predictor to achieve only low mean absolute error (MAE) or, high Persons correlation coefficient (PCC). Rather, we defined a multi-objective function (OBJ), combining PCC and MAE, and carried out the optimization of weights, β, for maximizing OBJ to achieve high performance both in terms of PCC and MAE. The equations for OBJ, PCC and MAE are as follows:

$$OBJ = PCC + (1 - MAE) \tag{13}$$

Furthermore, we integrated a post processing of predicted ASA values within GA to avoid negative values of the predicted ASA. To keep the ASA values practicable, the predicted negative values (because of the natural extension of the equation towards the non-admissible region) were replaced by zero.

3.3.1 Implementation and Availability

We implemented the REGAd^{3}p tool in C. The software is developed and tested on Linux platform. It is dependent on two external packages, namely PSI-BLAST[15] and NR database[16], which are publicly available. The software is available online[17] with a user manual.

3.4 Evaluation of REGAd^{3}p

Here we evaluate of our approach based on obtained results. First, we presented the result of the secondary structure prediction performed internally within the ASA prediction framework. Later, we present the results obtained through parameterization and development of the ASA predictor.

3.4.1. Results of Secondary Structure Prediction

At first, we predicted secondary structure of a residue of our dataset so that we can use the predicted secondary structure probabilities as features for ASA prediction. We explored four classifiers: (*i*) Logistic Regression (LogReg) using LIBLINEAR [270], (*ii*) Random Forest (RDF), (*iii*) Artificial Neural Network (ANN) using WEKA [271], and (*iv*) support vector machine (SVM) using LIBSVM [184]. Note that, the main objective of this work is to predict ASA and utilize the predicted ASA to improve 3DIGARS [239] energy function.

[15] PSI-BLAST link: ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/
[16] NR database link: ftp://ftp.ncbi.nlm.nih.gov/blast/db/
[17] REGAd^{3}p link: http://cs.uno.edu/~tamjid/Software/REGAd3p/REGAd3p.tar.gz

We collected the eight state secondary structure annotation from DSSP program [51] which includes α-helix (H), 3-helix or, 310-helix (G), 5-helix or, π-helix (I), residue in isolated β-bridge (B), residue in extended beta or, β-ladder (E), hydrogen bonded turn (T), bend (S) and random coil or, loop (blank) for all residues in SSD1299 dataset. We converted this eight state annotation into three state annotation [175] by coding H, G and I as H (helix), B and E as E (beta) and T, S and blank as C (coil). **Table 13** gives an illustration of the performance of the four classifiers for both the feature plans (***see*** **Table 12**) in terms of accuracy (total number of correctly predicted residues) of three class classification. All the classifiers were trained on SSD_TR1001 dataset for three class (helix, beta and coil) classification and tested on SSD_TS298 dataset. The superior performance of SVM model on both the feature plans motivated us to select it as our predictor of secondary structure. We used *radial basis function* (RBF) as the kernel for the applied SVM. We used LIBSVM [184] package for building SVM model and used the default parameter values provided within the package. The default values of misclassification cost of SVM and gamma parameter of RBF were one and $(1/N_{feature})$, respectively.

Table 13. Performance of secondary structure prediction by four classifiers on SSD_TS298 dataset.

Classifier	LogReg (%)	RF (%)	ANN (%)	SVM (%)
Feature-Plan # 1	73.46	72.3	72.8	**75.31**
Feature-Plan # 2	73.51	72.7	72.53	**74.86**

Bold: indicates the obtained best values.

3.4.2. Result of Accessible Surface Area Prediction

In section, we report the results we obtained during the design process of REGAd^{3}p.

3.4.2.1. Feature Plan Selection

We evaluated the performance of REGAd^{3}p on both the training (SSD_TR1001) and test (SSD_TS298) dataset. **Table 14** presents the performance of REGAd^{3}p in predicting absolute ASA values in terms of PCC and MAE for feature plan # 1 and plan # 2. As a result of inclusion of three dimensional features (MG and BGs) and structural features (short, long disorder probabilities), PCC value is increased (**0.28%**) as well as MAE is decreased (**0.16%**). This result validates the correlation of residue exposure with protein's flexibility (disorder) and usefulness of these features in ASA prediction. In addition, it also motivates us to do further experiments only on feature plan # 2.

Table 14. Prediction quality of ASA for different feature plans with 1st order polynomial as kernel.

Dataset	SSD_TR1001		SSD_TS298	
Features	**MAE**	**PCC**	**MAE**	**PCC**
Plan # 1, non-optimized	27.27	0.655	25.44	0.711
Plan # 1, optimized	26.53	0.661	24.57	0.717
Plan # 2, non-optimized	27.05	0.665	24.45	0.711
Plan # 2, optimized	**26.31**	**0.670**	**24.53**	**0.719**

Bold: indicates the obtained best values.

3.4.2.2. *Kernel Selection*

We extended the kernel from degree 1 polynomial (linear) up to 4 to determine the optimal polynomial function to be utilized, so that the model best fits the ASA values with feature plan # 2. **Table 15** summarizes the results. PCC is increased by **2.03%** and MAE is decreased by **2.3%** as we extended the kernel from 1st order polynomial to 3rd order polynomial function. **Table 15** also shows **4.63%** and **5.13%** fall in performance in terms of PCC and MAE, respectively, when the kernel is extended beyond 3rd order polynomial. This behavior indicates that the predictor's performance can suffer from high dimensionality as a result of making the model too complex and motivated us to select 3rd order polynomial as the optimal kernel function.

Table 15. Prediction accuracy of ASA due to the extension of kernel function from linear to higher order polynomial (Feature-Plan # 2).

Dataset	SSD_TR1001		SSD_TS298	
Polynomial kernel	**MAE**	**PCC**	**MAE**	**PCC**
Degree 2, non-optimized	26.24	0.683	25.05	0.717
Degree 2, optimized	25.86	0.686	24.53	0.723
Degree 3, non-optimized	25.29	0.699	25.54	0.727
Degree 3, optimized	**25.19**	**0.702**	**23.97**	**0.734**
Degree 4, non-optimized	26.61	0.676	25.98	0.685
Degree 4, optimized	26.21	0.679	25.21	0.700

Bold: indicates the obtained best values.

Table 14 and **Table 15** compare the performances of REGAd^{3}p for both with and without the optimization and signify the importance of weight optimization for improving the performance. For the best model (with Feature-Plan # 2 and 3rd order polynomial kernel), the improvement due to optimization over un-optimized model was **0.96%** for PCC and **6.14%** for MAE. Genetic algorithm successfully enhanced the performance of classical regression method to make it competitive with several complex pattern recognition algorithms, like artificial neural network and support vector regression that we had tested extensively.

3.4.2.3. Result of Final Model

To verify the robustness of our best model further, we carried out 10-fold cross validation on the training dataset, SSD_TR1001. In statistical prediction, the following three cross-validation methods are often used to examine a predictor's effectiveness in practical application: independent dataset test, subsampling test, and jackknife test. However, of the three test methods, the jackknife test is deemed the least arbitrary that can always yield a unique result for a given benchmark dataset as elaborated in [272]. Accordingly, the jackknife test has been widely recognized and increasingly used by investigators to examine the quality of various predictors (e.g., [273], [274], [275], [276], [277], [278], [279], [280]). However, to reduce the computational time, we applied more commonly used approach, the 10-fold cross-validation. The result of 10-fold cross validation test is **0.69** and **25.43** in terms of PCC and MAE, respectively, which is consistent with the result of independent test (indicated with bold in **Table 15**).

Finally, the overall performance comparison among the models (for different feature plans as well as kernels) is shown in **Fig 17** on SSD_TS298 dataset with optimization, where the best model with plan # 2 and kernel # 3, is indicated.

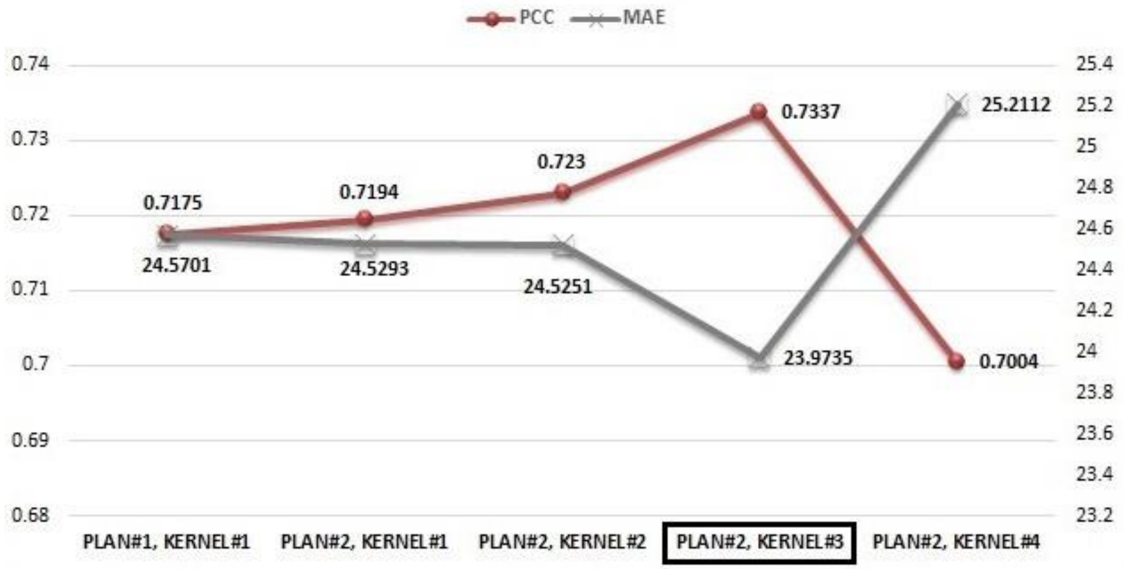


Fig 17. Overall comparison of performance for different feature plans and kernel functions on SSD_TS298 dataset with GA optimization. The *x*-axis and *y*-axis shows the model description and performance measure scores (PCC and MAE), respectively. The best model is marked.

3.4.3. Comparison of ASA Prediction with Existing Methods

We compared our regularized regression technique with optimization against top performing artificial neural network based predictor, SPINEX [175]. To avoid the inconsistencies raised in comparison due to different datasets, normalizing factors and evaluation measures, we ran SPINE-X on SSD1299 dataset and collected the absolute ASA for a fair comparison. **Fig 18** summarizes the result.

It shows that except in the case of MAE for SSD_TS298 dataset, REGAd^{3}p outperformed SPINE-X [175] in absolute ASA prediction. REGAd^{3}p gave **8.2%** and **1.45%** improved PCC score than SPINE-X [175] for SSD_TR1001 and SSD_TS298 datasets, respectively. Further, we evaluated the statistical significance of these improvements by t-test with R package [281] which shows that both of the improvements are significant. Moreover, we executed support vector regression (SVR) from LIBLINERA package [270] on SSD1299 dataset and found the better performances of REGAd^{3}p. SVR resulted a PCC value of 0.51 when trained on SSD_TR1001 dataset and tested on SSD_TS298 dataset for Feature-Plan # 2. To compare, REGAd^{3}p gave PCC value of **0.73** in case of the same dataset and feature plan.

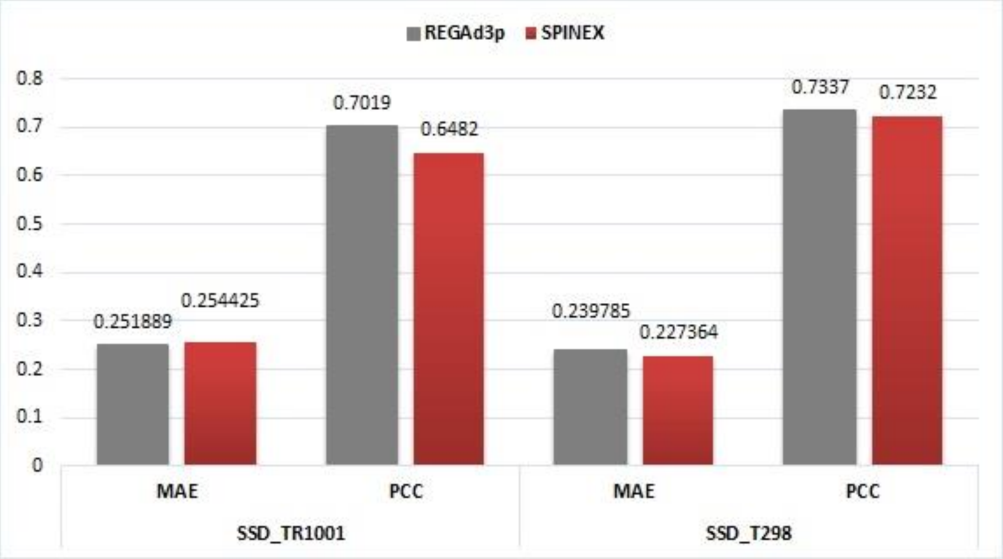


Fig 18. Comparison between REGAd^{3}p and SPINE-X [175] in absolute ASA prediction on SSD_TR1001 and SSD_TS298 dataset. The x-axis and y-axis represents the dataset and performance measure scores (MAE and PCC), respectively.

3.4.4. Case Studies with Individual Proteins

We selected two protein chains, (*i*) 1C9OA and (*ii*) 1DK8A of length 66 and 147, respectively, to investigate the sequence wise performance of REGAd^{3}p versus SPINE-X [175], in predicting absolute ASA values.

We plotted the actual annotation of ASA calculated from DSSP with residue wise predicted ASA values from REGAd^{3}p and SPINE-X [175] in **Fig 19**.

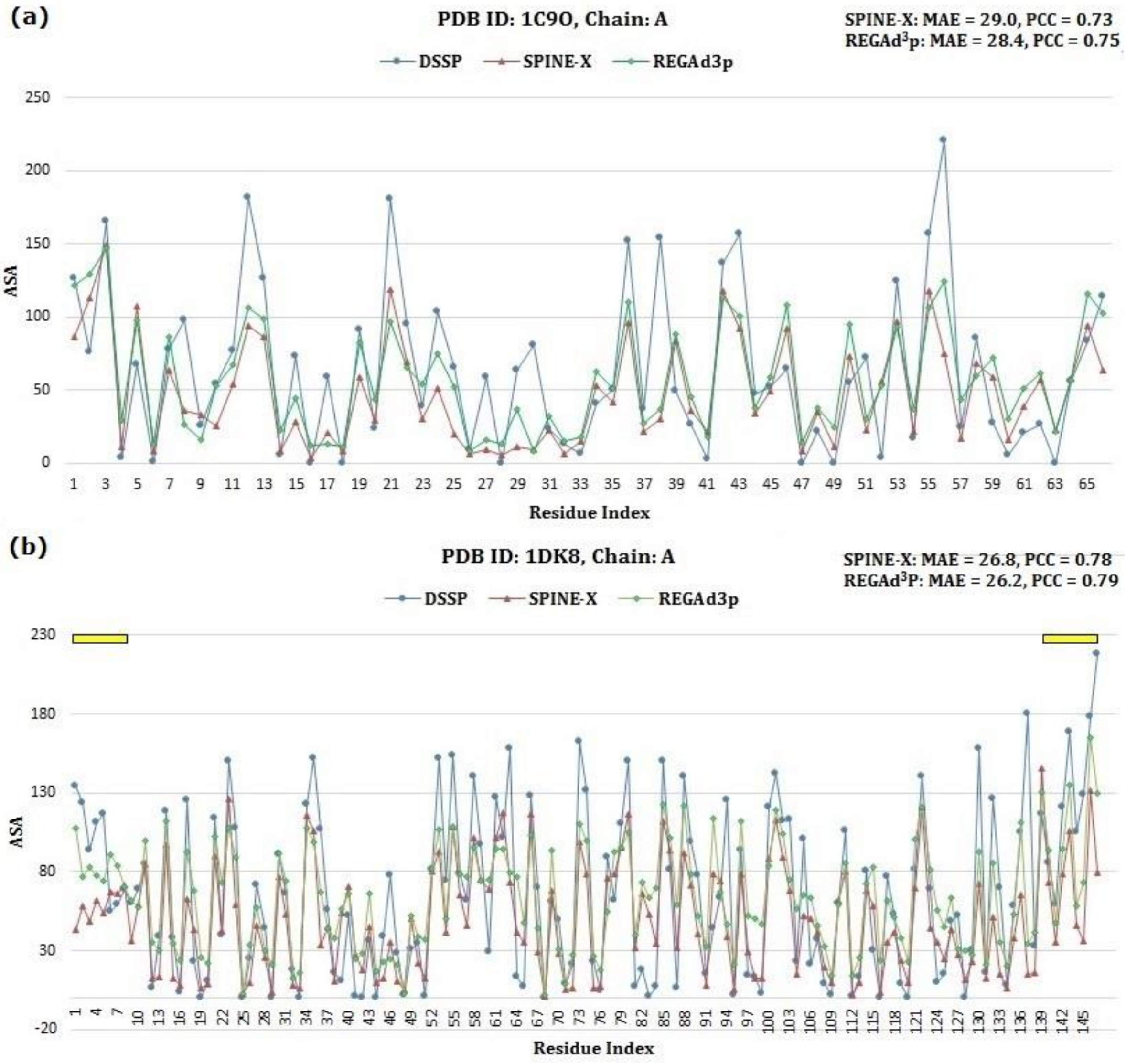


Fig 19. ASA prediction comparison for individual proteins. Plot of ASA for each residues of protein (i) PDB ID: 1C9O, Chain: A and (ii) PDB ID: 1DK8, Chain: A, given by DSSP (blue line with circle marker), SPINE-X [175] (red line with triangle marker) and REGAd3p (green line with diamond marker). For each plot, MAE and PCC scores between predicted ASA given by SPINE-X and REGAd3p with actual ASA from DSSP are shown on the top right corner. In PDB ID: 1DK8, the yellow bars represent two disorder regions [282] at the terminals. The x-axis and y-axis shows the residue index and ASA values, respectively.

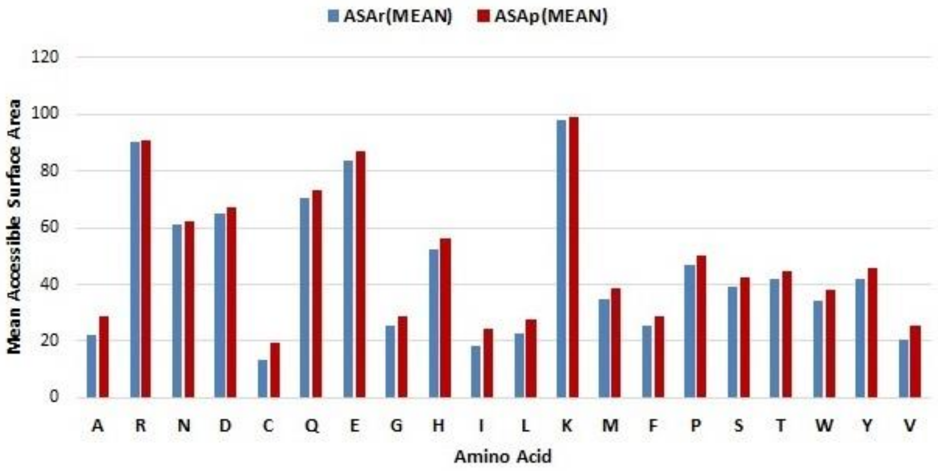
It is clear from the plots that both predictors lack in accuracy when exposure is high. To be specific, for the residue indices: 11 – 13, 21, 36, 38, 43 – 44, 55 – 56 of protein chain 1C9OA, both predictors under predict. To compare, REGAd^{3}p could result better prediction than SPINE-X for the residue index: 11 – 13, 15, 19, 23 – 3, 36 and 55 – 56 of protein chain 1C9OA.

For protein chain 1DK8A, we marked the predicted disorder region (residue indices 1 – 8 and 139 – 142) at the terminals collected from DisPredict [282]. Disordered regions are often characterized by dynamic conformation and high residue accessible surface areas. It is evident from the plots that REGAd^{3}p could better predict the ASA values at disordered regions than SPINE-X. We further summarized the residue-wise performance by sequence-wise MAE and PCC scores given by both the methods, reported in the top right corner of the plots. For both the proteins as well as measures (MAE and PCC), REGAd^{3}p outperformed SPINE-X. In the case of 1C9OA, REGAd^{3}p gave **2.1%** lower MAE and **2.7%** higher PCC than SPINE-X. At the same time, for 1DK8A, REGAd^{3}p outperformed SPINE-X by **1.5%** decrease in MAE and **1.3%** increase in PCC score.

3.5 ASA Prediction Analysis

We performed extensive analysis on the predicted ASA by REGAd^{3}p and actual ASA obtained from DSSP [51] to assess the quality of our proposed real value prediction framework. We used the prediction output on SSD_TS298 for the following analysis.

3.5.1 Amino Acid Specific Analysis

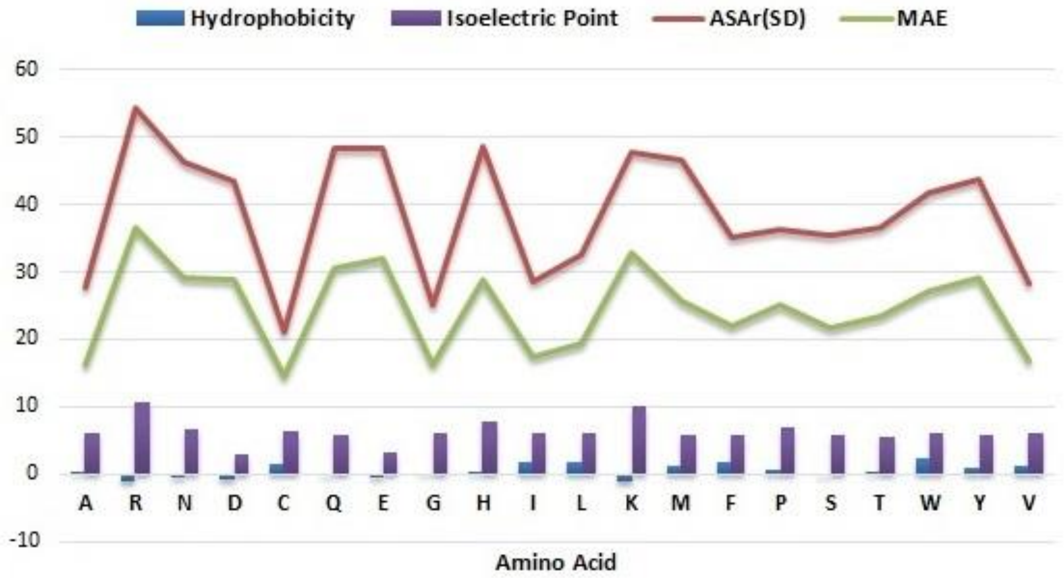


PCC(ASAr(MEAN), ASAp(MEAN)) = 0.99

Fig 20. Actual and predicted ASA of different amino acid residues of SSD_TS298 dataset. Amino acid specific comparison between mean actual ASA (blue bar) and mean predicted ASA (red bar) values. The x-axis and y-axis show the amino acid and ASA values, respectively.

This analysis is executed in terms of total number of residues within dataset (COUNT) per amino acid, maximum value of actual ASA within dataset (ASAr(MAX)), mean actual ASA (ASAr (MEAN)), standard deviation of actual ASA (ASAr(SD)), mean predicted ASA (ASAp(MEAN)) and MAE resulted from prediction. **Fig 20** illustrates that how the predicted and actual ASA values for each amino acid is highly correlated, with a PCC value equal to 0.99, without no significant over prediction or, under prediction.

Fig 21 represents the correlation between the mean absolute errors in prediction with the inherent variability (computed by standard deviation) of ASA values within the dataset. It shows a high correlation value of 0.97 which indicates that the internal fluctuation of ASA within the dataset determines the prediction quality. To this end, identification of flexible (disorder) residues with high net charge and low hydrophobicity is challenging, which are also characterized by large ASA values [204].



PCC（ASAr（SD）， MAE） = 0.97

Fig 21. Correlation of predicted ASA and physical properties of different amino acid residues. Correlation between MAE （green line） with actual standard deviation （red line） of ASA value within dataset, hydrophobicity （blue bar） and isoelectric point （purple bar） for each amino acid type.

It is also evident from **Fig 21**, as the disorder promoting amino acids (R, Q, E, K) [283] are likely to have highly variable ASA, therefore contributing in high MAE values. However, **Fig 20** shows that the average predicted ASA given by REGAd^{3}p for these amino acids are very close to the average actual ASA. Furthermore, **Fig 21** shows the prediction error for each amino acid with its respective hydrophobicity index and isoelectric point (charge index) [174]. The high errors in prediction are resulted for Arg (R), Lys (K), Glu (E), which have high charge and low hydrophobicity. These residues usually reside in the exterior of proteins and incorporate flexible ASA values with high deviation. **Fig 20** also shows that the low errors are found for Cys (C), Gly (G), Ile (I), Val (V), Ala (A), which are normally buried or, partially buried inside the core of the protein with higher hydrophobicity values.

3.5.2 Secondary Structure Specific Analysis

Fig 22 presents the distribution of actual and predicted ASA values for three different secondary structure types of residues (coil, helix and beta) with their corresponding MAE. It is evident from the distributions that for each type of residues, prediction of ASA is relatively easier for the unexposed residues, as the distributions overlap with low ASA values.

Coil residues have highest exposure and beta residues have lowest exposure (mean actual ASA for coil, helix and beta residues are 27.49, 45.87 and 56.61, respectively) which indicates that beta residues are mostly structured whereas coil residues are mostly flexible. This is also clear from the MAE value which decreases from coil to helix to beta residues.

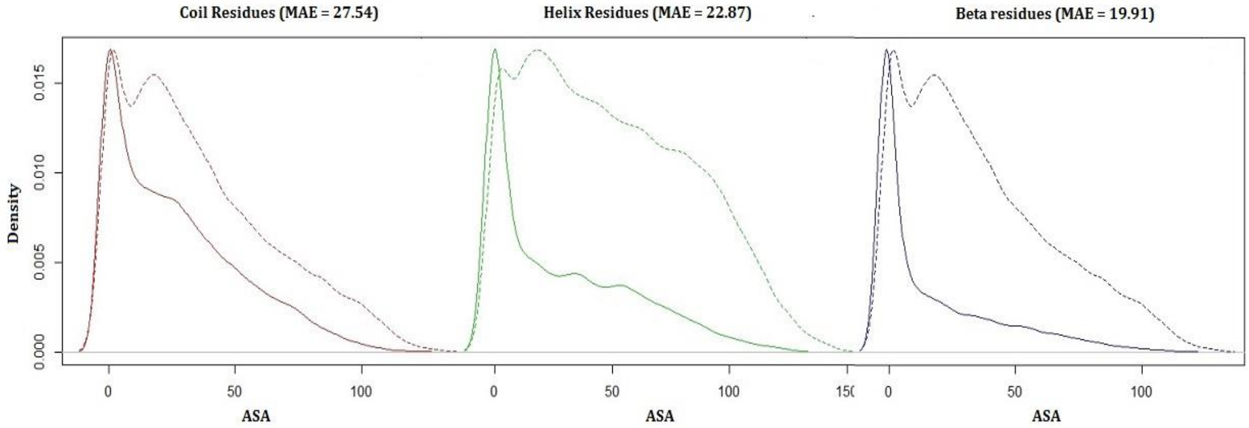


Fig 22. ASA prediction performance of REGAd^{3}p on different types of secondary structure residues of SSD_TS298 dataset. The Distribution of actual (solid line) and predicted (dotted line) ASA values for coil (red), helix (green) and beta (blue) residue are plotted. The x-axis and y-axis shows the ASA values and corresponding density. MAE is reported in the title of each type of residue’s curve.

3.5.3 ASA Range Specific Analysis

Fig 23 shows the mean absolute error in prediction resulted by REGAd^{3}p for different range, $[x_1 – x_2)$ where $x_1 \leq ASAr < x_2$ of actual ASA values. It also represents the count of total residues (COUNT), coil residues (Coil), helix residues (Helix) and beta residues (Beta) within a range. The graphical overview illustrates that for a wide range of actual ASA values, [0 – 105), REGAd^{3}p could predict with consistently low MAE which incorporates 87% of the total dataset. For the rest of the residues (13%), the MAE increases with the range which is justified, since these residues are highly exposed and often involved in protein-protein interactions which results in dynamic conformations.

Fig 23 also shows that only within the ASA range [0 – 15), there are more beta residues than other (helix or, coil) residues which are relatively easy to predict as beta residues are mostly structured. For rest of the range, there are more coil residues. However, REGAd3p could still predict the flexible coil residues for a wide range, [15 – 105), resulting consistent MAE with the structured beta residues. Thus, the prediction output of our REGAd^{3}p can be utilized as useful feature for flexible region prediction in future.

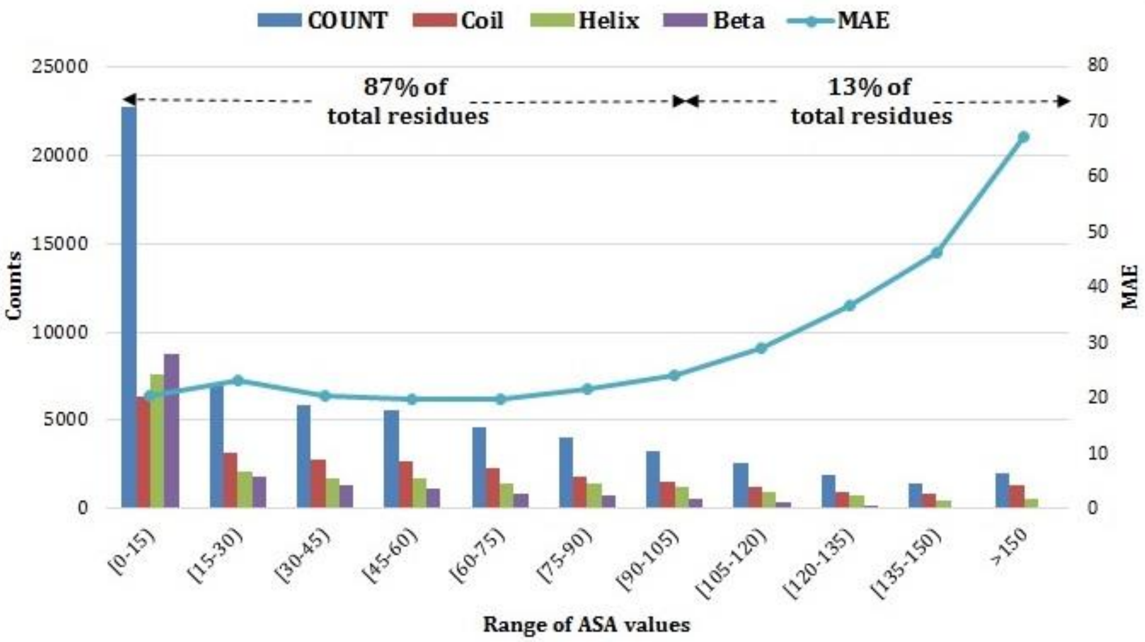


Fig 23. Variation of error in ASA prediction depending on the range of actual ASA values (dataset: SSD_TS298). The x-axis shows the range of actual ASA values in the form [x_1 – x_2) = $x_1 \le$ ASAr < x_2. The primary y-axis (left side) shows the count of all types of residues (blue bar), coil residues (red bar), helix residues (green bar) and beta residues (purple bar) within a range and secondary y-axis (right side) shows MAE (light blue line), respectively.

3.6 Application of Predicted ASA in the Energy Function, 3DIGARS2.0

An efficient energy function is the one which can identify more native from their decoy sets. Towards this goal, we have developed a new energy function (3DIGARS2.0) which is an improvement over the previous version (3DIGARS[18]) [239] for protein structure prediction. 3DIGARS2.0 is different from 3DIGARS in terms of using sequence based solvent-accessibility information. We obtained the sequence-based solvent-

[18] 3DIGARS: http://cs.uno.edu/~tamjid/Software.html

accessibility energy by modelling the error between actual and predicted accessible surface area (ASA). 3DIGARS2.0 is a linear combination of energy from 3DIGARS and energy from sequence-specific solvent-accessibility which is further optimized using Genetic Algorithm (GA).

3DIGARS2.0 outperforms DFIRE [233], RWplus [284], dDFIRE [285], GOAP [286] and 3DIGARS [239] energy functions based on the most challenging Rosetta and I-Tasser decoy sets with an improvement on weighted average of 80.78%, 73.77%, 141.24%, 16.52%, and 32.32% respectively based on count of correct identification of native from decoy sets (see Table 6).

3.6.1 Experimental Materials

In this section, we describe the experimental materials related to the development of the improved energy function.

3.6.1.1 *The 3DIGARS Potential*

3-Dimensional Ideal Gas Reference State (3DIGARS) [239] potential is based on hydrophobic-hydrophilic model (HP model) which computes three different energy score interaction tables, specifically, *i*) hydrophobic-hydrophilic (HP), *ii*) hydrophobic-hydrophobic (HH); and *iii*) hydrophilic-hydrophilic (PP). The interaction distance bin size is kept Δr = 0.5 Å each, with a cutoff distance of 15 Å, where *r* represents each distant bin with values ranging from 0.5 Å to 15 Å. The value of Δr_{cut} = 0.5 Å as all bin size are same.

The exception consider in 3DIGARS is based on the fact that – amino acid, based on their types are not distributed equally over the 3D structure of a protein to consider them in the same scale on an average by a single dimensional parameter, which can rather be segregated into at least 3 different categories based on the regular distributions within native conformations. 3DIGARS implements Genetic Algorithm (GA) to obtain the best fitted value of six parameters: 3D alpha (three different values of alpha represents three different interaction mentioned above and generate three different energy score tables) and 3D beta (three different values of beta represents the contributions of each of the group along with the z-score).

3.6.1.2 *Decoy Sets*

The performance of 3DIGARS2.0 has been compared with other state-of-art energy functions based on three most challenging decoy datasets Moulder, Rosetta and I-Tasser. Modular[19] dataset consists of 20 native proteins with 300 comparative decoy models generated using homologous template for each protein. Rosetta[20] dataset includes 58 proteins generated by Baker Lab. Each of the proteins contains 20 random

[19] Modular dataset: http://salilab.org/john_decoys.html

[20] Rosetta dataset: http://www.bakerlab.org/

models and 100 lowest scoring models from 10,000 decoys generated by ROSETTA *de novo* structure prediction [287] followed by all-atom refinement. I-Tasser[21] decoy set-II consist of 56 proteins. Each of these proteins contains 300 to 500 decoys. These decoys were generated first by using Monte Carlo Simulations and then refined by GROMACS4.0[22] MD simulation.

3.6.2 The 3DIGARS2.0 Potential

3DIGARS2.0 combines the sequence-specific solvent-accessibility energy, E^{ASA}, computed from the error between real and predicted accessible surface area (ASA) linearly with 3DIGARS. E^{ASA} is based on probability $P(\Delta ASA_i \mid AA_i)$ computed from the error modelling of our predicted solvent-accessibility $(\Delta ASA_i = ASA_i^{\text{Real}} - ASA_i^{\text{Pred}})$ for a given amino acid type AA_i . Here, for each residue *i* within each protein, ASA_i^{Real} (real accessible surface area) is computed using DSSP and ASA_i^{Pred} is computed using REGAd^{3}p methodology.

Table 16 summarizes the performance achieved by our SVM model on secondary structure prediction in terms of accuracy and REGAd^{3}p predictor on absolute accessible surface area prediction in terms of MAE as well as PCC. The best result was found for I-Tasser dataset in ASA prediction with PCC value equal to **0.76**.

Table 16. Performance of secondary structure prediction and ASA prediction by REGAd3p for Moulder, Rosetta and I-Tasser datasets.

Dataset	SS prediction	ASA prediction	
	Accuracy (%)	MAE	PCC
Moulder	66.56	23.07	0.68
Rosetta	72.04	24.93	0.71
I-Tasser	74.23	24.73	0.76

Now, keeping the prediction accuracy of REGAd^{3}p in mind as in **Table 16**, we wanted to model the error pattern in a useful way to aid in enhancing the accuracy of the energy function. With a view to this, we computed the error in ASA prediction for each of the residues of 1299 proteins and obtained the frequency distribution (FDT) of the error between real and predicted accessible surface area. While building frequency distribution we first calculated the max error ΔASA from the dataset of 1299 protein which was found to be 240. The error range from 0 to 195 was then divided by bin width of 5 to obtain 39 bins of equal size.

[21] I-Tasser dataset: http://zhanglab.ccmb.med.umich.edu/decoys/
[22] GROMACS4.0: http://www.gromacs.org/Downloads

Remaining of the error ranging from 195 to 240 is considered to fall in the last bin or the 40th bin of the frequency distribution. Thus, the 40th bin has the width of 45 (240-195 = 45). In our implementation of having each bin size of equal width of 5, we normalized the values of last bin by dividing each cell count by 9 (45/9 = 5). Thus, the final frequency distribution table consist of 20 rows (for 20 different types of amino acid) and 40 bins of equal size of 5. For each residue, frequency distribution table is updated as **Equation (14)**:

$$FDT\left(AA_i\,,bin_j\right) = FDT\left(AA_i\,,bin_j\right) + 1.0 \qquad (14)$$

Here, *AA_i* is the *ith* amino acid and *bin_j* is the *jth* bin. Index *i* ranges from 1 to 20 indicating twenty different amino acids and *j* ranges from 1 to 40 indicating bins. *bin_j* is defined by **Equation (15)**:

$$bin_j = abs(\Delta ASA_i = ASAr_i - ASAp_i)/bin\_size \qquad (15)$$

Here, *bin_size* = 5. Once the frequency table is obtained, cell whose frequency count is zero are replaced with a small value of 10^{-6}. After the frequency is computed, probability table is obtained by **Equation (16)**:

$$P(AA_i\,,bin_j) = FDT(AA_i\,,bin_j)/Tot\_Freq \qquad (16)$$

Here, Tot_Freq is the sum of the count of each amino acid type in frequency table. Finally, the energy score library for sequence specific solvent accessibility is obtain by **Equation (17)**:

$$E(AA_i\,,bin_j) = -\ln(Bin\_Count \times P(AA_i\,,bin_j)) \qquad (17)$$

Energy associate with each native protein as well as decoy protein is given by **Equation (18)**:

$$E^{ASA} = \sum_{K=1}^{N} E_k(AA_i\,,bin_j) \qquad (18)$$

The combined energy $E^{3DIGARS2.0}$ for each of the proteins including native as well as decoy sets are calculated using **Equation (19)**:

$$E^{3DIGARS2.0} = E^{3DIGARS} + (w \times E^{ASA}) \qquad (19)$$

3DIGARS2.0 potential combines 3DIGARS energy with the sequence-specific solvent accessible energy for each of the protein with weight w_1. The optimal value of weight *w*, ranging from 0 to 2, is obtained from using Genetic Algorithm (GA). The Genetic Algorithm (GA) parameters were of population size 300, elite-rate 5%, crossover-rate 90% and mutation-rate 50%. The stopping criteria to stop the optimization was set to maximum number of generations 2000. **Fig 24** shows the performance of GA where the value remains stable after around 3rd generation.

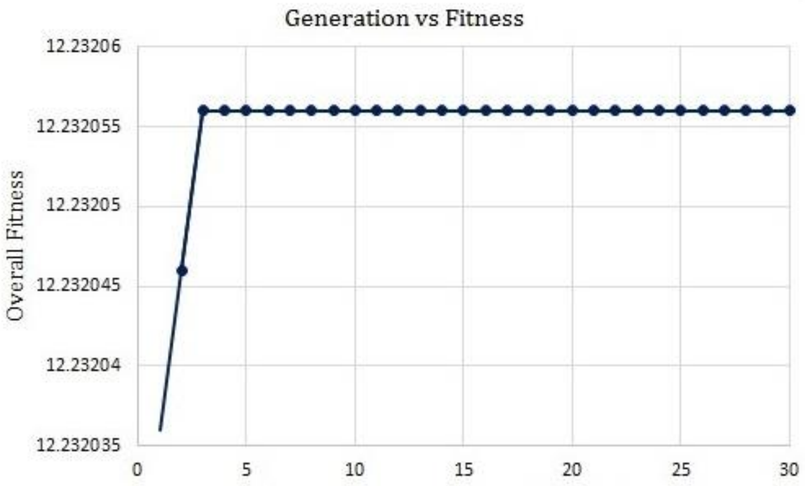


Fig 24. GA optimization result: generations versus fitness graph. Fitness increases sharply and remains constant over number of iterations indicating stable outcome.

Fig 25 shows the complete workflow of computing 3DIGARS2.0.

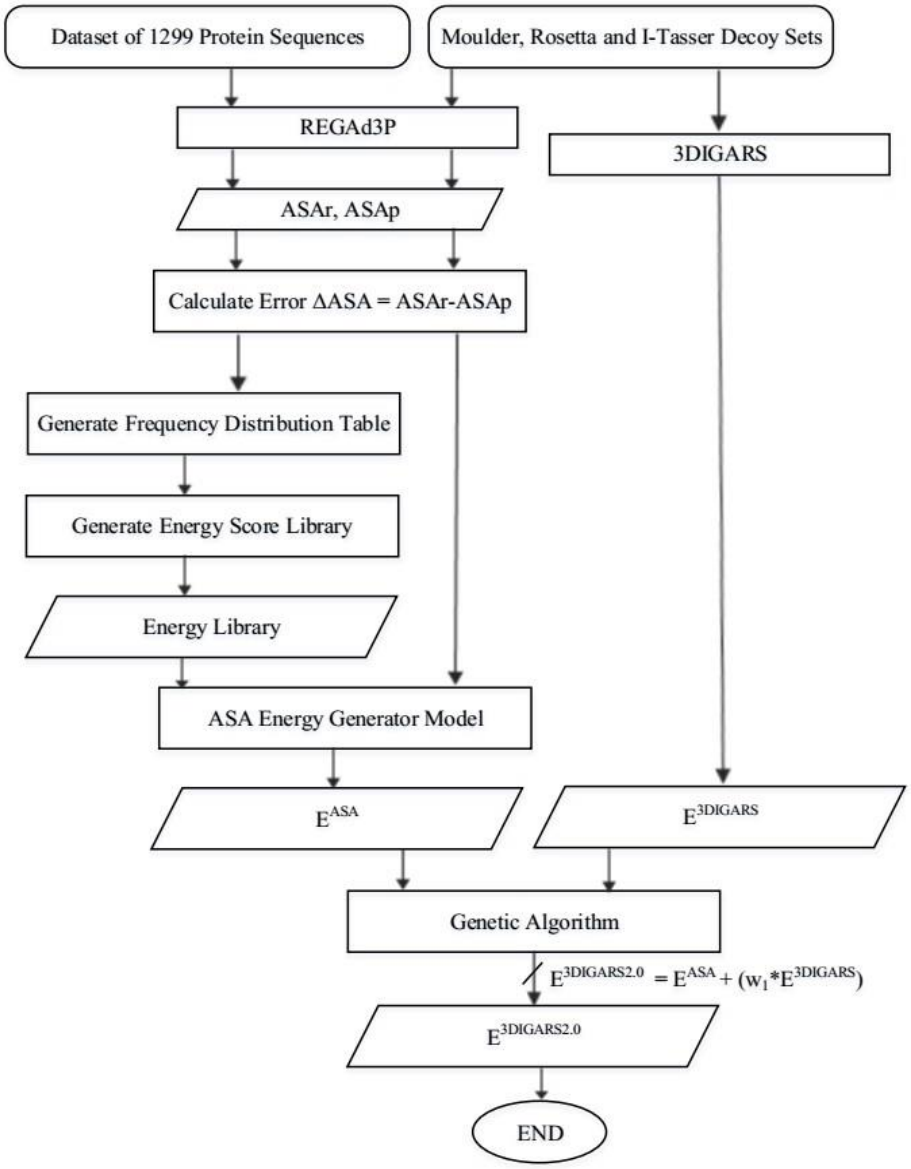

Dataset of 1299 Protein Sequences
Moulder, Rosetta and I-Tasser Decoy Sets
REGAd3P
3DIGARS
ASAr, ASAp
Calculate Error ΔASA = ASAr-ASAp
Generate Frequency Distribution Table
Generate Energy Score Library
Energy Library
ASA Energy Generator Model
E^ASA
E^3DIGARS
Genetic Algorithm
E^3DIGARS2.0 = E^ASA + (w1*E^3DIGARS)
E^3DIGARS2.0
END

Fig 25. Steps of computing 3DIGARS2.0 potential. The abbreviations used are explained in Section3.6.2.

3.6.3 Performance of 3DIGARS2.0

We compare the performance of 3DIGARS2.0 with the state-of-art-approaches DFIRE, RWplus, dDFIRE and GOAP using the most challenging three different decoy datasets, Moulder, Rosetta and I-Tasser. 3DIGARS2.0 is found to outperform all the state-of-art approaches including the previous version of 3DIGARS with high number of correctly identified native proteins from their decoy datasets. For example, based on Rosetta and Tasser decoy-sets 3DIGARS2.0 improved on an average over DFIRE, RWplus, dDFIRE, GOAP, 3DIGARS are **79.64**%, **72.5**%, **162.48**%, 17.77%, and 31.86% respectively. In **Table 17** the results for DFIRE, RWplus, dDFIRE and GOAP are obtained from [286] and 3DIGARS from [239].

Table 17. Comparison between DFIRE, RWplus, dDFIRE, GOAP, 3DIGARS and 3DIGARS2.0 based on correct selection of native from their decoy-set and z-score.

Decoy Sets (No. of targets)	DFIRE	RWplus	dDFIRE	GOAP	3DIGARS	3DIGARS 2.0
Moulder (20)	**19** (-2.97)	**19** (-2.84)	18 (-2.74)	**19** **(-3.58)**	**19** (-2.998)	**19** (-2.6728)
Rosetta (58)	20 (-1.82)	20 (-1.47)	12 (-0.83)	45 **(-3.70)**	31 (-2.023)	**49** (-2.9871)
Tasser (56)	49 (-4.02)	**56** **(-5.77)**	48 (-5.03)	45 (-5.36)	53 (-4.036)	**56** (-4.2964)

Bold indicates best score. Values within the parenthesis are average z-scores of the native structure.

3.7 Summary and Conclusions

In this chapter, we described a new framework, namely REGAd^{3}p. The workflow includes a canonical exact regression technique, optimized further by genetic algorithm. The superior performance achieved by our proposed framework proves that integration of optimization by genetic algorithm can successfully enhance the capability of classical pattern recognition methods. The framework is generic and applicable for any real-value prediction application with appropriate tuning of the parameters. However, we have applied it for the prediction of absolute accessible surface area (ASA) of residues from protein sequence alone.

The accessible surface area (ASA) is often used as an important measure related to proteomic studies for describing the biophysical properties of a protein. We introduced a comprehensive feature set which could better characterize absolute ASA as we achieved better accuracy (PCC: 0.73) in the case of independent test compared to existing independent test results (PCC equal to 0.49 [256], 0.61 [262], 0.66 [263]). However, these results are subject to different datasets and different normalizing factors which make the comparisons often inconsistent. Therefore, under this work, we introduced a new benchmark dataset, SSD1299 and compared the performance of our methodology with the existing state-of-the-art predictor, SPINE-X [175] by running it on SSD1299 dataset. Our test results for multiple datasets (SSD_TR1001, SSD_TS298, Moulder, Rosetta, I-Tasser) further prove our method is robust. As demonstrated in a series of recent publications (e.g., [272], [273], [274], [275], [276], [279]) in developing new prediction methods, user-friendly and publicly accessible web-servers significantly enhance their impacts [280]. We will make efforts in our future work to provide a web-server for the prediction method presented in this chapter.

For this work, we avoided the common practice of predicting normalized ASA [178] and then de-normalizing since the normalized ASA varies depending on normalizing factor. Current research is still determining a suitable reference value for normalizing ASA [288] which indicates that inaccurate normalizing factor can lead to misinterpretation of the absolute accessible surface area of residue within a protein conformation.

We followed the normalized ASA calculation adopted in [178] and computed the MAE for normalized ASA prediction with respect to the SSD1299 dataset by our proposed framework. We found that the correlation between the error in prediction in case of with and without normalization for twenty different amino acids is poor with PCC value equal to 0.41 which also proves the inconsistency between absolute and normalized ASA values. Therefore, we predicted absolute ASA values and justified the quality of our prediction through exhaustive analyses of different amino acids along with their physical properties, residues with different type of secondary structure and multiple range of ASA values. Through these analyses, we could show that flexible and higher ASA values are harder to predict. However, for a wide range of ASA values, [0 – 105), REGAd^{3}p can predict with consistently lower error and higher correlation which suggests that our methodology can be useful and consistent in measuring parameters in protein's dynamic or, flexible structure prediction and function identification.

To establish a concrete instance of the claim, we extended 3DIGARS energy functions by optimally combining the ASA error model based energy generated by REGAd^{3}p, namely 3DIGARS2.0, which significantly outperformed all the state-of-the-art energy functions based on the most challenging decoy datasets. We have also extended the application of sequence based predicted ASA towards developing of a sequence base energy score, which is described in **Chapter 4**. Moreover, in a separate work conducted by us in balanced secondary structure prediction [196], the predicted ASA by REGAd^{3}p served as one of the major feature.

Energy function is one of the key component of *ab initio* protein structure prediction, which is an important method to predict proteins' 3D structure from the given amino acid sequence only. In cases, where homologous proteins are absent, the *ab initio* protein structure prediction approach becomes essential. Therefore, the proposed predictor of real value ASA (i.e., REGAd^{3}p) and energy function (i.e., 3DIGARS2.0) can be very useful for emerging research of proteomics and related fields.

Chapter 4

PSEE: Position Specific Estimated Energy

— An Energy Score to Characterize the Stability of Protein and its Application in Disorder Prediction

Protein folding is the process by which a protein chain acquires its 3-dimensional (3D) structure. It is the physical process by which a polypeptide, a linear chain of amino acid residues, folds into its characteristic and functional 3D structure from random coil. Amino acids interact with each other to produce a well-defined 3D structure, the folded protein, known as the native state. The energy landscape describes the folding pathways in which the unfolded protein can adopt its native and stable state at minimal free energy. Therefore, Energy acts as a measure of a protein's structural stability. Lower free energy (especially negative energy) is favorable for stabilizing the folded state of a protein, whereas an unstructured protein cannot achieve such a gain in enthalpy, therefore remains neutral in terms of energy. The stable 3D structure is determined by the amino acid sequence or primary structure, explained by Anfinsen's dogma [3]. Thus, the possible extraction of energy contribution of amino acid residues from protein sequence alone has its underlying hypothesis and will have crucial implications as a feature to characterize protein structure and stability in inducing a machine learning model that is capable of accurately predicting 3D protein structural descriptors. Solutions for existing protein structure prediction problems need features that can capture the complexity of molecular level interactions.

In this chapter, we propose a novel approach to estimate position specific energy (PSEE) of a residue using contact energy and predicted relative solvent accessibility (RSA). Furthermore, we demonstrate PSEE can be reasonably estimated based on sequence information alone. PSEE is found useful in identifying the structured as well as unstructured or, intrinsically disordered region of a protein by computing favorable and unfavorable energy respectively, characterized by appropriate threshold. The most intriguing finding, verified empirically, is the indication that the PSEE feature can effectively classify disordered versus ordered residues and can segregate different secondary structure type residues by computing the constituent energies. PSEE values for each amino acid strongly correlate with the hydrophobicity value of the corresponding amino acid. Further, PSEE can be used to detect the existence of critical binding regions that essentially undergo disorder-to-order transitions to perform crucial biological functions.

Towards an application of disorder prediction using the PSEE feature, we have rigorously tested and found that a support vector machine model informed by a set of features including PSEE consistently outperforms a model with an identical set of features with PSEE removed. We have synonymously mentioned the improved predictor as DisPredict (version 2.0) or DisPredict2. In addition, the new disorder predictor, DisPredict2, shows competitive performance in predicting protein disorder when compared with six existing disordered protein predictors. The outline of this chapter is as follows.

- We start by giving the background information about thermodynamic forces that stabilizes protein fold and energy landscape of structured and disordered proteins in Section 4.1. Here, we discuss our motivation behind estimating energy score from protein's primary sequence and our contribution.
- In Section 4.2, we describe the technique to compute of position specific estimated energy (PSEE) per residue from protein sequence.
- Section 4.3 highlights the capacity of PSEE to score protein's stability and capture different structural properties of protein residues, therefore its importance as a feature to predict protein structure prediction problems.
- In Section 4.4, we discuss the design and development of an improved version of the initially developed disorder predictor (described in Chapter 2), DisPredict (version 2.0), including the experimental materials, such as dataset collection, feature set preparation and performance evaluation metrics.
- In Section 4.5, we evaluate and compare the performance of DisPredict2 statistically and with case studies.
- Finally, we conclude in Section 4.6 with brief future research directions.

4.1 Background and Motivations

Protein exists as an unfolded polypeptide or random coil without any stable 3D structure when translated from a sequence of mRNA to a linear chain of amino acids. As the polypeptide chain is being synthesized by the ribosome, the linear chain begins to fold into its three-dimensional structure. Folding begins to occur even during translation of the polypeptide chain of amino acid residues, which interact with each other to produce a well-defined 3D structure. The correct 3D structure is essential to function, although some parts of functional proteins may remain unfolded [22], which performs dynamic function through heterogeneous conformations. Protein may fail to fold into native structure, which becomes inactive proteins or misfolded proteins in some instances, having toxic functionality. Several neurodegenerative and other diseases are believed to result from the accumulation of amyloid fibrils formed by misfolded proteins.

While a fully functional protein is usually the one that is appropriately twisted, coiled and folded into a specific three dimensional conformation, intrinsically disordered regions (IDRs) or, proteins (IDPs) [2, 7] remain unfolded under physiochemical conditions [2, 5, 7], discussed in Chapter 2. IDRs and IDPs become biologically active through disorder to structure transitions [5, 20, 44, 46, 53, 69, 70, 75]. The connection of IDPs with critical human diseases, such as cancer, cardiovascular diseases, neurodegenerative diseases, genetic diseases, diabetes, amyloidosis and others, has created research areas such as prediction of protein disorder, identification of induced folding region, or binding sites in disordered proteins and drug discovery.

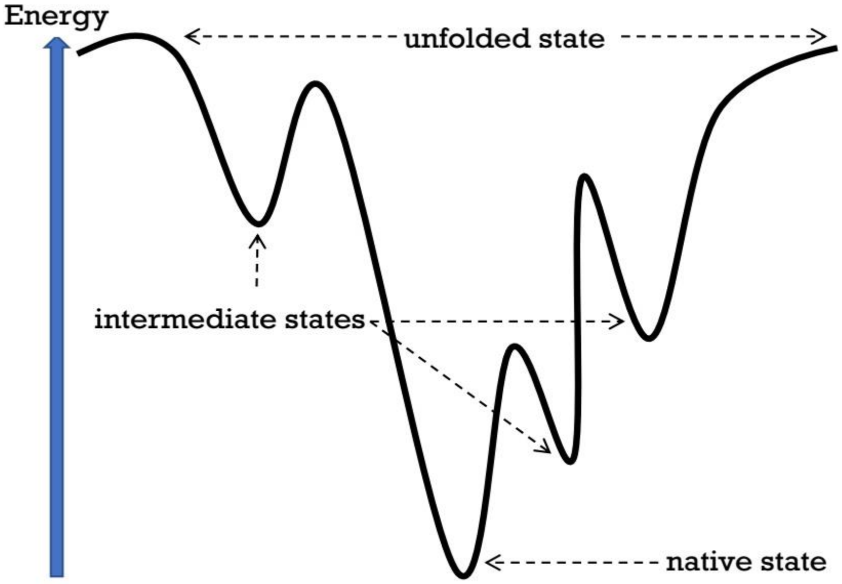


Fig 26. The Energy funnel: a sample energy landscape through which the unfolded linear chain of amino acid residues (primary structure) gains the 3D native structure. It shows that the unstructured proteins stay in less favorable energy state （less negative） and in native state, structured proteins stay in highly negative or favorable energy condition.

In late 1980s and early 1990s, Joseph Bryngelson and Peter Wolynes formulated the energy landscape theory of protein folding phenomenon [289]. This approach introduced the principle of minimal frustration [290]. The folding funnel landscape allows the protein to fold to the native state through any of many pathways and intermediates, rather than being restricted to a single mechanism. The computational simulations of proteins and experimental studies support this theory and is utilized by protein structure prediction and design methods.

A sample landscape is illustrated in **Fig 26**, which indicates that there are many initial possibilities, but only a single native state is possible; however, it does not reveal the numerous folding pathways that are possible. A different molecule of the same exact protein may be able to follow marginally different folding pathways, seeking different lower energy intermediates. Different pathways, having different thermodynamic favorability, can be utilized at different frequencies. Therefore, a pathway, being thermodynamically more favorable than another, is likely to be used more frequently in the pursuit of the native structure [291]. As the protein begins to fold and gets its various conformations, it seeks a more thermodynamically favorable structure than before and thus continues through the energy funnel. Thus, protein residues that are ordered (structured) can be assumed to contribute lower negative energy (favorable state) to the protein's 3D conformation whereas the disordered (unstructured) contribute less negative energy (unfavorable state).

The thermodynamic forces that stabilize the folded state of a protein comes from the formation of intramolecular non-covalent interactions, like *ionic bond*, *hydrogen bond*, *van der Waals attraction,* and a net force, *hydrophobic effect*. The ionic bonds in proteins, also called as salt bridges, occur due to highly favorable interaction between amino acids with side chains of opposite charge. A hydrogen bond is a strong form of dipole-dipole interaction between heteroatoms, which play a key role in the formation of secondary structure, such as alpha helices and beta sheets. The van der Waals' interaction, also known as London dispersion force, is the induced dipole-induce dipole interaction between nonpolar surfaces. The hydrophobic effect is the phenomenon in which the hydrophobic chains of a protein collapse into the core of the protein (away from the hydrophilic environment) [292]. Minimizing the number of hydrophobic side-chains exposed to water is an important driving force behind the folding process [293]. Furthermore, proteins will have limitations on their folding abilities by the restricted bending angles or conformations that are possible, described with a two-dimensional plot known as the Ramachandran plot, depicted with psi and phi angles of allowable rotation.

Protein folding is a spontaneous process, which is thermodynamically favorable within a cell, which follows the thermodynamic laws, thus the change in Gibbs free energy (ΔG) is negative. Gibbs free energy is defined in terms of enthalpy and entropy [292]. For a negative delta G to arise and for protein folding to become thermodynamically favorable, then either the change in enthalpy (ΔH) must be negative and dominant over an unfavorable entropy term, or the change in entropy (ΔS) must be positive and dominant over an unfavorable enthalpy term or, both terms must be favorable (negative change in enthalpy or positive change in entropy).

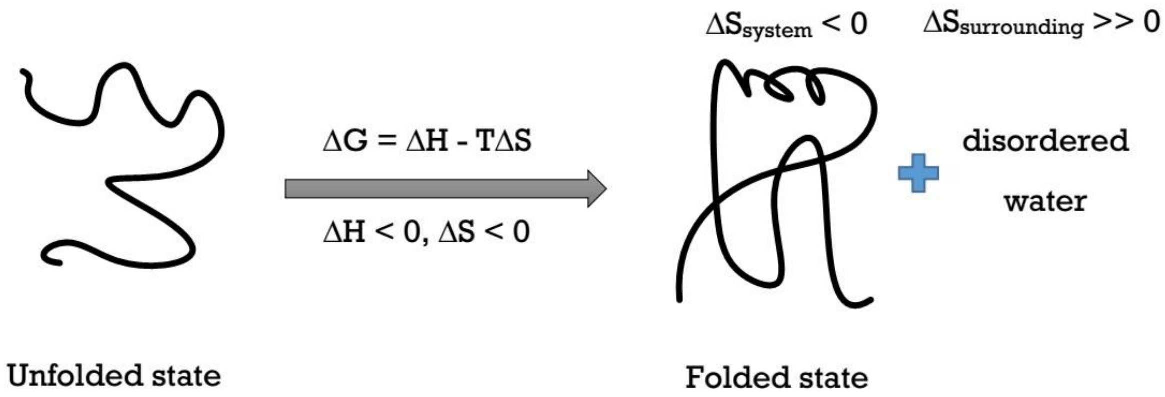

$\Delta S_{system} < 0$
$\Delta S_{surrounding} >> 0$
$\Delta G = \Delta H - T\Delta S$
$\Delta H < 0, \Delta S < 0$
disordered water
Unfolded state
Folded state

Fig 27. Thermodynamics of protein folding process. The spontaneous process of protein folding results in negative Gibbs free energy change（ΔG）. It includes negative change in enthalpy（ΔH） and decrease（positive change） in entropy of the system（ΔS_{system}）. However, the free energy causes an increase in the entropy of the surrounding environment（$\Delta S_{surrounding}$） resulting in disordered water and hence, a total increase of entropy.

Protein folding is an exothermic process, illustrated in **Fig 27**, with negative enthalpy change ($\Delta H < 0$) due to the formation of strong and short hydrogen bonds. The water molecules tend to aggregate around the hydrophobic regions or side chains of the protein, creating water shells of ordered water molecules [294]. The ordered water molecules around a hydrophobic region, driving hydrophobic collapse, decreases entropy of the system ($\Delta S_{system} < 0$). However, the heat given off by the exothermic process of folding causes the molecules in the surrounding to dance around more ($\Delta S_{surrounding} >> 0$), thus increases the total entropy ($\Delta S_{total} > 0$), following the 2nd law of thermodynamics.

For globular proteins, the contribution of interresidue interactions to total energy can be approximated by low resolution force fields, or statistical potentials, energy-like quantities derived from structured proteins based on the observed amino acid pairing frequencies [295, 296]. In deriving the actual potentials, different principles have been applied [295, 297-300]. The resulting empirical energy functions are well suited to assess the quality of structural models and have been used for fold recognition or threading [301, 302], in docking [303], ab initio folding [304] and predicting protein stability [305]. Their success in a wide range of applications suggests the existence of a common set of interactions, simultaneously favored in all native structures.

However, extraction of such energies is not possible to carry out for proteins whose structure is not known, like intrinsically disordered proteins (IDPs). To overcome these limitations, attempts have been made to predict the pairwise contact energy values among 20 different amino acids from sequence only [149] and found effective in characterizing ordered and disordered state [149] of protein residues. The underlying principle is the major milestone in protein science, the thermodynamic hypothesis of Christian Anfinsen, the primary structure of a protein, its linear amino-acid sequence, determines its native conformation. The amino acid composition is not as important as the sequence [292]. The essential fact of folding, however, is that the amino acid sequence of each protein contains the information that specifies both the native structure and the pathway to attain that state. Therefore, nearly identical amino acid sequences usually fold similarly [306].

4.1.1 Our Contributions

We propose a novel approach of predicting position specific residual energy contribution in the total energy of a protein. We predict this energy per residue from the protein's primary sequence alone unlike the energy functions [15, 179, 307] while the protein structure is given, and called as Position Specific Estimated Energy (PSEE) [11]. The computation of PSEE considers the potential contact partners (amino acids) and the contact energies in the neighborhood of the primary protein sequence as well as the relative burial of the target residues and its partners to capture the hydrophobic effect, which can be defined as the tendency of nonpolar (or hydrophobic) amino acids to become buried because that leads to increase the entropy of water. PSEE successfully characterized the ordered and disordered state of protein residues, different types of secondary structure residues, hydrophobic and polar amino acid residues.

An important implication of PSEE is its usability as a valuable feature for the development of sequence based predictors of disorder protein, secondary structure and accessible surface area and so on where 1D sequence information to 3D structural mapping is essential. As an application, we enhanced our disorder protein predictor, DisPredict [10] with the new PSEE feature, called DisPredict2 [11], and DisPredict2 outperforms DisPredict in predicting disorder residues more accurately.

4.2 Extraction of PSEE from Sequence

The free energy of a protein chain is a function of effective inter-residual contacts in its three-dimensional conformation. An iterative method is described by Thomas and Dill [297] to extract interaction potentials (ENERGI) from a database of protein structures obtained from Protein Data Bank (PDB) [308]. Initially, the 20 × 20 contact energy matrix in [297] is derived from known structures of 37 protein chains. A similar approach is applied in [149] to recalculate the contact energies between all possible pairs of 20 different amino acids using known structures of 785 proteins from PDB.

However, the amino acid composition in the primary structure of protein determines its native structure with favorable energy. Therefore, it is believed that the pairwise contact energy can be extracted from the amino acid sequence [149]. The predicted pairwise contact energies are derived in [149] using 674 protein's primary structure (amino acid sequence) by the least square fitting with the contact energies derived from tertiary structure of 785 proteins. The actual and predicted energies are found to have linear relationship, explained in [149]. The predicted energy matrix (*P*) derived in [149] is shown in **Table 18**.

Table 18. Predicted pairwise contact energy matrix derived in [149].

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	-1.65	0.98	0.66	1.16	-2.83	1.2	1.8	-0.41	1.9	-3.69	-3.01	0.49	-2.08	-3.73	1.54	-0.08	0.46	0.32	-4.62	-2.31
R	0.98	0.21	1.08	-2.02	-0.41	0.91	-3.13	0.84	0.19	2.05	-0.6	2.34	2.09	-0.4	1.06	0.95	0.98	-5.89	0.36	0.08
N	0.66	1.08	0.61	0.32	-4.18	1.28	0.2	-0.32	1.84	-0.07	0.97	1.12	0.21	0.73	1.15	0.29	0.46	-0.74	0.93	0.93
D	1.16	-2.02	0.32	0.84	-0.82	2.67	1.97	0.88	-1.07	0.68	0.23	-1.93	0.61	-0.92	3.31	0.91	-0.65	-0.71	0.9	0.94
C	-2.83	-0.41	-4.18	-0.82	-39.58	-2.91	-0.53	-2.96	-4.98	0.34	-2.15	-1.38	1.43	-3.07	-2.31	-2.33	-1.84	4.26	-4.46	-0.16
Q	1.2	0.91	1.28	2.67	-2.91	-1.54	0.1	1.11	2.64	-0.18	-0.58	0.43	1.9	0.77	-0.42	1.12	1.65	-2.06	-2.09	0.38
E	1.8	-3.13	0.2	1.97	-0.53	0.1	1.45	1.31	0.61	1.3	1.14	-2.51	2.53	0.94	1.44	0.81	1.54	-1.07	1.29	0.12
G	-0.41	0.84	-0.32	0.88	-2.96	1.11	1.31	-0.2	1.09	-0.65	-0.55	-0.16	-0.52	0.35	2.25	0.71	0.59	1.69	-1.9	-0.38
H	1.9	2.05	1.84	-1.07	-4.98	2.64	0.61	1.09	1.97	-0.71	-0.86	2.89	-0.75	-3.57	0.35	0.82	-0.01	-7.58	-3.2	0.27
I	-3.69	0.19	-0.07	0.68	0.34	-0.18	1.3	-0.65	-0.71	-6.74	-9.01	-0.01	-3.62	-5.88	0.12	-0.15	0.63	-3.78	-5.26	-6.54
L	-3.01	-0.6	0.97	0.23	-2.15	-0.58	1.14	-0.56	-0.86	-9.01	-6.37	0.49	-2.88	-8.59	1.81	-0.41	0.72	-8.31	-4.9	-5.43
K	0.49	2.34	1.12	-1.93	-1.38	0.43	-2.51	-0.16	2.89	-0.01	0.49	1.24	1.61	-0.82	0.51	0.19	-1.11	0.02	-1.19	0.19
M	-2.08	2.09	0.21	0.61	1.43	1.9	2.53	-0.52	-0.75	-3.62	-2.88	1.61	-6.49	-5.34	0.75	1.39	0.63	-6.88	-9.73	-2.59
F	-3.73	-0.4	0.73	-0.92	-3.07	0.77	0.94	0.35	-3.57	-5.88	-8.5	-0.82	-5.34	-11.25	0.32	-2.22	0.11	-7.09	-8.8	-7.05
P	1.54	1.06	1.15	3.31	-2.13	2.97	1.44	2.25	0.35	0.12	1.81	0.51	0.75	0.32	-0.42	1.12	1.65	-2.06	-2.09	0.38
S	-0.08	0.95	0.29	0.91	-2.33	0.85	0.81	0.71	0.82	-0.15	-0.41	0.19	1.39	-2.22	1.12	-0.48	-0.06	-3.03	-0.82	0.13
T	0.46	0.98	0.46	-0.65	-1.84	-0.07	1.54	0.59	-0.01	0.63	0.72	-1.11	0.63	0.11	1.65	-0.06	-0.96	-0.65	-0.37	1.14
W	0.32	-5.89	-0.74	-0.71	4.26	-0.76	-1.07	1.69	-7.58	-3.78	-8.31	0.02	-6.88	-7.09	-2.06	-3.03	-0.65	-1.73	-12.39	-2.13
Y	-4.62	0.36	0.93	0.9	-4.46	0.01	1.29	-1.9	-3.2	-5.26	-4.9	-1.19	-9.73	-8.8	-2.09	-0.37	-0.37	-12.39	-2.68	-3.59
V	-2.31	0.08	0.93	0.94	-0.16	-1.91	0.12	-0.38	0.27	-6.54	-5.43	0.19	-2.59	-7.05	0.38	0.13	1.14	-2.13	-3.59	-4.82

Here, we present a novel idea of extracting position specific estimated energy (PSEE) contribution of each residue in a protein from its sequence alone based on following two hypotheses.

Hypothesis 1: *The position specific energy for a protein residue includes the contact effects with different types of amino acid within a neighborhood along the primary sequence*.

The preliminary idea to predict pairwise energies in [149] agrees with the above hypothesis that the energy contribution of a residue depends on the amino acid type of that residue as well as the types of its partners in the sequence. Therefore, we utilize the energy matrix (P) derived in [149] to include the effect of having variable count of different amino acid type residues that can form favorable contacts with the target residue.

Hypothesis 2: *The position specific energy contribution of a protein residue is related to the relative solvent accessibility (RSA) of the target residue and the residues within its neighborhood region, which can essentially capture the hydrophobic effect on the 3D state.*

The RSA of a residue is used to determine its proportional exposure ($pExp$) or, burial ($pBurr$), and hence the effective contact surface that can characterize the local environment of that residue in the tertiary structure. In protein folding process, the hydrophobicity of the amino acid, having less $pExp$, acts as a driving force to develop the core in the tertiary structure and the hydrophilic residues usually stay on the surface of the protein with high $pExp$. Thus, $pExp$ (or, $pBur$) of a residue can provide useful information in capturing the local solvent effects and can help computing favorable (negative) energy contribution of that residue in the native structure.

Let, AA_i is the i^{th} amino acid residue of the protein sequence, where $i \in \{1, \dots, L\}$ and L is the length of that protein sequence. N_i is the neighborhood region around AA_i that consists of the contact partner residues of AA_i. N_i includes contact radius (CR) number of residues on the either side of target residue (AA_i). Thus, the size of N_i is equal to $2CR$. The predicted pairwise contact energy between AA_i and AA_j is denoted by $P(AA_i, AA_j)$, where AA_j belongs to N_i. We weight this contact potential by the proportional burial of the contact partners to capture the essential contact effect in the estimation of position specific energy of the target residue AA_i. Therefore, $PSEE(AA_i)$ is formulated as,

$$PSEE(AA_i) = pBur(AA_i) \left[\frac{\sum_{AA_j \in N_i} P(AA_i, AA_j) \times pBur(AA_j)}{2CR}\right] \tag{21}$$

4.2.1 Computation of Proportional Exposure (or Burial)

RSA of a protein residue is calculated by normalizing the accessible surface area (ASA) of that residue by the surface area of the same type of residue in a reference state. We used the ASA normalizing values derived in [309] using Gly-X-Gly tripeptide as the reference state for a given residue X. Therefore, the proportional exposure ($pExp$) and burial ($pBurr$) can be expressed by the following equations.

$$pExp(AA_i) = \frac{predicted\ ASA\ (AA_i)}{ASA\ (AA_i)\ in\ the\ conformation\ Gly - AA_i - Gly} \tag{22}$$

$$pBur(AA_i) = 1 - pExp(AA_i) \tag{23}$$

Table 19. ASA normalization values for 20 amino acids in Å^2, proposed in [309].

Amino Acid (AA)	ASA normalization value	Amino Acid (AA)	ASA normalization value
Alanine (A)	129.0	Leucine (L)	201.0
Arginine (R)	274.0	Lysine (K)	236.0
Asparagine (N)	195.0	Methionine (M)	224.0
Asparatate (D)	193.0	Phenylalanine (P)	240.0
Cysteine (C)	167.0	Proline (P)	159.0
Glutamine (Q)	225.0	Serine (S)	155.0
Glutamate (E)	223.0	Threonine (T)	172.0
Glycine (G)	104.0	Tryptophan (W)	285.0
Histidine (H)	224.0	Tyrosine (Y)	263.0
Isoleucine (I)	197.0	Valine (V)	174.0

The ASA normalization values are listed in **Table 19**. We utilized a new ASA predictor framework, REGAd^{3}p [15], described in **Chapter 3** to generate predicted ASA of the residues. REGAd^{3}p [15] is a new real-value ASA predictor from protein sequence alone that showed maximum Pearson correlation coefficient (PCC) value of 0.76 on a blind dataset

4.2.2 Determining Contact Radius (CR)

PSEE of a residue serves as a measure of structural stability of a residue being in a specific position. The structurally stable proteins, so as the residues of proteins, gains energetically favorable (negative) condition compared to the unstructured counterparts. The quantification of PSSE by **Equation 21** involves the determination of the contact radius (CR) of the neighborhood around the target residue. It is assumed that the target residue forms effective local contacts with the CR number of residues on its either side.

To determine the CR parameter for the computation of PSEE, we applied PSEE as a feature to characterize the structured (ordered) and unstructured (disordered) residues. We allowed the minimum CR value equal to 4 to maximum of 30. We performed this experiment on the DisProt database [100] of disordered proteins that stores manually curated annotations of ordered and disordered residues. The recent release of DisProt version 6.02 contains 694 proteins with 1539 disordered regions. We excluded three chains from this set, Id: DP00688, DP00195, DP00642, as they have unknown amino acids, such as X, B and Z. Furthermore, the Cysteine (C) amino acid, being highly reactive due to its sulfhydryl group, caused abnormal PSEE values for some residues of 11 more protein sequences which we have discarded for the mentioned reason. A very high Cysteine-Cysteine pairwise interaction energy is also explicit in **Table 18**. Thus, we excluded these 11 chains while tuning the value of CR. This purification resulted a list of 680 protein chains, called as DisProt680 dataset, from DisProt database [100]. After that, we computed mean PSEE, formulated by **Equation 24**, of DisProt annotated ordered (o) and disordered (d) residues.

$$\overline{PSEE}(o) = \frac{\sum PSEE(o)}{n_o} \text{ and } \overline{PSEE}(d) = \frac{\sum PSEE(d)}{n_d} \tag{24}$$

Here, n_o and n_d are the total number of ordered and disordered residues, respectively. We computed $\overline{PSEE}(o)$ and $\overline{PSEE}(d)$ for CR values of 4 to 30. For each value of CR, we define the threshold, $t(PSEE)$, for PSEE based identification of ordered and disordered residues as the value that is equally distant from $\overline{PSEE}(o)$ and $\overline{PSEE}(d)$. **Fig 28** shows the $\overline{PSEE}(o)$, $\overline{PSEE}(d)$ and $t(PSEE)$ for CR equal to 4 to 30.

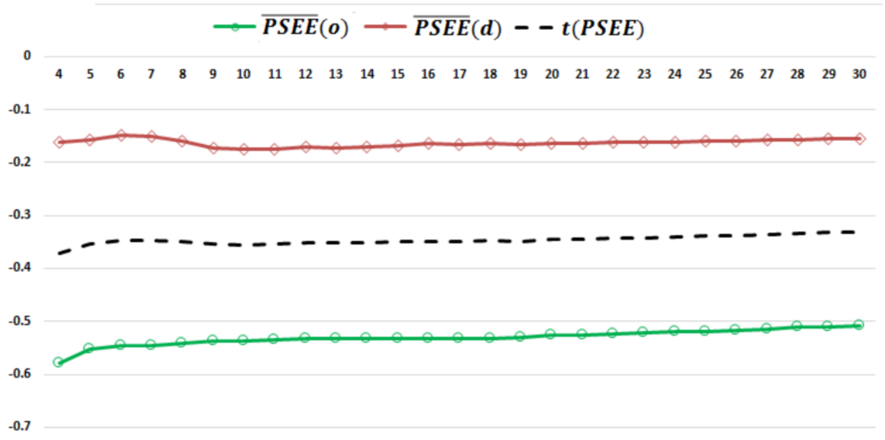
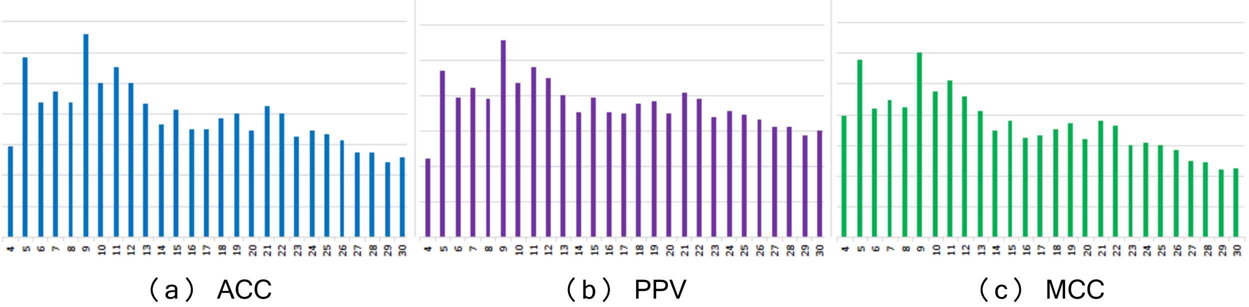


Fig 28. $\overline{PSEE}(o)$, $\overline{PSEE}(d)$ and $t(PSEE)$ for different contact (CR) values. Mean PSEE for ordered and disordered residues, indicated by *green line with circle marker* and *red line with diamond marker* respectively, of DisProt680 dataset for CR values of 4 to 30. The separation line or, threshold (t(PSEE)) is drawn with a black dashed line. The *x*-axis and *y*-axis show the CR and mean PSEE values, respectively.

Fig 28 illustrates that PSEE identifies the energetically induced gap between the structured and unstructured residue and clearly draws the separation line in terms of $t(PSEE)$ for all values of CR. For

CR value equal to 4 to 30, $\overline{PSEE}(o)$ ranges from -0.51 to -0.58, whereas $\overline{PSEE}(d)$ ranges from -0.13 to -0.15. Therefore, PSEE could recognize the energetically favorable (negative) condition of a structured residues. Now, we utilize $t(PSEE)$ of corresponding CR values to classify ordered versus disordered residues to determine the best CR value that can generate the most distinguishing PSEE values to classify ordered versus disordered class most effectively.



（a） ACC

（b） PPV

（c） MCC

Fig 29. Performance of ordered and disordered residue classification based on per residue PSEE value calculated using different contact radius (CR) values. Classification performance is shown in terms of （a） balanced accuracy, ACC （*blue bar*），（b） precision, PPV （*purple bar*） and （c） Matthews correlation coefficient, MCC （*green bar*） for CR values equal to 4 to 30. The *x*-axis and *y*-axis show the CR values and the performance metric values, respectively.

We plot the PSEE based disorder classification performance in terms of balanced accuracy (ACC), precision (PPV) and Matthews correlation coefficient (MCC) in **Fig 29**. We carried out this preliminary classification based on PSEE only to identify the effective CR value, thus we ignore the actual numerical values of the performance metrics here. **Fig 29** shows that PSEE values calculated with CR value 9 performs the disordered residue classification most accurately based on the DisProt680 dataset. Thus, we obtained the best CR value 9 and we used the same of rest our experiments in this work.

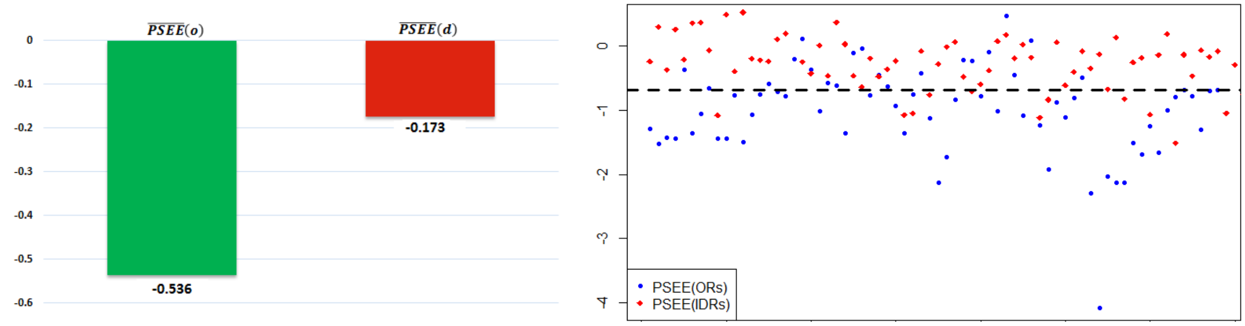
4.3 Performance of PSEE in Determining Structural Property

In this section, we highlight the usefulness of PSEE to characterize the structural stability of protein residues. Our results show that PSSE can effectively distinguish ordered and disordered residues, residues including three different types of secondary structures (helix, beta and coil) as well as residues with different physical property (hydrophobic and hydrophilic). Therefore, PSEE can effectively extract useful biological information from sequence that makes it a useful feature for machine learning based computational tools for disorder prediction, secondary structure prediction, residue exposure prediction, contact prediction, binding region prediction etc.

4.3.1 Ordered and Disordered Residues

Fig 30(a) shows the mean PSEE of ordered and disordered residues of DisProt680 dataset with contact radius of 9 on the either side of the target residue. The absolute gap between $\overline{PSEE}(o)$ and $\overline{PSEE}(d)$ is 0.363 that is reasonable to use PSEE feature for ordered versus disordered residue classification.

Further, we investigated the PSEE values at the region level. **Fig 30(b)** plots the PSEE values for IDRs and ordered regions (ORs) computed as the average PSEE values of the respective residues of the regions. The average PSEE value for all IDRs is -0.391 and that for ORs is -1.00. The black dashed line in **Fig 30(b)** shows the separation line, computed as the middle value (-0.698) of the two average PSEE values for all IDRs and ORs. Therefore, the region below -0.698 is energetically favorable, whereas above it is the unfavorable region. It shows that PSEE values for some IDRs falls into the favorable region as well.



（a） PSEE of Order and Disordered Residues

（b） PSEE of Ordered and Disordered Regions

Fig 30. Order versus disorder characterization of PSEE in residue and region level. （a） Mean PSEE for ordered （*green bar*） and disordered （*red bar*） residues of DisProt680 dataset. The bars are label with the respective mean PSEE values. （b） PSEE values for ordered regions （*green circle*） and disordered regions （*red diamond*）. The separation line between the average PSEE of all ordered and disordered region is indicated by black dashed line. The *x*-axis and *y*-axis represent the region index and the corresponding PSEE values.

To investigate it further, we segregate the IDRs into four types depending on the length of IDRs; IDRs with ≤ 5 residues, (5 – 20] residues, (20 – 40] residues and ≥40 residues. Then we compute the average PSEE for all IDRs having similar length range.

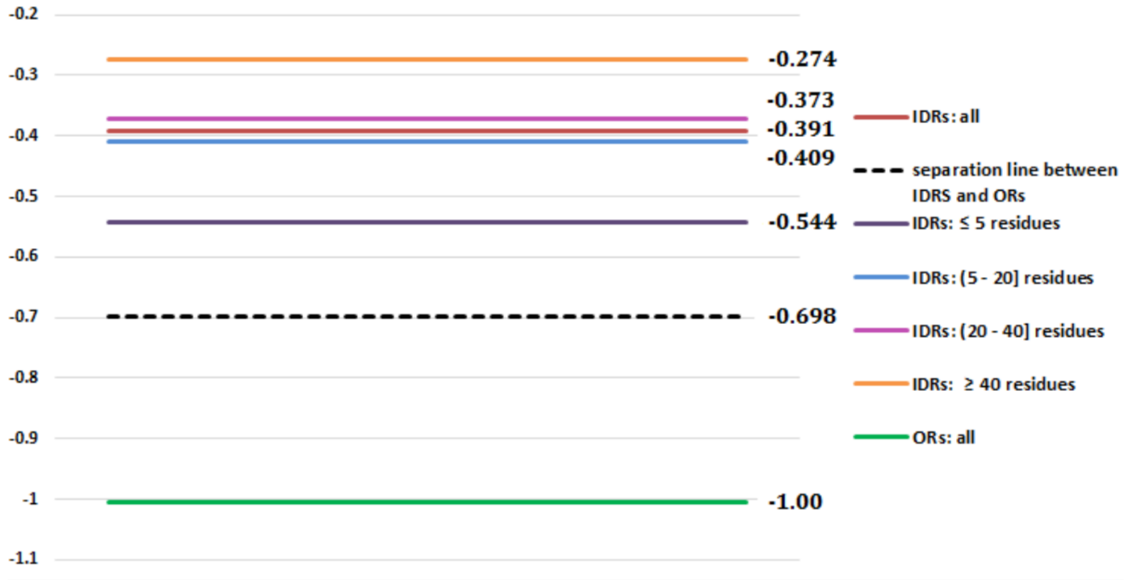


Fig 31. PSEE of different length disordered regions and all ordered regions. Average PSEE of different protein regions of DisProt680 dataset; ORs （*green*）， IDRs （*red*） IDRs with ≥ 40 residues （*orange*）， IDRs with （20 – 40] residues （*pink*）， IDRs with （5 – 20] residues （*blue*）， IDRs with ≤ 5 residues （p*urple*） and the separation line between all IDRs and ORs is shown by black dashed line. The lines are labeled by the corresponding numerical values of PSEE.

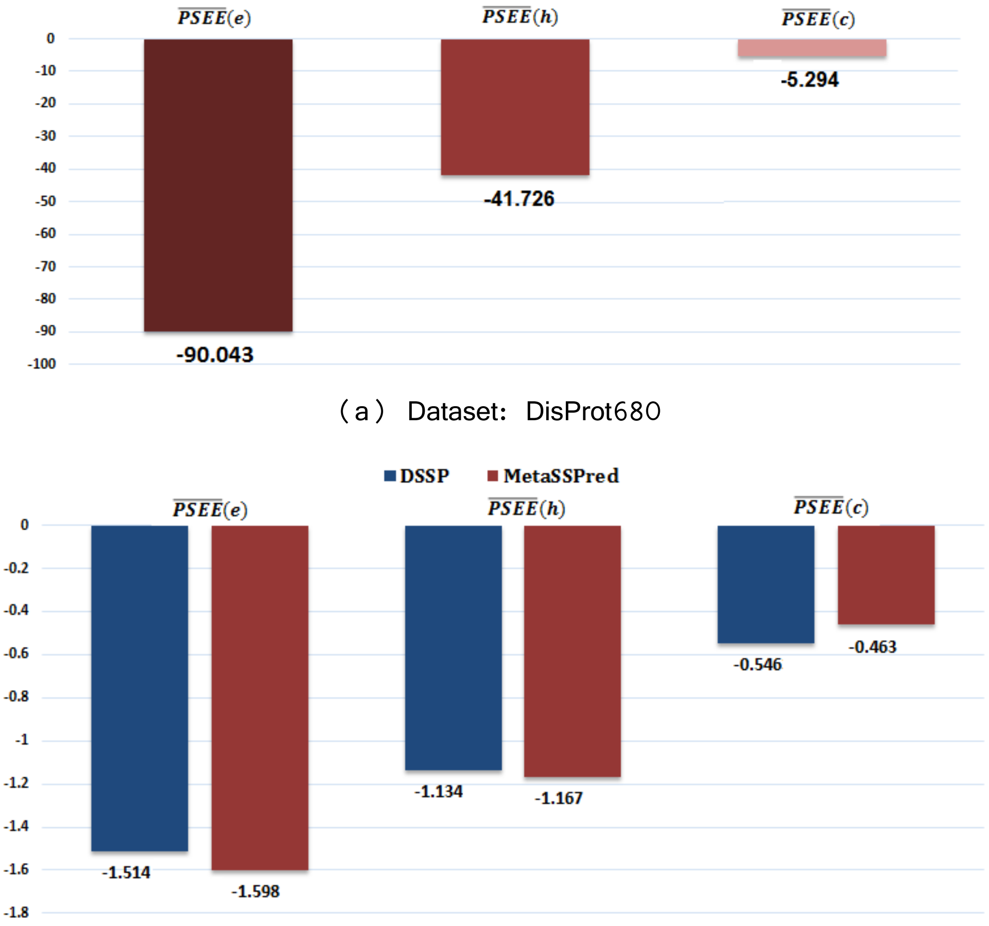
Fig 31 shows the average PSEE for all ORs, IDRs, 4 different types of IDRs along with separation line shown in **Fig 30(b)**. The relatively longer IDRs with (20 – `40] and ≥ 40 residues have PSEE values, -0.373 and -0.274, which are more unfavorable (less negative) than that of considering all IDRs, -0.391. Therefore, PSEE is useful in identifying long disordered regions. It is important to note that the average PSEE for shorter IDRs with ≤ 5 residues, -0.544, is close to the separation line, -0.698, and thus tends to have favorable energy. These short disordered regions are often called as binding sites which are biologically important, as they undergo disorder to order transition by interacting with various partners. Identifying the binding sites in disordered regions are one of the most recent research areas due to their functional importance. Our result shows that PSEE values for short disordered regions reflect the usefulness of PSEE in binding site prediction as well.

4.3.2 Helix, Beta and Coil Residues

To capture the performance of PSEE in capturing the structural differences of three different types of secondary structure residues (helix, beta and coil), we computed mean PSEE for helix (h), beta (e) and coli (c) residues using **Equation 25**.

$$\overline{PSEE}(h) = \frac{\sum PSEE(h)}{n_h}, \overline{PSEE}(e) = \frac{\sum PSEE(e)}{n_e}, \text{ and } \overline{PSEE}(c) = \frac{\sum PSEE(c)}{n_c} \tag{25}$$

We applied a new secondary structure predictor, called MetaSSPred [310], to generate predicted annotations for helix(h), beta (e) and coil (c) residues . MetaSSPred [310] is a balanced secondary structure predictor that can overcome the under prediction of less dominating beta residues in the datasets. Helices and beta residues are preferably located in the core of the protein, having favorable energy. Beta residues are more structured compared to the helix residues. On the other hand, coil residues stays in the surface areas of proteins and highly flexible, having unfavorable energy.



（a） Dataset: DisProt680

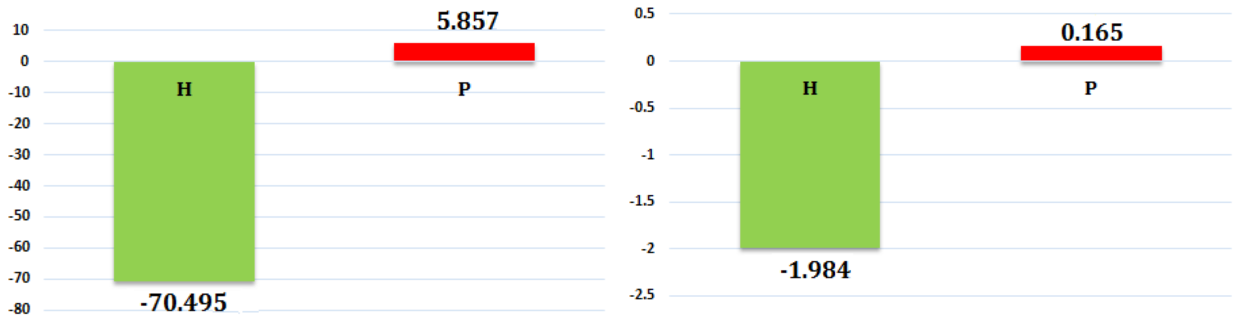
（b） Dataset: SSD1299

Fig 32. Secondary structure residue type characterization by PSEE. (a) Mean PSEE for beta （*dark brown bar*）, helix （*brown bar*） and coil （*light brown bar*） residues of DisProt680 dataset, predicted using MetaSSPred [196]. **(b)** Mean PSEE for beta, helix, and coil residues of SSD1299 dataset [15]. The *blue* and *brown* set of bars represent the actual and predicted annotations from DSSP [51] and MetaSSPred [196], respectively. The bars are label with the respective mean PSEE values.

Fig 32(a) shows the $\overline{PSEE}(h)$, $\overline{PSEE}(e)$ and $\overline{PSEE}(c)$ for residues of DisProt680 dataset. Beta residues have the highest negative PSEE and coils possess lowest negative energy, whereas helix residues stay in between beta residues and coil residues. This result is reasonable to validate the usefulness of PSEE in identifying different secondary structure residues. To further ensure this, we repeated the similar experiment on another dataset, generated by us in [15], specifically for secondary structure analysis. This dataset is called as secondary structure dataset (SSD) containing 1299 protein sequences with known structure from PDB. We ran DSSP [51] to generate the actual annotations of secondary structures for the residues of SSD1299 dataset and MetaSSPred [310] for the predicted annotations. The eight class annotations provided by DSSP are converted into three classes using the similar mapping given in [15, 310]. The mean PSEE values for the residues in SSD1299 dataset is shown in **Fig 32(b)**. PSEE consistently distinguished the three types of residues annotated by DSSP as well as MetaSSPred for SSD1299 dataset. Therefore, PSEE will serve as a useful feature for secondary structure prediction.

4.3.3 Hydrophobic and Hydrophilic Residues

Hydrophobic (H) amino acids build up the core of the protein and the hydrophilic or, Polar (P) ones are preferentially cover the surface of the proteins and are in contact with solvent due to their ability to form hydrogen bonds. Therefore, the hydrophobic residues gain energetically favorable condition compared to hydrophilic residues. Hydrophobic amino acids are A, G, I, L, M, F, P, W, Y and V, whereas the hydrophilic amino acids are R, N, D, C, Q, E, H, K, S, T. We computed mean PSEE for the H and P type residues of both DisProt680 dataset and SSD1299 dataset.

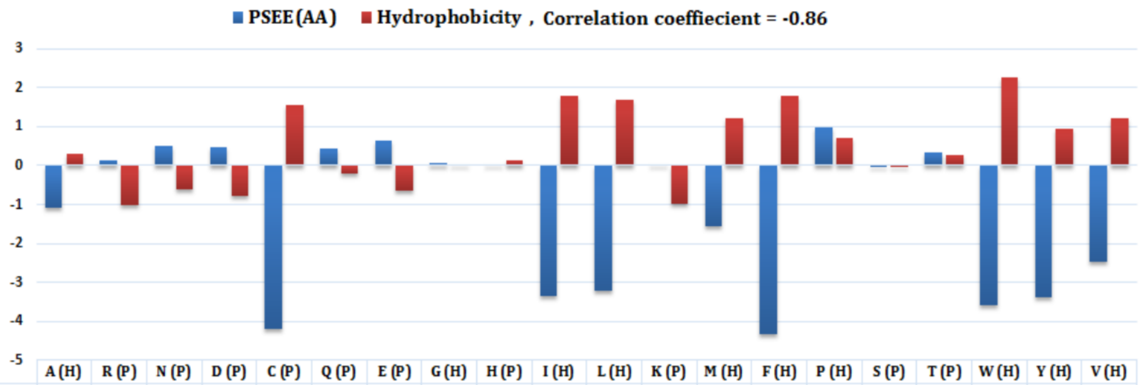


（a） Dataset: DisProt680 （b） Dataset: SSD1299

Fig 33. Mean PSEE of hydrophobic and hydrophilic residues. PSEE for hydrophobic （*green bar*） and hydrophilic （*red bar*） residues of **(a)** DisProt680 dataset and **(b)** SSD dataset. The bars are label with the respective mean PSEE values.

Fig 33 shows that for both datasets, the mean PSEE values for hydrophobic and hydrophilic residues are negative and positive, respectively. Thus, PSEE effectively discriminates hydrophobic and hydrophilic residues. As the hydrophobicity of the residues are directly related to the ASA of the residues, PSEE can serve as a useful feature for ASA prediction [15].

We further collected the hydrophobicity index for 20 different amino acids from [174] and computed mean PSEE for 20 different amino acid residues of SSD1299 dataset. Essentially the residues with positive hydrophobicity should obtain negative mean PSEE. **Fig 34** shows the correlation between hydrophobicity index and mean PSEE of 20 amino acid type residues with the correlation coefficient (CC) equal to -0.86. This result emphasizes that (aggregated) PSEE is strongly correlated with the physical property, hydrophobicity, of the amino acid residues, which in turn, confirms that the proposed approach is not deviating from the statistics obtained in previous work significantly [174].



	A (H)	R (P)	N (P)	D (P)	C (P)	Q (P)	E (P)	G (H)	H (P)	I (H)	L (H)	K (P)	M (H)	F (H)	P (H)	S (P)	T (P)	W (H)	Y (H)	V (H)
🟦 PSEE(AA)	-1.076	0.132	0.520	0.468	-4.179	0.431	0.643	0.078	0.033	-3.343	-3.216	0.016	-1.563	-4.335	0.988	-0.004	0.331	-3.602	-3.386	-2.464
🟥 Hydrophobicity	0.31	-1.01	-0.6	-0.77	1.54	-0.22	-0.64	0	0.13	1.8	1.7	-0.99	1.23	1.79	0.72	-0.04	0.26	2.25	0.96	1.22

Fig 34. Correlation between mean PSEE and hydrophobicity index of 20 amino acids. Mean PSEE (*blue bar*) and hydrophobicity index (*red bar*) of 20 different types of amino acid residues of SSD1299 dataset. The data values are given in the data table under the plot.

The negative value of correlation coefficient (CC) is desirable as the high (positive) hydrophobicity of a residue resembles its structural stability, thus its favorable (negative) energy contribution. Proline (P) and Threonine (T) are the exceptions here. Proline is referred as hydrophobic; however, it is found more in turns (coils) with unstable structure than helix and beta sheets. Thus, it has positive hydrophobicity as well as positive PSEE that correspond to unstable structure.

4.4 DisPredict (version 2.0)

In this section, we describe the materials and methods of our proposed DisPredict2 [11], anonymously mentioned as DisPredict (version 2.0), which is an enhanced version of our initially developed DisPredict framework. DisPredict2 uses our proposed novel feature, PSEE [11], into the feature set of our existing predictor, DisPredict [10]. DisPredict2 is available at http://cs.uno.edu/~tamjid/Software/PSEE/PSEE.zip.

4.4.1 Datasets

4.4.1.1 *Training Set*

We trained DisPredict2 with the same dataset as was utilized to train DisPredict [10] to have an accurate assessment of the effectiveness of the novel feature PSEE. DisPredict2 is trained with 477 protein sequences of Short-Long (SL) [10, 172] dataset.

SL477 dataset contains protein chains from DisProt [100] database. 50% of the disorder regions in this dataset are short with less than or equal to 20 residues, and rests are long. The allowable similarity between protein sequence pairs is 25%. SL477 dataset consists of approximately 25%, 34% and 40% of residues annotated as disordered, ordered and unknown. The unknown residues are annotated as 'X'. We ignored X residues for training and evaluation purposes.

4.4.1.2 *Test Set*

We tested and compared the performance of DisPredict2 with that of DisPredict [10] based on four independent datasets, DD73 [10], CASP8, CASP9 and CASP10. DD73 dataset is prepared by us and used as the holdout dataset in [10].

While the training dataset, SL477, is extracted from the protein chains of DisProt database version 5.0, DD73 accommodates 48 proteins from DisProt database version 5.1 to 6.02. The rest of the 25 single chain proteins are extracted from PDB [308] with the following criteria: *i*) X-ray structures with resolution ≤ 3.0 Å, *ii*) length ≥ 50 residues, and *iii*) 30% sequence identity cut-off. Later we removed sequences with more than 25% pairwise sequence similarity using BLASTCLUST[23] from NCBI-BLAST package [173]. Among 73 protein chains, 37 are fully disordered, 23 are fully ordered and 13 have both ordered and disordered regions. For DisPredict2, we utilized DD73 dataset for both independent evaluation of the predictor and optimization of threshold for disordered residue classification. However, CASP datasets are kept completely independent, and we did not carry out any optimization with the CASP datasets.

[23] BlastClust: ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html

CASP8 dataset contains 122 protein chains, of which, 103 are X-ray derived protein structures and 19 are NMR structures. This dataset has approximately 11% disordered residues, and the rest of the residues are structured.

We used 111 protein chains of CASP9 dataset to test and compare DisPredict2 versus DisPredict [10]. For this dataset, only 10% of the total residues were annotated disordered. CASP9 dataset has approximately similar proportion of X-ray and NMR derived protein structures.

In CASP10, 94 protein chains were used to assess the disorder predictors. For all CASP datasets, a residue is considered as disordered if it lacks spatial coordinates or, shows a high conformational variability across different X-ray structures or, NMR models.

4.4.2 Feature Set

In DisPredict2, we supplied the same 56 per residue features used in DisPredict [10], elaborately described in **Section 2.3.3** along with PSEE. Therefore, we have 57 features per residue in DisPredict2. The residue level information includes:

(*i*) Amino acid type, encoded by one single value, as all the necessary information for the correct folding of a protein is encoded in its amino acid sequence [3];

(*ii*) Seven physicochemical properties of amino acid as different types, short or long, disordered regions in protein are found to have distinguished physicochemical properties;

(*iii*) Twenty PSSM's (position specific scoring matrix) indicating the evolutionary information conserved in each residue position of a protein sequence;

(*iv*) Three predicted secondary structure (helix, beta and coil) probabilities from SPINE-X [175], one predicted relative surface area [178] and two predicted backbone torsion angle (phi, psi) fluctuations [179] since disordered residues are characterized by lack of stable secondary structure, highly exposed area and higher fluctuations of torsion angle;

(*v*) One monogram and twenty bigrams computed from PSSM [265] representing the conserved evolutionary information in three dimensional structure level;

(*vi*) One indicator for terminal residues, five residues from N terminal and C terminal are indicated by -1.0 to -0.2 and +0.2 to +1.0 respectively with a step size 0.2 and

(*vii*) One position specific estimated energy (PSEE) value.

Finally, before feeding the features into the classifier, 10 neighboring residue's, on the either side of the target residue, information is aggregated using a sliding window of 21, resulting in 21 × 57 = 1197 features per residue.

4.4.3 Predictor Framework

We developed DisPredict2 using support vector machine (SVM) algorithm, following our initially designed DisPredict predictor. We kept the dataset and classification algorithm similar to DisPredict to be able to compare the contribution of the proposed PSEE feature. SVM with radial basis function (RBF) kernel simultaneously minimizes the empirical classification error (training error) and generalized error (test error) by maximizing the geometric margin of the separating hyperplane. The DisPredict2 predictor framework has three levels:

Parameter optimization: The first level is the parameter tuning that determines the optimal values of two parameters for SVM classifier, namely C and γ, where C is the cost of misclassification that penalizes the feature space points on the wrong side of the decision boundary and γ is the parameter of RBF kernel. The parameter selection is done by grid search using 5% of the training dataset, which is guided by 5-fold cross validation with the accuracy (fraction of correctly predicted residues) optimization. The best parameter values found by the grid search is, C = 0.5 and $\gamma = 0.0078125$.

Model development: The second level of DisPredict2 development involves the prediction model that generates both binary annotations and real valued probabilities of order versus disorder residues. The probability range, $0.5 \leq \text{range} \leq 1.0$, is considered as disorder probability and $0.0 \leq \text{range} < 0.5$ is considered as order probability. The first and second level development of the predictor is done using LIBSVM [32].

Threshold optimization: The third level of the predictor is to optimize the threshold for disorder classification and to reannotate the residues accordingly. We employed Youden's J statistic [311] to find the optimal threshold for disorder prediction by analyzing the receiver operating characteristic (ROC) curve using pROC package [183]. This statistic determines the optimal cut-off that maximizes the distance from the identity (diagonal) line. The optimality criterion is formulated as,

$$max(sensitivities + specificities) \tag{26}$$

To make our predictor robust, we carried out the threshold optimization with an independent test dataset, DD73. The best threshold value found is 0.79. Therefore, we curated the annotation output given by the SVM model using $0.79 \leq \text{range} \leq 1.0$ as disorder probability and $0.0 \leq \text{range} < 0.79$ as order probability. Further, we scaled the probability range [0.0, 0.79) into [0.0, 0.5) for the ordered residues and

[0.79, 1.0] into [0.5, 1.0] for the disordered residues to make the DisPredict2's output more natural for binary classification.

3.3.2 Implementation and Availability

We implemented the DisPredict2.0 tool in C. The software is developed and tested on Linux platform. It is dependent on two external packages, namely PSI-BLAST[24] and NR database[25], which are publicly available. The software is available online[26] with a user manual. Besides disorder prediction, the software generates the pre-residue PSEE values of the target protein within the 'Features' sub-directory.

4.5 Evaluation of DisPredict2

In this section, we report the predictive performance of DisPredict2 [11] that measures the benefits of using PSEE as feature in the application of structure (or, disorder) classification and prediction while compared with DisPredict (version 1.0) [10]. The superior performance of DisPredict2 validates effectiveness of the proposed PSEE feature.

4.5.1 Performance Measures

The binary outputs given by DisPredict2 is evaluated and compared using the measures listed in **Table 20**. MCC is considered as the most balanced measure for binary classification. Moreover, we computed AUC, considered as the measure for the probability assignment. We further plotted the ROC curves and Precision-Recall curves. The AUC values and the curves are generated using ROCR package [312].

For a comprehensive comparison, we separately ranked the predictors in terms of balanced accuracy (ACC), Precision (PPV), Mathews Correlation Coefficient (MCC) and Area Under ROC curve (AUC). We gave same rank to all predictors having similar score. We assigned a cumulative score (S_c) as a summation of ranks according to different metrics and determined the final rank according to that cumulative score.

[24] PSI-BLAST link: ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/
[25] NR database link: ftp://ftp.ncbi.nlm.nih.gov/blast/db/
[26] REGAd^{3}p link: http://cs.uno.edu/~tamjid/Software/REGAd3p/REGAd3p.tar.gz

Table 20. Name and definition of performance measuring parameters.

Name of metric	Definition
True positive (TP)	Number of correctly predicted disordered residues
True negative (TN)	Number of correctly predicted ordered residues
False positive (FP)	Number of incorrectly predicted disordered residues
False negative (FN)	Number incorrectly predicted ordered residues
Balanced accuracy (ACC)	$\frac{1}{2}\left(\frac{TP}{TP+FN}+\frac{TN}{TN+FP}\right)$
Precision (PPV)	$\frac{TP}{TP+FP}$
Mathews correlation coefficient (MCC)	$\frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

4.5.2 Comparison with Other Predictors

We compared the performance of DisPredict2 with 7 others state-of-the-art disorder predictors. These predictors include our initial disorder predictor, DisPredict [10], SPINE-D [123], MFDp [164], MFDp2 [165], Espritz [313], IUPred-Long (IUPred-L) and Short (IUPred-S) [149].

SPINE-D [123] is a two-layer neural network based technique that was initially developed for three state prediction (disordered residues in short and long regions, ordered residue) and later reduced into two state prediction (disordered vs ordered residues). Espritz [313] is a high throughput predictor that uses recursive neural network.

MFDp [164] and MFDp2 [165] are meta predictors that combine different complementary disorder predictor’s output to have further curated prediction. MFDp [164] combines four predicted disorder probabilities from IUPred-L [148, 149], IUPred-S [148, 149], DISOPRED2 [128] and DISOclust [314], while it’s incremental version, MFDp2 [165], further incorporates sequence based predicted disorder content from DisCon [141].

IUPred-L [148, 149] and IUPred-S [148, 149] predict disordered residues in long and short regions, respectively, using predicted interaction energies. The formulation used in [148, 149] includes sequential local environment by involving interactions with potential partners. Our formulation of PSEE further improvises the pairwise energy based feature by strategically combining the proportional burial information of the potential partners that determines the structural local environment.

The result highlights that DisPredict2 is well competitive with different neural network based methods, meta-predictors as well as predictors that uses predicted pairwise energy as feature. Moreover, the comparative performance analysis of DisPredict2 versus DisPredict is provided to focus the utility of PSEE as feature for disorder prediction.

4.5.2.1 *Comparison on DD73 Dataset*

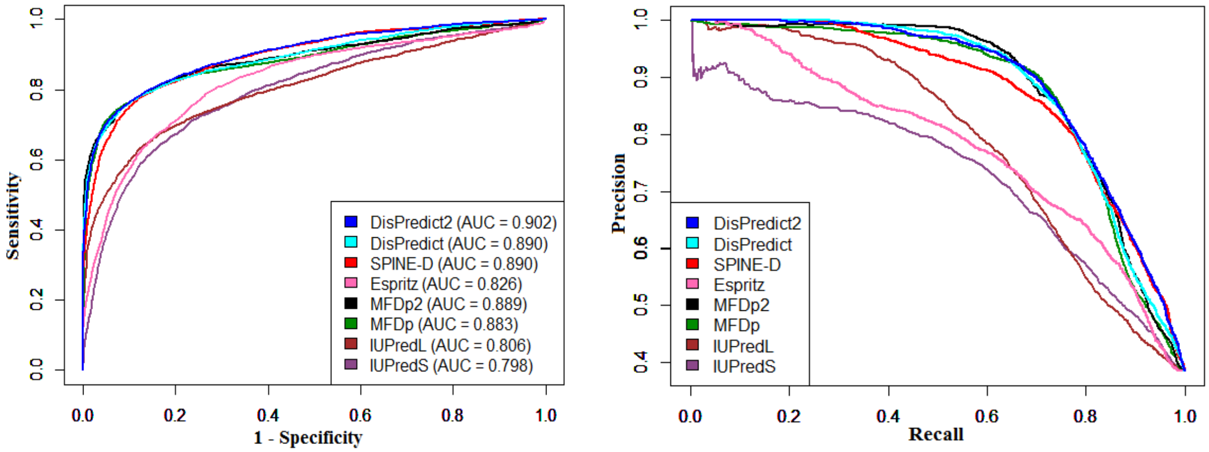
Table 21 shows the performance comparison based on DD73 dataset. This dataset is collected from both DisProt [100] and PDB [61] which is independent from the training dataset, SL477. DisPredict2 was assigned rank 1 in terms of ACC, MCC and AUC as well as achieved highest S_c with final rank of 1. MFDp2 gave the highest PPV only, however finally ranked 2 according to the overall performance. Moreover, DisPredict2 provided 0.41%, 6.35%, 3.48% and 1.36% improvement over DisPredict in terms of ACC, PPV, MCC and AUC under the ROC curve, respectively. These improvements focus the benefits of using PSEE as feature.

Table 21. Disorder prediction performances of 8 disorder predictors based on DD73 dataset.

Methods	Targets	ACC	PPV	MCC	AUC (ROC)	Ranks				Cumulative Score (S_c)	Final Rank
						ACC	PPV	MCC	AUC		
DisPredict2	73	**0.832**	0.857	**0.680**	**0.902**	**1**	2	**1**	**1**	**5**	**1**
DisPredict [10]	73	0.829	0.806	0.663	0.890	2	5	3	2	12	2
SPINE-D [123]	73	0.822	0.766	0.639	0.890	4	8	5	2	19	4
Espritz [313]	73	0.715	0.817	0.494	0.826	7	3	7	6	23	5
MFDp [164]	73	0.828	0.796	0.658	0.883	3	6	4	5	18	3
MFDp2 [165]	73	0.821	**0.873**	0.675	0.889	5	**1**	2	4	12	2
IUPred-L [149]	73	0.742	0.812	0.532	0.806	6	4	6	7	23	5
IUPred-S [149]	73	0.708	0.787	0.471	0.798	8	7	8	8	31	6

Best performances are marked by **bold**.

Fig 35 compares the ROC curves and precision-recall curves given by the predictors. DisPredict2, DisPredict and SPINE-D gave comparable ROC curves outperforming the others, while DisPredict2, DisPredict, MFDp and MFDp2 gave better precision for the recall range 0.3 to 0.8 than those of others.



(a) ROC Curves (b) Precision-Recall Curves

Fig 35. ROC and precision-recall curves given by 8 disorder predictors for DD73 dataset. Comparison of disorder predictors in terms of **(a) ROC curves** and **(b) precision-recall curves** on DD73 dataset. The area under ROC curves are given in the plot (a).

Table 22. Disorder prediction performances of 8 disorder predictors based on CASP8 dataset.

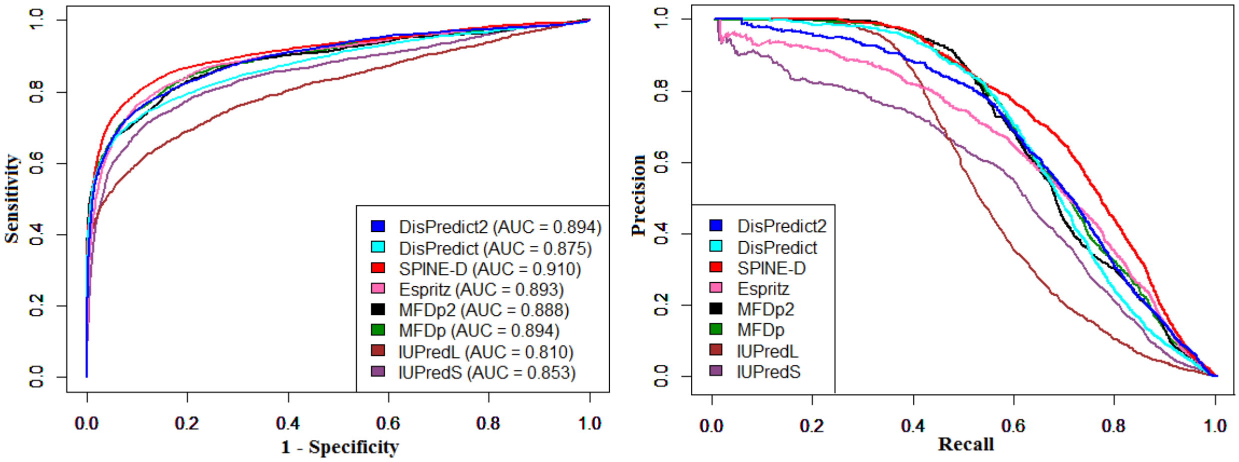
Methods	Targets	ACC	PPV	MCC	AUC (ROC)	Ranks				Cumulative Score (S_c)	Final Rank
						ACC	PPV	MCC	AUC		
DisPredict2	122	0.807	0.628	0.600	0.894	3	5	2	2	**12**	**1**
DisPredict [10]	122	0.810	0.529	0.551	0.875	2	7	6	6	21	6
SPINE-D [123]	122	**0.849**	0.504	0.576	**0.910**	**1**	8	5	**1**	15	4
Espritz [313]	122	0.797	0.636	0.592	0.893	5	3	4	4	16	5
MFDp [164]	122	0.806	0.634	0.601	0.894	4	4	3	2	13	2
MFDp2 [165]	122	0.774	**0.758**	**0.622**	0.888	6	**1**	**1**	5	13	2
IUPred-L [149]	122	0.722	0.700	0.531	0.810	8	2	8	8	26	7
IUPred-S [149]	122	0.766	0.624	0.551	0.853	7	6	6	7	26	7

Best performances are marked by **bold**.

4.5.2.2 *Comparison on CASP8 Dataset*

Table 22 shows the performance of the predictors based on CASP8 dataset. SPINE-D stood first in terms of ACC and AUC scores, however gave 33.5% and 7.4% lower PPV and MCC than those MFDp2 whose rank is 1 according to these two scores. DisPredict2 showed comparable performance in terms of all the metrics and attained the best cumulative score and finally ranked 1. Thus, the overall performance of DisPredict2 is promising. Furthermore, DisPredict2 provided 0.38% lower ACC than that of DisPredict while resulted 18.73%, 8.81% and 2.17% higher PPV, MCC and AUC than those of DisPredict.

Fig 36 compares the ROC curves and precision-recall curves. SPINE-D, Espritz, DisPredict2 and MFDp2 gave competitive ROC curves, while the SPINE-D resulted the best precision-recall curve.



（a） ROC Curves

（b） Precision−Recall Curves

Fig 36. ROC and precision−recall curves given by 8 disorder predictors for CASP8 dataset. Comparison of disorder predictors in terms of **(a) ROC curve** and **(b) precision−recall curve** on CASP8 dataset. The area under Roc curves are given in the plot （a）.

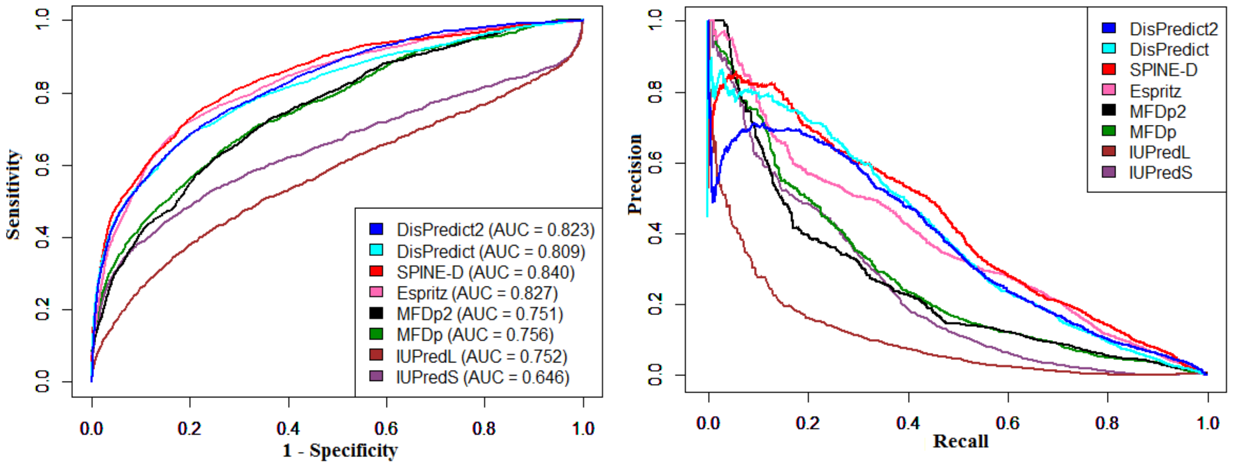
4.5.2.3 *Comparison on CASP9 Dataset*

The comparative performances of the predictors on 111 protein chains of CASP9 dataset are reported in **Table 23**. CASP9 dataset is a highly imbalanced dataset with approximately 10% of the residues are characterized as disordered. MCC is regarded as the best measure in evaluating prediction performance on such imbalanced dataset as it does not favor over prediction of dominating class. DisPredict2 resulted the best MCC and precision (PPV) score on CASP9 dataset, while ranked 3rd according to ACC and AUC. On the other than, SPINE-D gave the best ACC and AUC, however provided 26.5% lower precision than that of DisPredict2. DisPredict2 obtained the 1st position in final ranking with cumulative score difference of 2 and 4 from Espritz and SPINE-D respectively in 2nd and 3rd. Moreover, DisPredict2 with PSEE performed 20%, 5.76% and 1.69% better than DisPredict in terms of PPV, MCC and AUC, respectively, with slightly lower (2.66%) accuracy.

Table 23. Disorder prediction performances of 8 disorder predictors based on CASP9 dataset.

Methods	Targets	ACC	PPV	MCC	AUC (ROC)	Ranks				Cumulative Score	Final Rank
						ACC	PPV	MCC	AUC		
DisPredict2	111	0.699	**0.471**	**0.407**	0.823	3	**1**	**1**	3	**8**	**1**
DisPredict [10]	111	0.718	0.389	0.385	0.809	2	4	3	4	13	4
SPINE-D [123]	111	**0.745**	0.346	0.385	**0.840**	**1**	7	3	**1**	12	3
Espritz [313]	111	0.683	0.466	0.386	0.827	4	2	2	2	10	2
MFDp [164]	111	0.651	0.361	0.299	0.756	5	6	5	5	21	5
MFDp2 [165]	111	0.616	0.399	0.276	0.751	7	3	7	6	23	6
IUPred-L [149]	111	0.561	0.259	0.147	0.572	8	8	8	8	32	8
IUPred-S [149]	111	0.633	0.466	0.386	0.827	6	5	6	7	24	7

Best performances are marked by **bold**.



（a） ROC Curves　　（b） Precision-Recall Curves

Fig 37. ROC and precision-recall curves given by 8 disorder predictors for CASP9 dataset. Comparison of disorder predictors in terms of **(a) ROC curves** and **(b) precision-recall curves** on CASP9 dataset. The area under ROC curves are given in the plot （a）.

Fig 37(a) shows the ROC curves given by DisPredict2, DisPredict, SPINE-D and Espritz were competitive at different points as a result of different thresholds, whereas SPINE-D resulted the most consistent precision-recall curve. We observed a sharp drop of precision (PPV) in **Fig 37(b)** at a very low recall value for SPINE-D, DisPredict and DisPredict2. A precision-recall curve essentially plots the PPV and recall scores of a predictor at different threshold values. Therefore, these drops can be the result of having decreasing PPV values (truly positive results out of total positive test outcomes) at some threshold values. However, the PPV values had an increasing trend afterwards.

4.5.2.4 *Comparison on CASP10 Dataset*

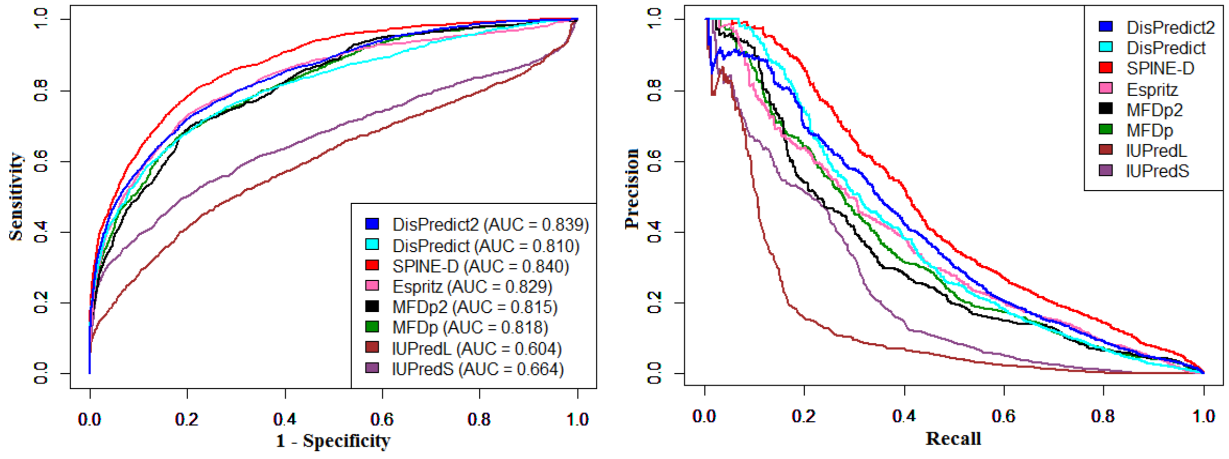
Table 24 illustrates the performance comparison on CASP10 dataset. This dataset has only 6.2% of the residues annotated as disordered. DisPredict2 achieved reasonable ranks, however not the best, in terms of all the scores. On the contrary, SPINE-D gave highest ACC and AUC values with very low precision (ranked 7). Similarly, MFDp2 showed the best precision with low ACC (ranked 6) and Espritz gave best MCC with low ACC (ranked 5). The cumulative rank of Dispredict2, SPINE-D and Espritz were same, therefore all three of them were finally ranked 1. Moreover, the performance of DisPredict2 is 39.06%, 15.73% and 3.58% higher in terms of PPV, MCC and AUC, respectively. Therefore, DisPredict2 turn out to be better disorder predictor than DisPredict [10] using PSEE as the only additional features.

Table 24. Disorder prediction performances of 8 disorder predictors based on CASP10 dataset.

Methods	Targets	ACC	PPV	MCC	AUC (ROC)	Ranks				Cumulative Score	Final Rank
						ACC	PPV	MCC	AUC		
DisPredict2	94	0.719	0.347	0.370	0.839	3	4	2	2	**11**	**1**
DisPredict [10]	94	0.734	0.249	0.320	0.810	2	7	6	6	21	6
SPINE-D [123]	94	**0.774**	0.269	0.366	**0.840**	**1**	6	3	**1**	**11**	**1**
Espritz [313]	94	0.674	0.441	**0.374**	0.829	5	2	**1**	3	**11**	**1**
MFDp [164]	94	0.677	0.359	0.336	0.818	4	3	4	4	15	4
MFDp2 [165]	94	0.636	**0.453**	0.332	0.815	6	**1**	5	5	17	5
IUPred-L [149]	94	0.569	0.238	0.160	0.604	8	8	8	8	32	8
IUPred-S [149]	94	0.635	0.331	0.278	0.664	7	5	7	7	26	7

Best performances are marked by **bold**.

Fig 38 shows that SPINE-D resulted better ROC and precision-recall curve consistently with the highest AUC and ACC values in **Table 24**, whereas the curves of DisPredict, DisPredict2 and Espritz were comparable.



(a) ROC Curves

(b) Precision-Recall Curves

Fig 38. ROC and precision-recall curves given by 8 disorder predictors for CASP10 dataset. Comparison of disorder predictors in terms of **(a) ROC curves** and **(b) precision-recall curves** on CASP10 dataset. The area under ROC curves are given in the plot (a).

4.5.3 Feature Correlation Plots with PSEE

Here, we further discuss the capacity of PSEE to capture multiple structural properties of the residues with DisProt680 dataset. We performed a similar analysis in Chapter 2 with coil probability and exposure of ordered and disordered regions to focus the possible noise within the annotation of disorder available in current state-of-the-art databases, which are utilized to train a disorder predictor. In this chapter, we carried out the analysis with PSEE and other structural properties.

Fig 39 shows the correlation between $pExp$ (or, $pBur$) and PSEE of disordered and ordered regions. The vertical dashed line is the separation (-0.698) of PSEE for ORs and IDRs, and the horizontal dash-dotted line indicates separation for exposed or, buried residues. We assume that residues with relative exposure, computed by **Equation 22**, less than 25% are buried. We collected ASA for the residues of DisProt680 dataset by running REGAd^{3}p [15]. Therefore, the left of the vertical line is the energetically favorable regions, and most of the ordered regions (*blue circle*) have PSEE in this region and most of the disordered regions (*red diamond*) have PSEE on the right side. Specifically, the first quadrant (top-right corner) of the plot is the major distribution area of the disordered regions with unfavorable (positive) energy and higher exposure. On the other hand, the third quadrant (bottom-left corner) of the plot is the essential region for ordered regions with favorable (negative) energy and lower exposure.

It is explicit in **Fig 39** that the PSEE values of most of the disordered regions are in the first quadrant. Therefore, PSEE can capture the exposure-property of the residues and at the same time can categorize them as ordered or, disordered. However, the other quadrants also contain some disordered regions.

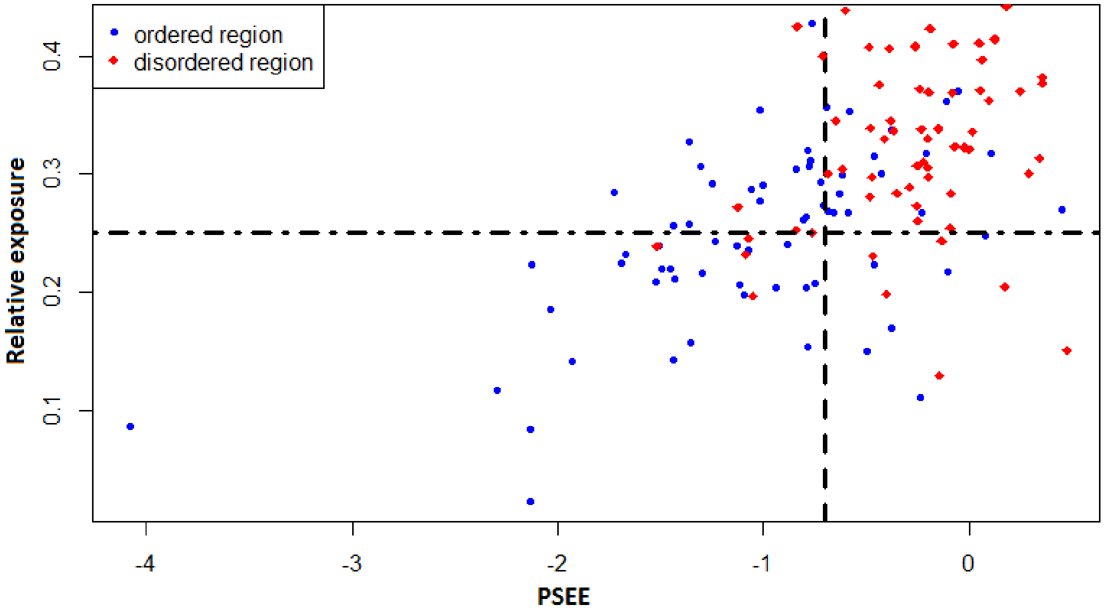


Fig 39. Correlation between PSEE and relative exposure of ordered and disordered regions. PSEE and relative exposure of ordered regions are shown by *blue circles* and those of disordered regions are shown by *red diamonds*. The vertical dashed line separates the average PSEE of ordered and disordered regions and the horizontal dash-dotted line separates the ordered and disordered regions with more and less 25% exposure.

Fig 40 shows the similar correlation analysis between the coil-like tendency and PSEE of disordered and ordered regions. We collected coil probability of the residues of DisProt680 dataset by running MetaSSPred [310] and assume that the residues with higher than 50% coil probability have flexible structure. Therefore, the first quadrant (top-right corner) of the plot is the essential area for disordered regions with unfavorable (positive) energy and high coil probability. On the other hand, the third quadrant (bottom-left corner) of the plot is the essential region for ordered regions with favorable (negative) energy and low coil probability.

Fig 40 shows that most of the PSEE values for ordered regions fall in the third quadrant, where as those of disordered regions fall in the first quadrant. However, for both **Fig 39** and **Fig 40**, the other quadrants also contain some disordered regions. This can be caused by mis-annotation of disorder [10] from DisProt database or, the disorder to order transition of binding sites.

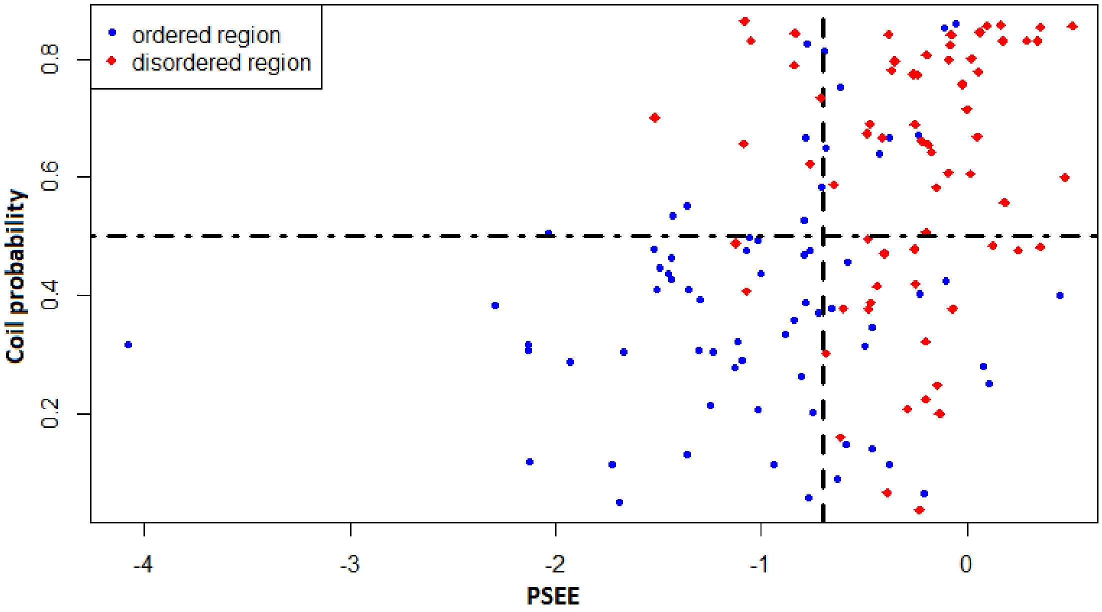


Fig 40. Correlation between PSEE and coil probability of ordered and disordered regions. PSEE and coil probability of ordered regions are marked by blue *circles* and those of disordered regions are marked by *red diamonds*. The vertical black dashed line separates the average PSEE of ordered and disordered region and the horizontal dash−dotted line separates the ordered and disordered regions with more and less 50% coil probability.

To further investigate, we searched for possible PDB models of each of the 694 sequences from DisProt disorder protein database of version 6.01. Each of the disorder protein sequences were crosschecked with approximately ~300,000 protein sequences from PDB (March, 2016) and we found that 155 IDPs or Proteins with IDRs of DisProt have structures in PDB. Specifically, 155 sequences of DisProt were mapped to 1226 sequences from PDB where some protein sequences have multiple structures in PDB for exactly same FASTA sequence. Therefore, our investigation validates the possibility of disorder-to-order transition of disordered proteins. This results also highlight the rationale behind the order-like characteristics of several IDRs reported in DisProt and the overlap found in feature correlation plot analysis.

4.5.4 Amyloidogenic region (AR) prediction by DisPredict2

To emphasize the biological significance of the outputs provided by DisPredict2, we attempted to evaluate it can detect the aggregation of amyloids. The proteins with amyloidogenic regions (ARs) are insoluble, however can improperly interact to fold and form amyloids. ARs play important role in protein aggregation, and they are directly linked with critical human diseases such as neurological disorder**.**

We collected 7 sequence from AMYPdb [315] and computed disorder probabilities of the residues by DisPredict2. **Fig 41** shows the location ARs, mean (drp_{mean}) and standard deviation (drp_{stdev}) of disorder probabilities of the residues of ARs, along with probability plot for the proteins.

4.5.4.1 *UniProtKB – P61769 (B2MG_HUMAN)*

P61769 is a human B2M (Beta-2-microglobulin) protein, a component of the class I major histocompatibility complex (MHC) and involved in the presentation of peptide antigens to the immune system. The AR is located within

residues 21 – 119. The drp_{mean} and drp_{stdev} given by DisPredict2 are 0.324 and 0.181, respectively.

4.5.4.2 *UniProtKB – P61626 (LYSC_HUMAN)*

P61626 is a human LYZ (Lysozyme C) protein, associated with the monocyte-macrophage system and enhance the activity of immune-agents. The AR is located within residues 19 – 148. The drp_{mean} and drp_{stdev} given by DisPredict2 are 0.352 and 0.205, respectively.

4.5.4.3 *UniProtKB – P0DJI8 (SAA1_HUMAN)*

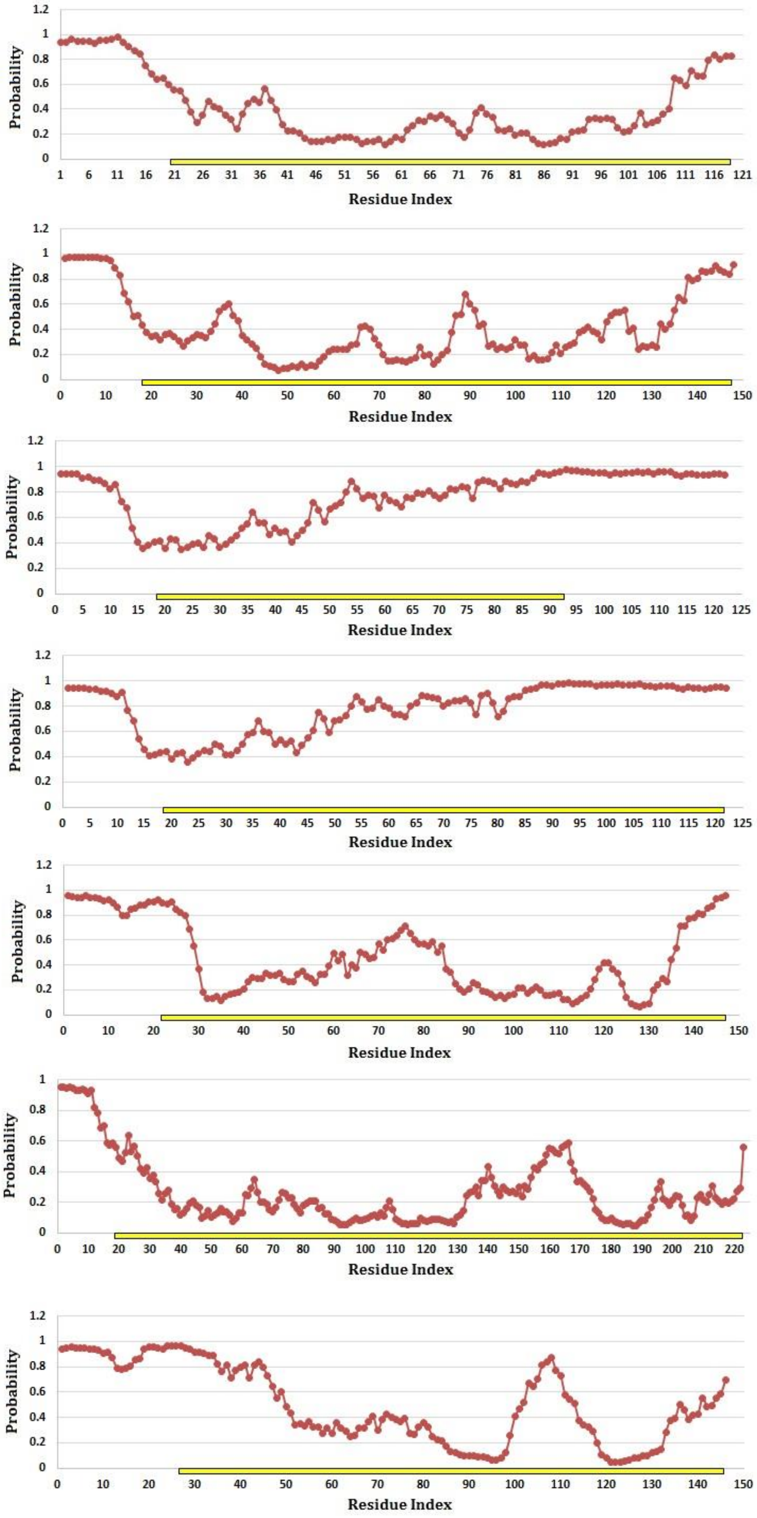
P0DJI8 is a human SAA1 (Serum amyloid A – 1) protein. Extracellular accumulation of SAA1 protein causes secondary amyloidosis, which is associated with disruption of tissue structure and lung cancer like diseases. The AR is located within residues 19 – 94. The drp_{mean} and drp_{stdev} given by DisPredict2 are 0.680 and 0.194, respectively.

4.5.4.4 *UniProtKB – P0DJI9 (SAA2_HUMAN)*

P0DJI9 is a human SAA2 (Serum amyloid A – 2) protein. Extracellular accumulation of SAA2 protein causes secondary amyloidosis, which is associated with disruption of tissue structure and compromise functions. The AR is located within residues 19 – 122. The drp_{mean} and drp_{stdev} given by DisPredict2 are 0.773 and 0.196, respectively.

4.5.4.5 *UniProtKB – P02766 (TTHY_HUMAN)*

P02766 is a human TTR (Transthyretin) protein, binds with thyroid hormone and transports thyroxine from the bloodstream to the brain. Dissociation of tetramer and partial unfolding leads to the formation of aggregates and amyloid fibrils. The AR is located within residues 21 – 147. The drp_{mean} and drp_{stdev} given by DisPredict2 are 0.383 and 0.245, respectively.



（a） UniProtKB – P61769 （B2MG_HUMAN） drp_{mean} = 0.324 drp_{stdev} = 0.181

（b） UniProtKB – P61626 （LYSC_HUMAN） drp_{mean} = 0.352 drp_{stdev} = 0.205

（c） UniProtKB – P0DJI8 （SAA1_HUMAN） drp_{mean} = 0.680 drp_{stdev} = 0.194

（d） UniProtKB – P0DJI9 （SAA2_HUMAN）， drp_{mean} = 0.773， drp_{stdev} = 0.196

（e） UniProtKB – P02766 （TTHY_HUMAN） drp_{mean} = 0.383 drp_{stdev} = 0.245

（f） UniProtKB – P02743 （SAMP_HUMAN） drp_{mean} = 0.213， drp_{stdev} = 0.238

（g） UniProtKB – P01034 （CYTC_HUMAN） drp_{mean} = 0.419 drp_{stdev} = 0.266

Fig 41. Disorder probability plots for proteins with amyloidogenic regions (ARs) given by DisPredict2. The *yellow bar* indicates the ARs and the *red line* shows the disorder probability of each residue indicated by circle marker. The description of protein, location of ARs are given on the label of each plot along with the mean and standard deviation of disorder probability for the AR.

4.5.4.6 *UniProtKB – P02743 (SAMP_HUMAN)*

P02743 is a human APCS (Serum amyloid P-component) protein, found in basement membrane and associated with amyloid deposits. It can interact with DNA and histones and may scavenge nuclear material released from damaged circulating cells. The AR is located within residues 20 – 223. The drp_{mean} and drp_{stdev} given by DisPredict2 are 0.213 and 0.238, respectively.

4.5.4.7 *UniProtKB – P01034 (CYTC_HUMAN)*

P01034 is a human CST3 (Cystatin-C) protein, an inhibitor of cysteine proteinases and serves an important physiological role as a local regulator of this enzyme activity. Cystatin C amyloid deposition in the cerebral vessels results in cerebral amyloid angiopathy, cerebral hemorrhage and premature stroke. The AR is located within residues 23 – 146. The drp_{mean} and drp_{stdev} given by DisPredict2 are 0.419 and 0.266, respectively.

Fig 41 shows that the mean disorder probabilities (drp_{mean}) for seven amyloidogenic regions range from 0.213 to 0.776, with an average of 0.45 (approximately in the middle of the probability range) and high standard deviation (drp_{stdev}) of 0.203. Therefore, DisPredict2 identified the flexibilities associated with the disorder (without amyloid formation) to order (with amyloid formation) transitions and the associated structural flexibilities of amyloidogenic regions.

4.6 Discussion

In this chapter, we describe the extraction of position specific estimated energy, named as PSEE, for each residues of a protein, based on sequence information alone. The quantification of PSEE includes the interaction effect of the target residue within a neighborhood in terms of pairwise contact energies between different amino acid types. We define the estimated neighborhood size in terms of number of residues on either side of the target residue with which it can form favorable contacts. Further, it utilizes the predicted relative exposure (or, burial) of a residue to approximate the local three-dimensional conformational position and stability of the residue. The source code to compute PSEE is written in C and the code is publicly available in open source format https://github.com/tamjidul/DisPredict2_PSEE.

Our result shows that PSEE is very effective in characterizing ordered (structurally stable) and disordered (structurally unstable) residues as well as regions in protein sequences. A fine-grained analysis further highlights that the average PSEE of the residues of binding site in disordered regions is well separable from those of disordered or, ordered regions. Therefore, PSEE detects the existence of critical binding regions in disordered proteins that undergo disorder to order transition and perform crucial biological functions [316]. Moreover, PSEE is effective in distinguishing the residues of two different datasets with three different types of secondary structures (helix, beta and coil). The residues with complementary physical properties, such as hydrophobic and hydrophilic, are promisingly identified by PSEE. Moreover, it strongly correlated with the respective hydrophobicity index of 20 different types of amino acid.

This promising correlation among different structural properties and PSEE of protein residues motivated us to propose PSEE to be utilized as a feature for the development of predictive tools in the area of bioinformatics and computational biology. To validate our argument, we construct DisPredict2, a new disorder protein predictor, integrating PSEE in the feature set of an existing disorder protein predictor, DisPredict [10]. DisPredict2 is implemented in C and the code is publicly available in open source form at https://github.com/tamjidul/DisPredict2_PSEE.

DisPredict2 showed improved performance over DisPredict [10] on 4 different datasets including CASP8, CASP9 and CASP10 datasets. Moreover, the disorder probability output given by DisPredict2 resembles the flexible structural transformation of amyloidogenic regions of proteins. Therefore, we believe that the new position specific residual feature, PSEE, and the disorder predictor, DisPredict2, both will be effective in understanding several insights of protein structures and hence the respective functions.

Chapter 5

PBRpredict: A Peptide-Binding Residue Predictor

— A Framework using Stacked Model

Protein-protein interactions (PPIs) play a key role in the biological processes as well as pathogenic processes in a living cell through physical interactions among multiple proteins within a complex. A major portion of the PPIs involve recognition of linear peptides by globular *Peptide Recognition Domain* (PRD) that induce binding with peptides and can form transient complexes. Human proteome contains millions of peptide motifs that are typically part of disordered regions and bind with appropriate partners through disorder-to-order transition. While in contact with the binding partners, transiently interacting peptide-protein complexes are involved in a wide range of molecular activities. Therefore, it is crucial to identify the peptide motifs in proteome and link the motifs to the domains that recognize them. Identification of peptide-binding residues in proteins that promote transient interactions is a pre-requisite for identifying peptide motifs. Specifically, peptide-binding tendency of proteins with different PRDs can be utilized to scan a proteome to identify the peptides likely to bind a particular PRD. Thus, recognition of peptide-binding residues is crucial for assembling peptide-mediated interactomes.

In this chapter, we computationally study the two-player complex process of induced-binding between peptides and protein with peptide-recognition domains [14]. With a view to this, we propose a new computational framework to predict peptide-binding residues (PBR) of receptor proteins in peptide-protein complex, called *PBRpredict* (**P**eptide-**B**inding **R**esidues **Predict**or) [317]. PBRpredict classifies binding and nonbinding residues from protein sequence alone as well as generates a probability score. PBRpredict is developed by stacking different learning models. To develop the model, we explored six different machine-learning algorithms as base learners: support vector machine, gradient boosting, bagging, random forest, extra tree and k-nearest neighbor classifier. The outputs of the base learners were aggregated using a meta-learner, which was logistic regression classifier.

For this study, a set of protein complexes with a wide range of peptide-binding domains was collected from PDB and the sequences with domains were annotated with interaction information based on atomic distances from peptide residues in the structure. Using a comprehensive set of sequence-based features including chemical and evolutionary profile, secondary structure, surface area and local backbone profile, flexibility and an energy based profile, we guide our predictor to learn about peptide-binding residues. We carried-out a rigorous performance evaluation using statistical metrics and case studies. After careful analysis of the prediction performance, we tuned the classification thresholds of the base-level and the meta-level learners of the stacking approach to trade-off between the true positive rate and false positive rate. Finally, we established three different PBRpredict models of a similar framework that apply different thresholds for segregating binding and non-binding residue under the name PBRpredict-Suite. The results manifest that PBRpredict-Suite models, provide well-balanced and biologically relevant outputs for proteins of different lengths and with a wide variety of PRDs.

Further, we computed the score, with a novel approach, named *PSBE* (**P**osition **S**pecific **B**inding **E**nergy), to approximate the binding energy contribution of the hot spot residues of peptide on the peptide-protein interaction surface. To extract PSBE from sequence alone, we utilized a similar concept that we used to compute Position Specific Estimated Energy (PSEE) as described in **Chapter 4**. PSBE is found effective in recognizing the prevalent amino acids in the hot spots of the peptides. The outline of this chapter is as follows.

- In section 5.1, we start by giving the background information about peptide-protein interactions, peptide-recognition domains and disorder-to-order transition of peptides. Moreover, we have defined the problem under consideration and reviewed the relevant literature in this section.
- In Section 5.2, we describe the experimental materials, including the definition of peptide-binding residues and regions, data collection and mining process, input features used to train the predictor, and the criteria to evaluate and compare the predictor.
- Section 5.3 describes the design and development of the predictor, PBRpredict-Suite models.
- We described the performance evaluation report for window selection, feature selection, parameter selection and comparison of PBRpredict-Suite models with existing predictors in Section 5.4.
- In Section 5.5, we describe the formulation of the position specific binding energy (PSBE) score and its effectiveness in recognizing the hot spots of peptide.
- Finally, in Section 5.6 we draw conclusions with brief future directions.

5.1 Background and Motivation

Interactions between proteins are essential for the vast majority of biological processes of a living cell through physical contacts among multiple proteins within a complex [318, 319]. Proteins carry signals and are the primary controller of the cell functionalities, including gene expression, cell growth, proliferation, morphology and intercellular communication. While proteins can independently function, a majority of the proteins interact with others for biological activity and correspondence [320]. These interactions can occur in different pace, such as long-time stable interactions in homo-oligomers and transient interactions between short linear peptides with globular protein receptors [321, 322].

The short peptides usually originate in intrinsically disordered proteins or regions in proteins (IDPs/IDRs) [68, 129, 323] that remain unstructured in an unbound stage, however, undergo conformational changes upon transient interactions only at the presence of a suitable binding partner. At about 40% of the PPIs involve recognition of linear peptides (about 5 – 25 amino acid long) by a globular receptors that have a peptide recognition domain (PRD) and can induce binding [324] with peptides, and promote formation of transient complexes [320]. While in contact with the binding partners (synonymously called ‘receptors’ in this chapter), transiently interacting peptides are involved in a wide range of molecular activities, including protein scaffolding, modification, transport, folding, signaling, and cell cycling. Moreover, 22% of human disease mutations occur in disordered segments of proteins with such motifs. Therefore, fast identification of regions in globular receptor proteins that promote transient interactions would be crucial for assembling potential interactomes and mapping signaling network.

At the present time, a growing number (around 200 [325]) of modular protein-interaction domains that mediate peptide-protein interactions have been identified, i.e., SH2, 14-3-3, Chromo and Bromo, SH3, Tudor, MBT, VHS, CW, PDZ (PDZ 1 and PDZ 2), PTB, WW, Glycine-Tyrosine-Phenylalanine (GYF) and MHC domains [326]. PRDs such as SH2, 14-3-3, Chromo and Bromo domains serve to recognize post-translational modifications (PTMs) of amino acids (such as phosphorylation, acetylation, methylation etc.) [327] and translate these into discrete cellular signals. Other domains such as SH3 and PDZ recognize linear peptide epitopes and serve to organize protein complexes based on localization and regions of elevated concentration. In both cases, the ability to nucleate specific signaling complexes is in large part dependent on the selectivity of a given peptide-recognition domain for its cognate peptide ligand.

5.1.1 Role of *In Silico* Techniques in Peptide-Binding Residue Prediction

Peptide is an interesting class of molecule that shows strong activity, low toxicity and few drug-drug interactions, therefore it is worth investigating their interactions with globular partners to develop new therapeutic agents [328, 329]. Compared to the size of known interactomes [330], a relatively lower proportion (approximately 20%) of the human protein interactions have been explored using experimental techniques, like peptide or protein arrays, phase display, mass spectrometry, HTP technique etc. [325].

High-throughput (HT) experimental techniques such as yeast two-hybrid and tandem affinity purification have been developed and applied to discover protein-protein interactions (PPIs) in multiple organisms on a genome-wide scale [331]. However, these approaches have inherent limitations and can provide substantial false positive rate [331, 332] with many interactions likely undiscovered due to high rates of false negatives [331, 333, 334]. The development of reliable computational approaches to identify PPIs is therefore an important alternative to HT experimental techniques [335, 336]. Moreover, for only a few major PRDs such as PDZ and SH3 domains, HT experimental techniques [326, 337, 338] such as phage display have been used to derive binding preferences.

However, a computational predictor of peptide-binding residues or regions of wide variety of PRDs will have useful implications, which can subsequently be used to scan a genome to identify proteins that are likely to bind a given PRD. Although this is challenging to build an accurate predictor of peptide-binding residues from protein sequence alone, it is vital to cope with the sequencing speed and demand which motivated us towards the study carried out in this chapter. We predicted regions of interaction in partner proteins as well developed a score to identify the hot spots of peptide surface that mostly contribute in binding energy, which will be useful to fast-scan a peptide across those regions and identify candidate sites of interaction.

5.1.2 Problem Definition

The problem of studying peptide-protein interactions and their identification can be defined in the following two following ways:

Coarse-grain Definition: Given one protein with peptide-recognition domain and a peptide, the problem is to identify whether they will interact or not, shown in **Fig 42**.

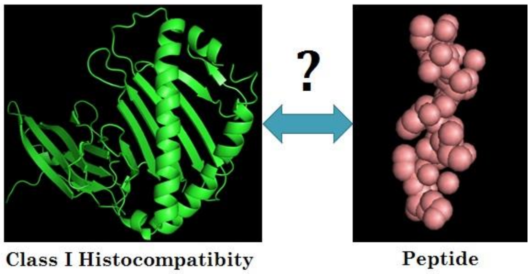


Fig 42. Coarse-grained view of the underlying problem. For example, given a MHC molecule (*green*) and a peptide (*tint*), it is to predict whether they will interact and form a complex or not.

Fine-grain Definition: Given one protein with peptide-recognition domain and a peptide, the problem is to identify the peptide-binding residues or regions in the protein that recognize the peptide. Moreover, characterize the hot spots on peptide surface that primarily contribute in the energy needed for binding. A sample fine-grain problem definition is shown in **Fig 43**. In our study, we aim to develop computational tools to solve problems under this fine-grain definition.

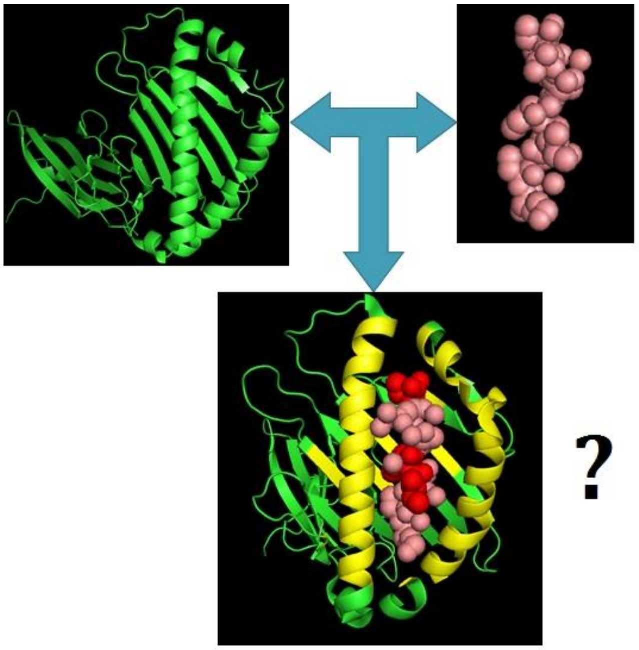


Fig 43. Fine-grained view of the underlying problem. For example, given a MHC molecule （*green*） and a peptide （*tint*）, it is to predict the peptide-binding residues （*yellow*） in MHC molecule and residues on peptide surface （*red*） those contribute higher binding energy.

5.1.3 Literature Review

An adequate literature exists that investigates the underlying strategies behind peptide-protein binding [321, 339], exploring how peptides can recover the entropic loss and achieve enthalpy gain involved in the process of binding. A rigorous study in [321] with known peptide-protein complexes shows that peptides usually bind to the largest pocket on the protein surface, and result in more packed interface than that of protein-protein interface. Moreover, the presence of 'hot spot' residues that make the major contribution of the energy in binding, has been conceptualized at the interface of both protein-protein [340-342] and peptide-protein complexes [321].

Discoveries of new peptide-protein interactions are challenging [326, 343]. Attempts have been made to predict PDZ domain-peptide interactions from protein sequence [344], analyze and predict interactions of SH3 domain [345, 346], predict SH2 domain interactions in a genome-wide scale using structure information [347]. Some other computational predictions of peptide-mediated interactions include structure-based modeling of binding specificity [348] to identify farnesylation and identification of interactions with Bcl-2 proteins [349]. The database of the eukaryotic linear motif (ELM) [350] provides consensus sequence patterns for peptide motifs that bind to many different PRDs. The study of available structures of protein-peptide complexes in the PDB have also identified potential peptide-protein interactions [351]. However, the experimental or computational efforts are focused on limited range of PRDs, therefore the methods that enable predictions for a larger number of PRD families, are needed. Recently, a computational framework is developed to predict peptide-protein interactions using a Bayesian approach that integrates knowledge from the ELM database, domain-peptide structures from the PDB, and non-structural information [352].

While previously discussed studies are focus on identifying interactions, attempts have been made to identify the potential peptide-binding sites as well. Computational tools to predict protein-peptide binding regions from structures are Pepsite [353], Peptimap [354], PepBind [355]. However, accurate prediction of binding regions from sequence only has further implications as it can be applied in proteome-scale to assemble potential interactome. Despite much progress, the sequence-based computational efforts have been taken to predict a few PRDs, *i.e.*, MHC molecules [356, 357]. To the best of our knowledge, there exists only one scholarly article in the literature called SPRINT [358] that predicts peptide-binding sites on few PRDs, *i.e.*, MHC, PDZ, SH2, and SH3 from sequence, but often overpredicts.

5.1.4 Our Contributions

In this study, we develop a new computational tool to predict peptide-binding residues, named PBRpredict, of proteins with peptide-recognition domain from protein sequence alone. We collected a new dataset of peptide-protein complexes with a large variety of domains, like MHC I and II, PDZ, SH2, SH3, WW, 14-3-3, Chromo and Bromo, Polo-Box, PTB, enzyme inhibitor, from Protein Data Bank (PDB) [61]. A set of partner (receptor) proteins was generated from the complexes, and the protein chains were annotated with interaction information based on the atomic distances from peptide residues in the structure.

Using a comprehensive set of sequence-based features including residual profile, chemical and evolutionary profile, secondary structure and local backbone profile, surface area and an energy based profile, we guided our predictor to learn about the peptide-binding regions. This work investigates ‘Model Stacking’ [359], an effective machine-learning technique, in this challenging application of proteomics that requires the ability to capture the atomic interaction level feature of protein molecule form its sequence alone. Furthermore, we develop two complementary versions of the initial model by tuning the classification thresholds, keeping the other parameters and overall framework the same, to improve the model's capacity to recognize potential binding sites. The final three models are called *PBRpredict-strict*, *PBRpredict-moderate* and *PBRpredict-flexible,* which are combined in the *PBRpredict-Suite*.

The competitive performances of PBRpredict-Suite models support the strength of our predictor framework. When compared with the current state-of-the-art method, the proposed models showed a reasonable, well-balanced and biologically relevant performance. We analyzed the competence of the strict, moderate and flexible PBRpredict models on different case-studies, *i.e*., structure-specific sequence with known and unknown domains from PDB, and full-length sequence with unknown domain from UniProt. The outputs validate the usefulness of the 3 models in the PBRpredict-Suite in different cases. Thus, PBRpredict will have further implication in solving challenging problems of computational biology, like binding affinity prediction, hot spot regions and residue prediction, and peptide binding site prediction.

Furthermore, we extracted an energy score, position specific binding energy (PSBE) from sequence only to characterize the amino acids that are prevalent in the hot spots of peptide surface. The hot spot residues are known to have major contribution in the binding energy. The PSBE was found effective in identifying the amino acid composition that are likely to be hot spots.

5.2 Experimental Materials

In this section, we describe the definition of peptide-binding residues and regions, data collection process, peptide-binding domains in the dataset, aggregation of input features, and the criteria to evaluate and compare the peptide-binding residue prediction tasks.

5.2.1 Datasets

In this study, our focus is to capture the residue-patterns of different peptide-recognition domains (PRDs) from protein sequence alone. Therefore, we intended to collect a set of globular protein receptors that were experimentally found to bind with short peptide chains (5 to 25 residues long) in a complex. The residues of these receptor proteins (or, partner proteins), which were involved in peptide-binding, were then annotated as binding ('b') or non-binding ('n'). For our experiments, we explored PDB [61], accessed on September 2016, to assemble a set of peptide-protein complex structures using the following criteria:

(*i*) Experimental method, x-ray crystallography;

(*ii*) Molecule type, protein (no DNA, RNA or hybrid);

(*iii*) Number of chains (both asymmetric unit and biological assembly), greater than or equal to 2;

(*iv*) Structures that contain at least one 5 to 25-residue long chain;

Our initial search with above criteria resulted in 6,043 protein complexes which contain total 25,557 chains. We filtered the set to remove complexes that have one or more subunit chains with unknown amino acid residues, 'X' or 'Z', because the necessary chemical features [174] are not available for these residues. Moreover, a multimeric protein (homomeric and heteromeric) can contain multiple entries of identical chains. In such cases, we kept only one unique copy of a chain that maximizes the number of peptide-binding residues. In the feature generation steps, we used SPINE-X [175] to generate predicted values of the two backbone angles, phi and psi. We removed those chains for which SPINE-X failed to produce the required features. In the final step, we clustered the remaining sequences at sequence identity below 40%. From each cluster, a representative sequence with maximum peptide-binding residues was chosen in the non-redundant dataset of 644 protein receptors, named as *rcp644*, available within the software package.

The rcp644 dataset contains 98 chains (around 15%) of length ≤ 25 whereas 546 chains are longer with > 25 residues. Out of 116,489 number of residues, around 17% were binding residues (positive class) while the rest of the 83% served as the negative samples.

5.2.1.1 *Peptide-Recognition Domains in the Dataset*

A wide range of peptide-recognition domains (PRDs) were included in our collection of receptor sequences with peptide-binding residues or regions that mediate peptide-protein interactions [326], listed in the following:

(*i*) **M**ajor **H**istocompatibility **C**omplex (**MHC I** and **II**) domain that recognizes peptide fragments derived from pathogen of length 8 – 12 residue [360].

(*ii*) **PDZ** domain, generally binds to short peptide motifs at C-terminal of other proteins [361].

(*iii*) **S**rc **H**omology **2** (**SH2**) domain and **P**hospho-**T**yrosine **B**inding (**PTB**) domains that recognize phosphorylation of tyrosine (pTyr or pY), such as SH2 binds to a core motif pY-X-X-P/L [362, 363]. Moreover, PTB domain can bind to motif, *i.e*., N-P-X-Y.

(*iv*) **S**rc **H**omology **3** (**SH3**) domain, binds to Pro-rich peptides [364], peptide motifs such as R-X-X-K [365] and also to the surface of ubiquitin [366].

(*v*) **14-3-3**, **WW**, **Polo-box**, **BR**CA1 **C** **T**erminus (**BRCT**), **F**ork**H**ead-**A**ssociated (FHA) domain that recognize different type phosphorylation or post-translational modifications (PTMs) of threonine (pThr or pT) and serine (pSer or pS) [367-370]. Further, WW domain binds to pro-rich motifs.

(*vii*) **Chr**omatin **o**rganization **mo**difier (**Chromo**), **Bromodomain** and **Tudor** domain that bind to methylated or acetylated peptides, such as Tudor domain can recognize PTMs on lysine (meLys or meK) and arginine (meArg or meR) by methylation. Chromo domain can also recognize meLys and Bromo domain recognize PTMs on lysine by acetylation (acLys or acK).

(*viii*) Enzyme/inhibitor complexes with hydrolase, kinase, isomerase, phosphatase, protease.

(*ix*) Complexes with antibody-antigen, amyloid fibrils, membrane or transmembrane proteins.

(*ix*) Nuclear receptor complexes and others.

Therefore, our dataset rcp644 captures the sequence-patterns of a wide range of peptide-binding domains and their interactions with short peptides. The set contains 98 chains (around 15%) of length less than or equal to 25 whereas 546 chains are longer having greater than 25 residues. Out of 116,489 residues, around 17% were peptide-binding residues (positive class) while the rest of the 83% residues served as the negative samples in training.

For each class of peptide-recognition domains (PRDs), 25% sequences were accumulated in the independent test dataset and the rest were gather to form the training set. Further, the training set were divided into two folds, in which we carefully included 50% sequences of each types of PRDs. This distribution allows training and evaluating the predictor with information of all the domains. **Table 25** lists up the counts of different PRDs in the full dataset (644 chains), full training set (475 chains) and two different folds of the training set (243 and 232 chains), and the test set (169 chains).

5.2.1.2 *Training Set*

The training set is composed of 475 receptor protein chains, named rcp_tr475. It contains 400 relatively longer chains (> 25 residues) and 75 shorter chains (≤ 25 residues). The 475 chains consist of 89,512 residues of which 16.5% (14,748) were peptide binding and rests (74,764) were non-binding residues. The count of different PRDs included in the training set along with its two different folds are listed in **Table 25**.

5.2.1.3 *Test Set*

The independent test set contains 169 chains with different PRDs (**Table 25**), called as rcp_ts169. The test set has 146 long chains and 23 short chains. Moreover, it has total of 26,977 residues of which 5,162 residues are peptide-binding residues and the rest of the 21,815 are non-binding residues.

Table 25. Name and count of the peptide-recognition domains included in the datasets.

Peptide Recognition Domains (PRDs)	Full Set	Full Training Set (Fold 1 + Fold 2)	Test Set
MHC I/II	81	60 (30 + 30)	21
PDZ	37	27 (14 + 13)	10
SH2	49	36 (18 +18)	13
SH3	54	40 (20 + 20)	14
14-3-3	35	26 (13 + 13)	9
WW	7	5 (3 + 2)	2
Polo-Box	10	7 (4 + 3)	3
Tudor	15	11 (6 + 5)	4
PTB	10	7 (4 + 3)	3
Chromo	12	9 (5 + 4)	3
Bromo	27	20 (10 + 10)	7
BRCT	13	9 (5 + 4)	4
FHA	7	5 (3 + 2)	2
Enzyme/Inhibitor (hydrolase/kinase/isomerase /phosphatase/protease)	109	81 (41 + 40)	28
Antibody/antigen/FAB	25	18 (9 + 9)	7
Membrane/Transmembrane	27	20 (10 + 10)	7
Amyloid	36	27 (14 + 13)	9
Nuclear	20	15 (8 + 7)	5
others	70	52 (26 + 26)	18
Total	644	475 (243 + 232)	169

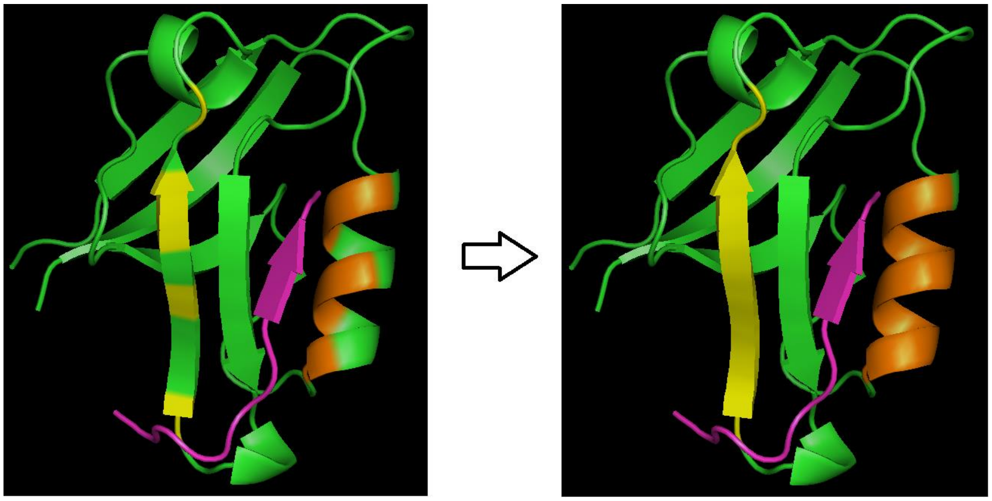
5.2.2 Annotation of Peptide-Binding Residues and Regions

A putative interaction between two amino acids is determined based on their atomic distances. Specifically, we annotated an amino acid as peptide-binding residue if at least one of its heavy atoms stays within 6Å distance from a heavy atom of a peptide residue [129]. Therefore, we did not consider hydrogen atom while determining interaction between amino acids. Further, we did not consider any interactions with two adjacent amino acids on either side of a residue to skip the covalent bonded interactions, and store only the transient interactions which is relevant in coupled-binding within peptide-protein complex [322].

After annotating the residues as either peptide-binding ('b') or non-binding ('n'), we applied a smoothing strategy to have regions of binding residues. We smoothed-out maximum 3-residue long non-interacting regions that fall within two consecutive interacting regions or residues. Therefore, we say that the resulting regions are the 'potential' areas that contain the residues of interaction. We call such labeling as *synthetic annotation*, which was assigned on top of the *actual annotation*. **Fig 44** shows a sample synthetic annotation of a chain with PDZ domain (PDB ID: 4JOE [371]).

...GISITGGKEHGVPILISEIHPGQPAD...NLRDTKHKEAVTILSQQRG...
...GISITGGKEHGVPILISEIHPGQPAD...NLRDTKHKEAVTILSQQRG...
...GISITGGKEHGVPILISEIHPGQPAD...NLRDTKHKEAVTILSQQRG...
...nnnnnnnnnnnbbbbbbbbbnnnnnn...nnnnnnnbbbbbbbbbnnnn...

(a) Actual and Synthetic Annotations on the Sequence



(b) Actual and Synthetic Annotations on the Structure

Fig 44. Actual and synthetic annotation of a PDZ domain (green) bound to peptide (pink) (PDB ID 4JOE [371]). (a) The actual (*2nd line from top)* and synthetic annotations (*3rd line from top*) of peptide-binding residues are shown on the protein's primary sequence (*1st line*). Two potential regions of interactions, residues: 28 – 36 and residues: 66 – 74) are highlighted in *yellow* and *orange* respectively. We used two different colors to highlight two different structural regions (helix and beta). The last line shows the annotated sequence with labels ('b' for peptide-binding and 'n' for non-binding). **(b)** The annotations are mapped onto the structure of the same protein (*green*) bound to a peptide (*pink*). The two highlighted regions in (a) are marked on the structure in *yellow* (beta region) and *orange* (helix region), respectively. Before smoothing, the binding residues were disjoint (*left*), whereas in synthetic annotation (*right*), the binding residues are contiguous. We viewed the 3D structures using PyMOL[50] and the secondary structure was assigned using DSSP [51].

The rationale behind generating such synthetic annotation is: we have disjoint residues of interaction with non-interacting residues in between due to the geometrical orientation of the side chain atoms. Notwithstanding, it is hard to capture these 3D structural details from 1D primary sequence alone and subsequently guide a machine learning algorithm. To reduce the complexity, we localized the binding residues in a region so that the prediction algorithm can be better informed about their characteristics from the sequential environment. In this way, we have less chance of missing a binding residue as the contiguous residues can reinforce the residue-level as well as region-level prediction.

5.2.3 Feature based Sequence Representation

We encoded the residues of the primary protein sequence using 60 features (f_1 – f_{60}) of 6 categories per residue to characterize the peptide-binding properties as described below.

Residue profile (f_1, f_{60}): The residue profile was created with two information: the amino acid (AA) type and the terminal (t) region indicator. Twenty different types of amino acids were encoded using 20 different numbers. Thus, AA contributes 1 feature per residue, which is useful to capture the amino acid compositions of peptide-recognition domains that primarily contribute to peptide-protein binding. Further, to distinguish the terminal residues that show higher tendency of binding than the central ones, we encoded five residues of N-terminal as (−1.0, −0.8, −0.6, −0.4, −0.2) and C-terminal as (+1.0, +0.8, +0.6, +0.4, +0.2), whereas rest of the residues were labeled as 0.0

Chemical profile (f_2–f_8): Seven physicochemical properties (PP) of amino acids, namely steric parameter, normalized van der Waals volume, hydrophobicity, isoelectric point, helix and sheet probabilities were collected [174], and fed as features to capture the chemical description of the protein residues that can transiently interact with peptides [372].

Conservation profile (f_9–f_{28}, f_{37}–f_{57}): Peptide-recognition domains can be conserved [373] and also undergo divergence for functional adaptation [326]. Sequence alignment based conservation score was extracted from Position Specific Scoring Matrix (PSSM). We executed three iterations of PSI-BLAST [173] against NCBI’s non-redundant database to generate PSSM of size sequence length × 20, which gave us 20 features per residue. These 1D scores given by PSSM were further extended to higher dimension by computing monogram (MG, 1 feature) and bigram (BG, 20 features) which are found to be effective in protein fold recognition [176, 374]. We used PSSMs, MG and BGs as conservation profile to predict peptide-binding residues.

Structural profile (f_{29}–f_{34}): We used six predicted structural properties; 3 secondary structure (SS) probabilities, specifically helix (H), beta (B) and coil (C), 2 backbone angles, phi (ϕ) and psi (ψ), and 1 solvent Accessible Surface Area (ASA) to construct the structural profile. The SS profile was predicted using a meta predictor, MetaSSpred [196] that gives balanced predictions of all three classes. The ASA and backbone angles were respectively predicted using tools, REGAd^{3}p [15] and SPINE-X [175].

Flexibility profile (f_{35}–f_{36}, f_{58}): We created a flexibility profile with 2 backbone angle fluctuations, such as dphi ($\nabla\phi$) and dpsi ($\nabla\psi$) and 1 disorder probability (drp). The backbone angle fluctuations were predicted using DAVAR [179] and the probability of a residue being disordered was predicted using DisPredict [10]. These features are useful to capture the pattern of conformational changes that may result from coupled-binding between a short peptide and a globular receptor [321].

Energy value (f_{59}): The transient bonds between peptide and receptor involve formation and dissolution of atomic interactions as well as structural changes that require change in free energy [321]. To capture the state of free energy contribution of the residues in peptide-protein interaction, we computed per-residue Position Specific Estimated Energy (PSEE) [11] from sequence. PSEE is a recently introduced concept in the literature, discussed in Chapter 4, that estimates free energy contribution of the residues from sequence by modeling pairwise contact energy and predicted solvent accessibility. Here, the pairwise interaction energy captures the sequential environment around the residue whereas the solvent accessibility captures the residue's state in the respective 3D structure.

We further computed the relative importance of these 60 features in predicting the peptide-binding residue by computing the Gini importance with extra tree classifier [375]. The output of this experiment suggested that all these features are useful (*see* **Section 5.4.1**). Thus, we used all 60 residue-wise features. Finally, we applied a sliding window of size 25 centering the target residue to include the properties of 12 residues on either side of the target, describing the local environment. Thus, we fed 60 × 25 = 1,500 features per residue to train the predictor model. The window size was selected through a separate set of experiments and the results are reported in **Section 5.4.2**.

5.2.4 Evaluation Criteria

The binary classification output is evaluated and compared using the measures listed in **Table 26**. Here, peptide-binding residues belong to the positive class and non-binding residues belong to the negative class. Recall is the measure to identify a predictor's completeness in classifying the positive class (peptide-binding residues), precision measures the predictor's exactness. Therefore, the harmonic mean of recall and precision called F1 score measure a classifier's overall correctness. The miss rate and fall-out rate measure two complementary types of incorrect predictions, respectively the misclassification of binding residue as non-binding and non-binding residue as binding. MCC is considered as another balanced measure to evaluate binary classification.

Table 26. Name and definition of performance measures to evaluate peptide-binding residue prediction.

Name of metric	Definition
True positive (TP)	Number of correctly predicted peptide-binding residues
True negative (TN)	Number of correctly predicted non-binding residues
False positive (FP)	Number of incorrectly predicted peptide-binding residues
False negative (FN)	Number of incorrectly predicted non-binding residues
Recall/Sensitivity	$True\ Positive\ Rate\ (TPR) = \frac{TP}{TP + FN}$
Specificity	$True\ Negative\ Rate\ (TNR) = \frac{TN}{FP + TN}$
Fall-out (or over prediction) Rate	$False\ Positive\ Rate\ (FPR) = \frac{FP}{FP + TN}$
Miss Rate	$False\ Negative\ Rate\ (FNR) = \frac{FN}{FN + TP}$
Balanced accuracy (Mean of Specificity and Recall)	$\frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right)$
Precision	$\frac{TP}{TP + FP}$
F1 Score (Harmonic mean of precision and Recall)	$\frac{2TP}{2TP + FP + FN}$
Mathews correlation coefficient (MCC)	$\frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$

Moreover, Area under ROC curve (AUC or AUROC) is considered as the measure for probability assignment. We further plotted the ROC curves and Precision-Recall curves. The AUC values and the curves are generated using pROC [376] and ROCR packages [312] in R.

5.3 PBRpredict Framework

We applied stacked generalization [377] to develop the peptide-binding residue predictor (PBRpredict). Stacking is an ensemble technique to minimize the generalization error and has been successfully applied in several machine learning tasks [378-380]. To the best of our knowledge, this study has first explored stacking for identifying the pattern of protein sequence that induces binding with peptides.

Stacking framework involves two-tier learning. The classifiers of the first tier and the second tier are called *base-learner* and *meta-learner* respectively. Multiple base-learners are employed in the first tier. In the second tier, the outputs of the base-level learners are combined using another meta-level learner. Here, the underlying idea is that different base-learners can incorrectly learn different regions of the feature space, effectively due to the no-free-lunch theorem [381]. A meta-learner is then applied, usually non-linearly, to accumulate the outcomes of the first-tier learners that are better-trained for differ feature-space regions. Therefore, it is desirable to choose classifiers that can generate uncorrelated prediction outputs for the base-level training.

The second tier, that combines the outputs from the first tier, makes the stacking technique different from other ensemble methods like, bagging and boosting as those techniques apply a particular cost function such as, weighted average or majority vote to combine the existing outputs from the first tier. On the other hand, stacking employs another classifier to learn from the outputs from the first tier and generates the final prediction.

5.3.1 Learning Algorithms

We explored six different machine leaning algorithms as base-learners and used logistic regression as meta-learner to combine probability distributions generated at the base-level. The included algorithms are briefly discussed below.

Support Vector Machine (SVM): We used radial basis function (RBF) kernel based support vector machine (SVM) [111] as one of the base-learners. SVM is an effective algorithm for binary prediction in high-dimensional space that minimizes both the empirical classification error in the training phase and generalized error in the test phase. SVM classifies by maximizing the separating hyperplane between two classes and penalizes the feature space points on the wrong side of the decision boundary using a cost parameter, c.

The parameter of RBF kernel, γ and the cost parameter, c were optimized to achieve best accuracy using time-intensive grid search along with 5-fold cross validation step. We conducted this parameter optimization using a subset of 60% samples from the training set, specifically 386 chains out of 644 chains of the dataset. The found best values of the parameters are, $\gamma = 2^{-7}$ and $c = 2^{3}$, and those are used as representative parameter values for training with the full dataset. The optimal setup of the penalty parameter (c) and RBF parameter (γ) make the SVM model effective for classification problem with imbalanced dataset and high-dimensional feature space, such as, problem that is attempted under this work. The generation of SVM model and parameter optimization is done using libSVM [184] package.

Random Decision Forest: The random decision forest (RDF) [382, 383] is an ensemble algorithm which is used to generate a base-learner in this study. RDF operates by constructing a multitude of decision trees on various sub-samples of the dataset and outputting the mean prediction of the decision trees. Therefore, the trees of RDF work on the subspaces of the full data. We used bootstrap samples to construct 1,000 tress in the forest to develop the RDF learner using scikit-learn [384].

Extra Tree Classifier: The extremely randomized tree or extra-tree classifier (ET) [385] is another class of ensemble methods and explored as a base-learner in this work. ET works by constructing randomized decision trees from the original learning sample. The best split is determined randomly from the range of values at each split. We used scikit-learn [384] to construct the ET model with 1,000 tress and the quality of a split was measured by Gini impurity index.

Gradient Boosting Classifier: Another machine learning technique that we used to develop a base-learner is the gradient boosting [386]. The gradient boosting classifier (GBC) works by combining weak learners into a single learner in an iterative fashion. Using scikit-learn [384], we applied 1,000 boosting stages where a regression tree was fit on the negative gradient of the deviance loos function. The learning rate was set to 0.1 and the maximum depth of each regression tree was set to 3. GBC gives robust performance to over-fitting with higher number of boosting stages, and we observed that 1,000 stages were giving competitive performance for this application.

K Nearest Neighbors: The *k* nearest neighbors (KNN) classifier [387] operates by learning from the *k* closest training samples in the feature space around a target point. The classification decision is produced based on the majority votes coming from the neighbors. In this work, the value of k was set to 9 and all the neighbors were weighted uniformly for generating the KNN model using scikit-learn [384].

Bagging Classifier: The bootstrap aggregation or bagging (BAG) [388] is another ensemble method that is particularly useful for reducing variance in the prediction. We developed bagging classifier model using scikit-learn [384] that essentially fits multiple subsets of data with repetitions on 1,000 decision trees, and combines output by weighted averaging.

Logistic Regression: To develop the meta-learner that combines the output probabilities generated by the base-learners, we used logistic regression (LogReg) [389] with L2 regularization. The LogReg classifier estimates the probability of interacting versus non-interacting residues based on the confidence or probability distributions produced by multiple independent base-learners.

5.3.2 Training and Test of Base Learners

The six potential classifiers (SVM, RDF, ET, GBC, KNN and BAG) to be used as base-learners were trained on the full rcp_tr475 dataset using M = 60 × 25 features. These models were then used to predict peptide-binding residues in 169 chains of the test set (rcp_ts169) and evaluated using statistical measures. Let us assume, N_{train} is the total number of residues of 475 chains in the training set (rcp_tr475). Then the size of the feture matrix used for training was $N_{train} \times M$. Guided by the performance of these six algorithms (*see* **Section 5.4.3**), we finally employed SVM, GBC and KNN in the base-level of the stacking performed to develop the predictor-models under PBRpredict-Suite (**Fig 46**). These base-level models are denoted by $MODEL_{SVM}$, $MODEL_{GBC}$ and $MODEL_{KNN}$, and the per-residue feature vector is $\boldsymbol{X'} = (f'_1, f'_2, \dots, f'_{60\times25})$.

5.3.3 Training and Test of Meta Learner

We created the feature matrix of N_{train} residues to train the meta-learner through blending, shown in **Fig 45**. For this, we divided the train set of 475 chains into two folds of 243 and 232 chains (**Table 25**) so that $N_{train} = N_{fold1} + N_{fold2}$. Here, N_{fold1} and N_{fold2} are the total number of residues in 243 and 232 chains, respectively. Further, we found that the inclusion of 60 features (discussed in **Section 5.2.3**) of the target residue in addition to the three probabilities generated by the three base-learners makes the meta-learner more accurate (discussed in **Section 5.4.4.2**). Therefore, the number of features used to train the meta-learner was 63.

At first, N_{fold1} number of residues with M number of features were used to develop the three base models, $MODEL_{SVM}^{fold1}$, $MODEL_{GBC}^{fold1}$ and $MODEL_{KNN}^{fold1}$, which were used to predict the N_{fold2} number of residues (*see* **Fig 45**). Conversely, N_{fold2} number of residues with M number of features were used to develop another set of base models, $MODEL_{SVM}^{fold2}$, $MODEL_{GBC}^{fold2}$ and $MODEL_{KNN}^{fold2}$, and the predicted probability values for N_{fold1} number of residues were generated using these models. Thereafter, the independently predicted probabilities of N_{fold1} residues (in 243 chains) and N_{fold2} residues (in 232 chains) were combined to generate the feature matrix of size $N_{train}\times$ 63 to train the LogReg.

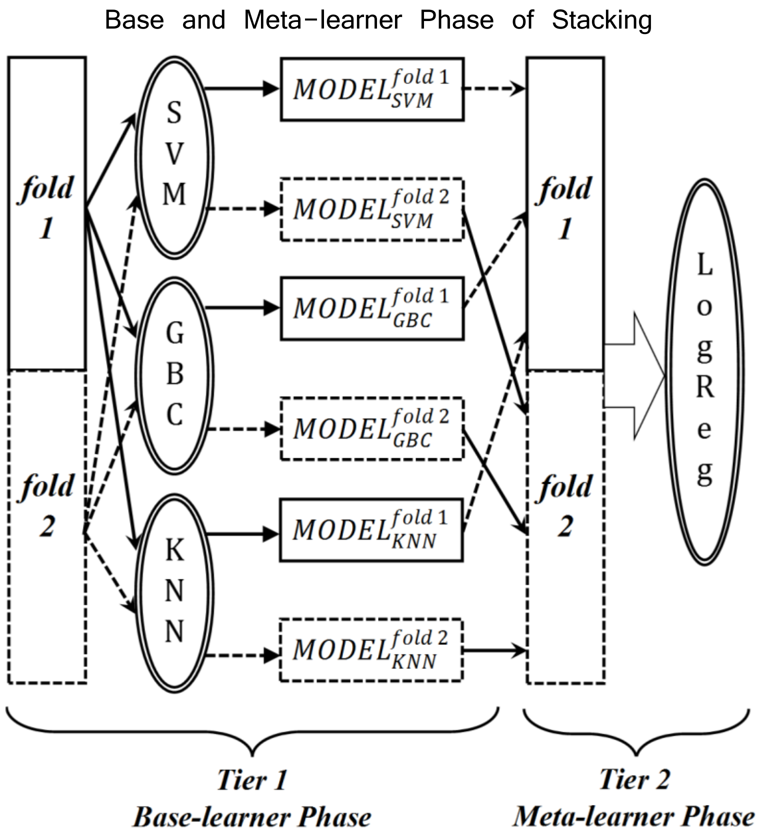

Base and Meta-learner Phase of Stacking

Fig 45. Two-tier training and validation in stacking. Blending of SVM, GBC and KNN to generate independent prediction outputs on two different folds of the full training set. These outputs are then used as training features for the meta-level LogReg classifier. The objects and arrows associated with fold1 and fold2 are indicated by solid line and dashed line, respectively.

To test the meta learner, we predicted 169 chains from the test set using $MODEL_{SVM}$, $MODEL_{GBC}$ and $MODEL_{KNN}$, which were trained using full training set but the test set. With the three output probabilities and the 60 features for the residues, we performed the meta-level predictions on these chains.

5.3.4 PBRpredict-Suite

PBRpredict-Suite is a collection of 3 peptide-binding residue predictor-models, namely PBRpredict-strict, PBRpredict-moderate and PBRpredict-flexible. The models are named according to their behavior in predicting the peptide-binding residues which is primarily determined by the classification thresholds used by the base-level and meta-level learners (**Section 5.4.5**). However, the learning algorithms and feature set combination used in both the levels of stacking were kept same for all three models in the suite.

Let us denote the set of thresholds used by SVM, GBC, KNN and LogReg to convert the probability outputs (or confidence score) into binary outputs as $(t_{SVM},\ t_{GBC}, t_{KNN}, t_{LogReg})$. With that the definition of the three models of PBRpredict-Suite are given below.

- **PBRpredict-strict**: The traditional value of 0.5 is used as thresholds by all the learners. Thus, $\left(t_{SVM},\ t_{GBC}, t_{KNN}, t_{LogReg}\right) = (0.5, 0.5, 0.5, 0.5)$. The three original probability values generated by SVM, GBC and KNN $(p_{SVM},\ p_{GBC}, p_{KNN})$ are used as features to train the LogReg. Later, the LogReg produces the final prediction output using 0.5 as threshold.
- **PBRpredict-moderate**: Here, we apply a moderate set of values as thresholds, $\left(t_{SVM},\ t_{GBC}, t_{KNN}, t_{LogReg}\right) = (0.3, 0.34, 0.35, 0.3)$. Thus, the probabilities of the predicted positive class (peptide-binding residue) can range from the modified threshold value to 1.0, *e.g.*, [0.3, 1.0] for SVM. Consequently, the range of the predicted negative class (non-binding residue) is [0.0, 0.3) for SVM. The original probabilities given by the base-learners are then scaled to [0.5, 1.0] for the positive class and to [0.0, 0.5) for the negative class from the corresponding ranges for different base-learners defined by the new thresholds. These three modified probability values for SVM, GBC and KNN $(p_{SVM}^{m},\ p_{GBC}^{m}, p_{KNN}^{m})$ are the used as features to train the LogReg, which produces the binary prediction output using 0.3 as threshold. Finally, the probabilities given by LogReg are scaled to the traditional range ([0.0 – 0.5) for the negative class and [0.5 – 1.0] for the positive class) from the one defined by the changed threshold ([0.0 – 0.3) for the negative class and [0.3 – 1.0] for the positive class).
- **PBRpredict-flexible**: In this model, the classification thresholds for all the learners are further loosened, $\left(t_{SVM},\ t_{GBC}, t_{KNN}, t_{LogReg}\right) = (0.17, 0.21, 0.21, 0.2)$. Like the PBRpredict-moderate, the output probabilities of the base-learners are scaled to modified range defined by these new thresholds and then used as features for training the meta-learner. Here, the LogReg produces the binary prediction output using 0.2 as threshold and the probabilities given by LogReg are scaled to the traditional range ([0.0 – 0.5) for the negative class and [0.5 – 1.0] for the positive class) from the one defined the changed threshold ([0.0 – 0.2) for the negative class and [0.2 – 1.0] for the positive class).

The framework of the PBRpredict-Suite is illustrated in **Fig 46**. The different threshold values for the learners of the moderate and flexible models of the suite were statistically chosen (*see* **Section 5.4.5**) to correct certain percentage of the false negative prediction outputs of the strict model. Altogether, these 3 models performed promisingly in different cases (*see* **Section 5.4**).

The PBRpredict-Suite with all 3 models is implemented as a single software package and available online[27]. The software is developed using C, Python and shell scripting, and tested on Linux platform. It includes a ReadMe file that lists up the external dependencies and guidelines to run the tool. The software outputs per-residue binary annotation and real-value probability given by 3 different models. It also generates a summary file that reports the peptide-binding tendency per-chain averaged over the predicted peptide-binding residues and all residues.

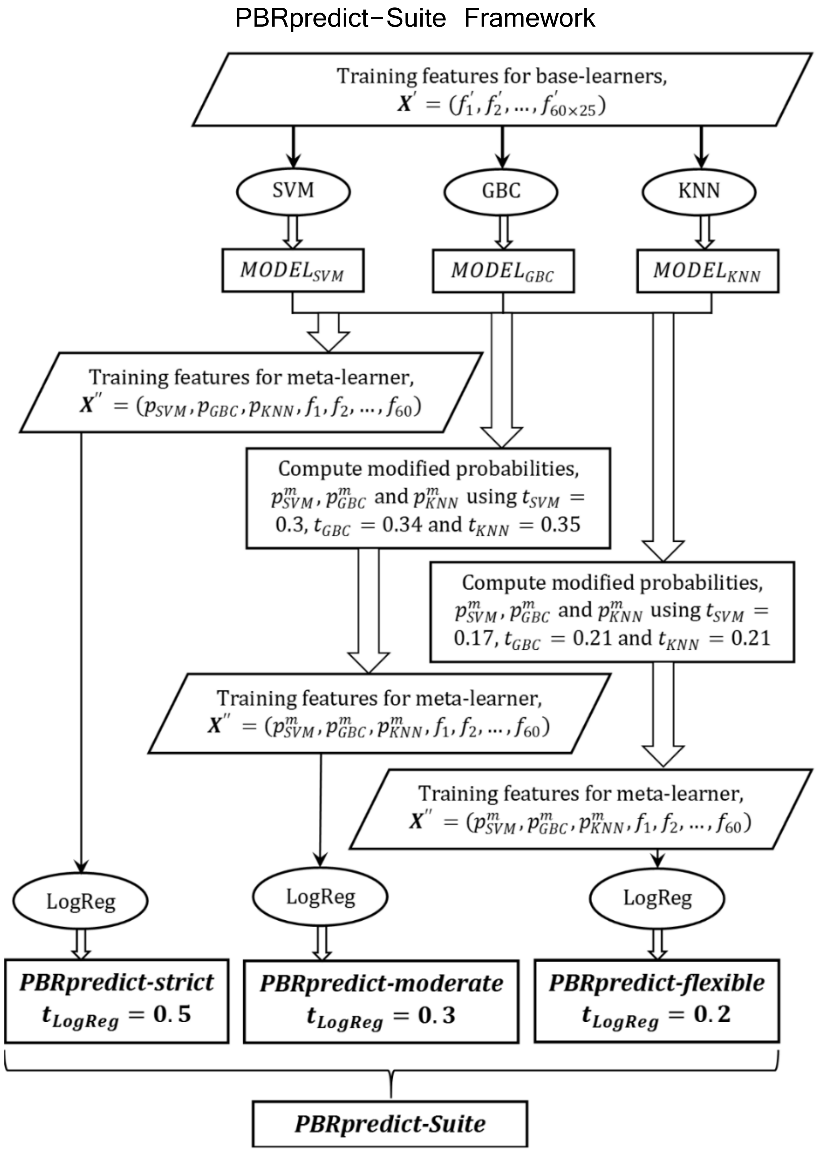

PBRpredict-Suite Framework
Training features for base-learners, $X' = (f_1', f_2', \ldots, f_{60\times25}')$
SVM
GBC
KNN
$MODEL_{SVM}$
$MODEL_{GBC}$
$MODEL_{KNN}$
Training features for meta-learner, $X'' = (p_{SVM}, p_{GBC}, p_{KNN}, f_1, f_2, \ldots, f_{60})$
Compute modified probabilities, p_{SVM}^m, p_{GBC}^m and p_{KNN}^m using $t_{SVM} = 0.3$, $t_{GBC} = 0.34$ and $t_{KNN} = 0.35$
Compute modified probabilities, p_{SVM}^m, p_{GBC}^m and p_{KNN}^m using $t_{SVM} = 0.17$, $t_{GBC} = 0.21$ and $t_{KNN} = 0.21$
Training features for meta-learner, $X'' = (p_{SVM}^m, p_{GBC}^m, p_{KNN}^m, f_1, f_2, \ldots, f_{60})$
Training features for meta-learner, $X'' = (p_{SVM}^m, p_{GBC}^m, p_{KNN}^m, f_1, f_2, \ldots, f_{60})$
LogReg
LogReg
LogReg
PBRpredict-strict $t_{LogReg} = 0.5$
PBRpredict-moderate $t_{LogReg} = 0.3$
PBRpredict-flexible $t_{LogReg} = 0.2$
PBRpredict-Suite

Fig 46. The workflow of BIRpredict-Suite framework including BIRpredict-strict, BIRpredict-moderate and BIRpredict-flexible. The symbols and abbreviations used are explained in Section 5.3.

[27] PBRpredict-Suite: http://cs.uno.edu/~tamjid/Software/PBRpredict/pbrpredictSuite.zip

5.3.5 Implementation and Availability

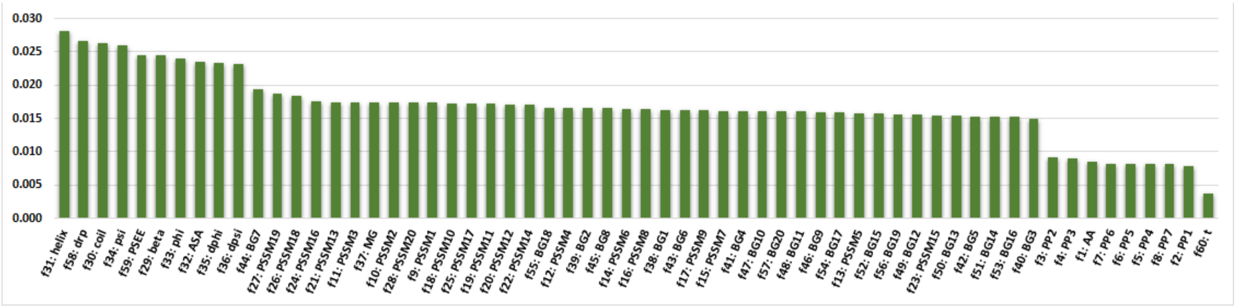
We implemented the PBRpredict-Suite using languages, like C, Python and shell scripting. The software is developed and tested on Linux platform. The software is available online[28] with a user manual.

5.4 Results

In this section, we report the results of each step of the development of the PBRpredict-Suite including the feature selection, window selection, base-learner selection, and tuning of thresholds to build strict, moderate and flexible models. The performance comparison among the PBRpredict-Suite models and a state-of-the-art predictor is discussed as well. We further analyze the biological significance of the predictor-models on multiple case studies with known and unknown peptide-recognition domains.

5.4.1 Feature Selection

Here, we report the results of the feature importance estimation using extra tree (ET) classifier. ET estimates the feature importance using a method described by Brieman [375] by maintaining impurity reduction for each feature [385, 390]. The information gain is attributed to each feature to measure total decrease of impurity. Finally, the classifier provides an importance value for each feature, known as Gini importance, which was used to rank the features.



Estimated Feature Importance

Fig 47. Feature importance estimation by ET classifier in peptide-binding residue prediction. The importance values are shown respectively by green bar. The *x*-axis shows the features in their abbreviated form according to Section 5.2.3. Multiple features of same category are indexed by their count, i.e., 20 PSSMs are indexed from 1 to 20.

[28] PBRpredict-Suite: http://cs.uno.edu/~tamjid/Software/PBRpredict/pbrpredictSuite.zip

Fig 47 presents the ranked features according to the importance values. The importance values can be interpreted as the fraction of the test samples that were correctly classified by that feature. The training and testing were done using rcp_tr475 and rcp_ts169 datasets with 60 features (window size 1). **Fig 47** shows that all the features have greater than zero importance, thus we used all 60 features to develop our predictor.

Fig 47 shows that the structural profile (f_{29}: beta, f_{30}: coil, f_{31}: helix, f_{32}: ASA, f_{33}: phi and f_{34}: psi), the flexibility profile (f_{35}: dphi, f_{36}: dpsi and f_{58}: drp) and the energy profile (f_{59}: PSEE) are the three most dominant feature categories. To understand the contribution of the dominant features, we further developed separate ET models by subsequently removing the six structural properties, three flexibility-related properties and one energy-based property from the feature set. Therefore, these ET models were developed based on 54 (60 – structural profile), 51 (60 – structural profile – flexibility profile) and 50 (60 – structural profile – flexibility profile – energy profile) features. The performance of these models and the one developed using all 60 features are reported in **Table 27**. The training and test were done using rcp_tr475 and rcp_ts169 datasets, and the window size was set to 1.

The results show that all MCC, F1 score, precision and recall continues to decrease with the removal of the dominant feature categories. Specifically, we observed no less than 5% decrease in MCC as we removed the structural, flexibility and energy profile. In addition, the F1 score is decreased by 6.2%, 4.4% and 4.8% after removal of the 6 structural properties, 3 flexibility-related properties including disorder probability and backbone angle fluctuations, and 1 position specific estimated energy (PSEE), respectively. These results validate the importance of the top features used to develop our predictor.

Table 27. Performance comparison of different feature sets (training set: rcp_tr475 and test set: rcp_ts169).

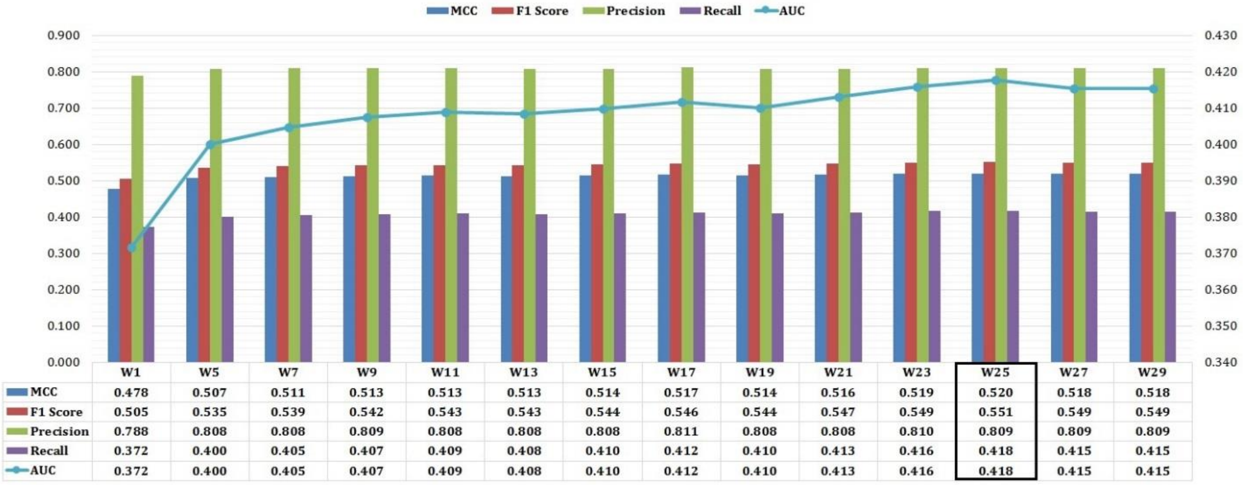
Metric	**60 Features**	**54 Features**	**51 Features**	**50 Features**
MCC	**0.478**	0.454	0.431	0.407
F1 Score	**0.505**	0.474	0.453	0.431
Precision	**0.788**	0.787	0.765	0.739
Recall	**0.372**	0.339	0.322	0.304

Best values are marked in bold.
60 features: all
54 features: all – structural profile
51 features: all – structural profile – flexibility profile
50 features: all – structural profile – flexibility profile – energy profile

5.4.2 Window Selection

In this section, we search for a suitable value of the sliding window size (W). The value of W approximates the number of residues around a target residue that may form the necessary local environment for inducing the peptide-protein transient interaction. We developed 14 different models with extra-tree (ET) classifier with 13 different window sizes (1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27 and 29). We chose ET classifier for this set of runs as this technique is relatively cheaper from a computational point of view and found to result comparable performance (*see* **Section 5.4.3**). The models were trained using rcp_tr475 dataset, and was independently tested using rcp_ts169 dataset.

The result of this experiment is shown in **Fig 48** in terms of recall, precision, F1 score, MCC and AUC score. We observed that all of the scores were improved with the increase of window size, which highlights that inclusion of neighborhood residue information better guides the predictor to learn about a target residue. We have also observed some irregular changes in the MCC and precision scores, which were not very significant. However, the score remained same for window size 19 and higher. Finally, we picked 25 as an optimum value of window as it gave better MCC, F1 score, recall and AUC values than the adjacent competing window sizes 23 and 27. Therefore, we took the features of 12 residues on either side of a target residue while determining whether the target residue is interacting or not.



	W1	W5	W7	W9	W11	W13	W15	W17	W19	W21	W23	W25	W27	W29
MCC	0.478	0.507	0.511	0.513	0.513	0.513	0.514	0.517	0.514	0.516	0.519	0.520	0.518	0.518
F1 Score	0.505	0.535	0.539	0.542	0.543	0.543	0.544	0.546	0.544	0.547	0.549	0.551	0.549	0.549
Precision	0.788	0.808	0.808	0.809	0.808	0.808	0.808	0.811	0.808	0.808	0.810	0.809	0.809	0.809
Recall	0.372	0.400	0.405	0.407	0.409	0.408	0.410	0.412	0.410	0.413	0.416	0.418	0.415	0.415
AUC	0.372	0.400	0.405	0.407	0.409	0.408	0.410	0.412	0.410	0.413	0.416	0.418	0.415	0.415

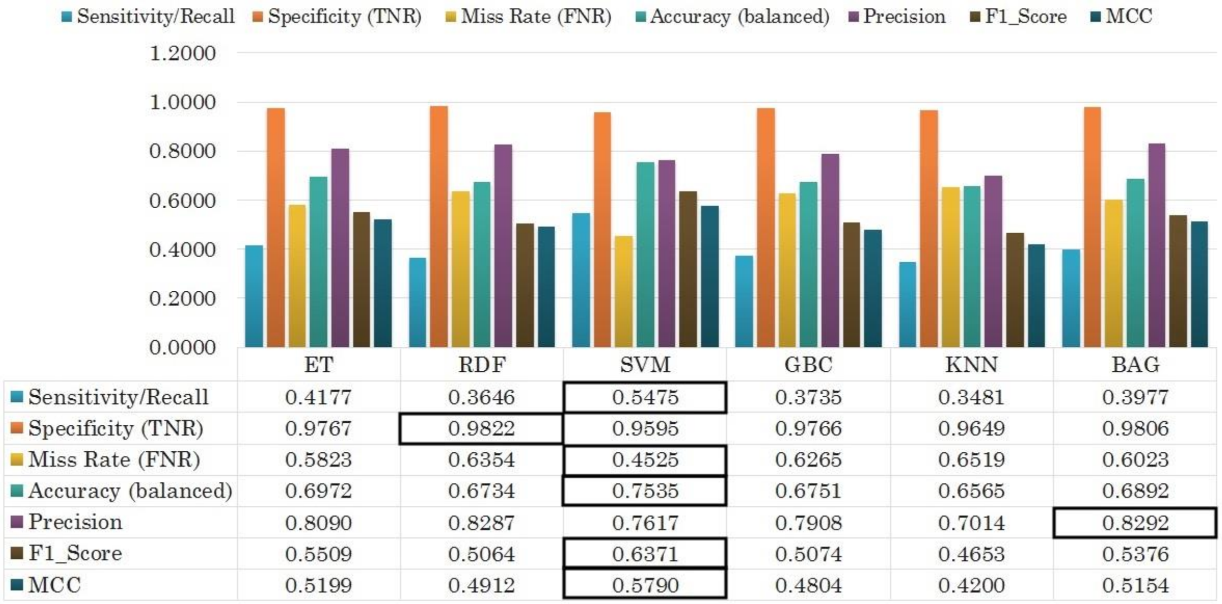
Fig 48. Performance comparison with different sliding window sizes for peptide-binding residue prediction using extra-tree classifier. The MCC, miss rate and recall scores are reported. The optimum size of window and the corresponding performance scores are marked by a black rectangle.

5.4.3 Performance Analysis of the Base Learners

In this section, we analyze the independent performances of the six base-learners, SVM, RDF, ET, GBC, KNN and BAG that we explored for binding-inducing region prediction in receptor proteins. The models were trained using rcp_tr475 dataset and were evaluated using independent test set, rcp_ts169. The predicted annotations were compared against the synthetic annotations of peptide-binding residues after smoothing that were used while training the models.

Fig 49 compares the binary prediction outputs of the learners and highlights that the optimized RBF-kernel SVM model gave outstanding performance in this application. The RBF-kernel SVM model gave the best recall (completeness of a classifier in predicting peptide-binding residues), miss-rate (rate of misclassifying a peptide-binding residue as non-binding), balanced accuracy (ACC) scores of values 0.547, 0.452 and 0.735. The closest competitor of RBF-SVM in terms of recall and ACC was the ET classifier.

The random forest (RDF) classifier performed the best for correctly predicting the non-binding residues in terms of specificity (0.982) and bagging (BAG) classifier gave the best precision score of 0.829 (correctness of a classifier in predicting binding-inducing residues). However, RBG-SVM model outperformed the other predictors in terms of two critical measures used to assess the performance of a binary classifier, MCC (regarded as the most effective measure for binary classification on an imbalanced dataset) and F1 score (balances between correctness and completeness of a classifier) with score values of 0.579 and 0.637, respectively. These scores are 11.35% and 15.62% better those provided by the closest competitor, ET classifier. On the other hand, GBC and KNN performed similarly, which were comparatively lower than the other predictors.

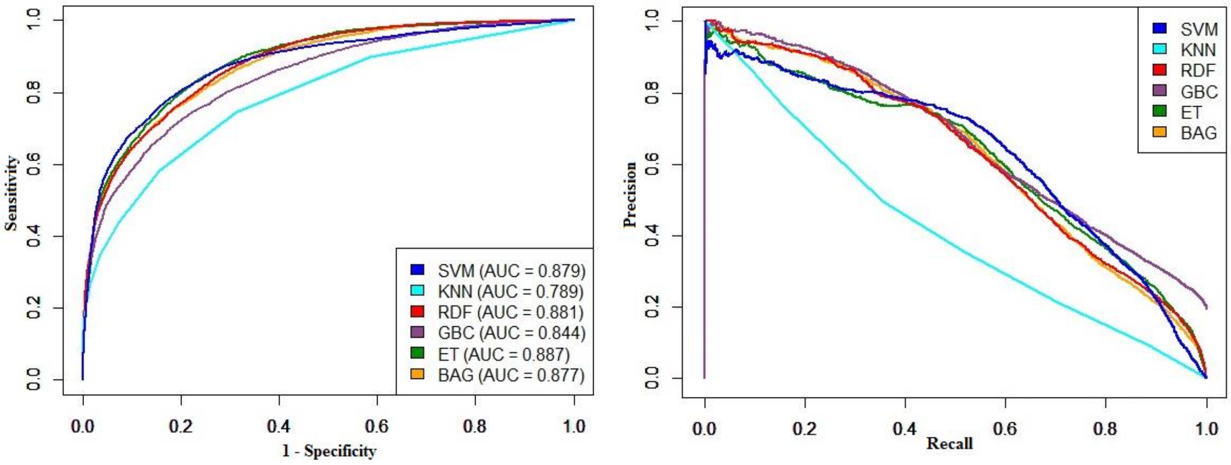


	ET	RDF	SVM	GBC	KNN	BAG
■ Sensitivity/Recall	0.4177	0.3646	0.5475	0.3735	0.3481	0.3977
■ Specificity (TNR)	0.9767	0.9822	0.9595	0.9766	0.9649	0.9806
■ Miss Rate (FNR)	0.5823	0.6354	0.4525	0.6265	0.6519	0.6023
■ Accuracy (balanced)	0.6972	0.6734	0.7535	0.6751	0.6565	0.6892
■ Precision	0.8090	0.8287	0.7617	0.7908	0.7014	0.8292
■ F1_Score	0.5509	0.5064	0.6371	0.5074	0.4653	0.5376
■ MCC	0.5199	0.4912	0.5790	0.4804	0.4200	0.5154

Fig 49. Peptide-binding residue prediction performance of the base-learners. The score outputs for all the measures are grouped together for each of the six base-learners. The best value in each scoring metric is marked by a black box. The *x*-axis and *y*-axis represent different learners and their corresponding score values respectively.

Fig 50 compares the ROC and precision-recall (PR) curves produced by the six methods that were tested as base-learners for this application. The ROC and PR curves can assess the performance of a classifier throughout its entire operating range by evaluating the probability distribution at different thresholds. The curves illustrate that the ET and RDF classifiers gave the highest and the second-highest AUC values of 0.887 and 0.881, respectively. The RBF-SVM was a close competitor with AUC value of 0.879. The KNN classifier provided the lowest AUC value of 0.789.

Fig 50(a) shows the complementary competitiveness of SVM with RDF and ET classifier at different points. Specifically, the recall/sensitivity of the SVM was lower than those of RDF and ET classifier at the range of high specificity (0.5 – 0.9), whereas at the range low specificity (0.0 – 0.45), SVM was better than RDF and ET. Another tree based ensemble learner, bagging (BAG) showed a similar performance to those of RDF and ET.



（a） ROC curves

（b） Precision-recall Curves

Fig 50. ROC and precision-recall curves given by 6 base-learners on peptide-binding residue prediction. Comparison of （a） ROC curves and （b） precision-recall curves on rcp_ts169 dataset by six different base-learners. The area under ROC curves （AUROC） are given in the plot （a）.

The PR curves in **Fig 50(b)** highlight that the precision of GBC, RDF and BAG were initially better than SVM and ET at the range low recall (0.0 – 0.4). However, SVM and ET gave better precision at higher recall (0.5 – 0.9). The gradient-boosting classifier (GBC) gave slightly different PR curve that involves sharp rise in precision value and it continued providing a reasonable precision for rest of the range of recall value. We also observed that the curves of KNN classifier were the least competitive.

5.4.3.1 *Correlation Analysis of the Base-Learners*

We further performed a pair-wise correlation analysis of the residue-wise probability outputs resulted on rcp_ts169 dataset by these six learners, reported in **Table 28**. To carry out this analysis, we computed the Persons correlation coefficient (ρ) between the two sets of probabilities given by two classifiers using following equation.

$$\rho = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}}$$

According to the working principle of stacking, as discussed in Section 2.3, it is useful to apply stacking using the base-learners that can capture diversifier regions of the feature space and therefore provides uncorrelated outputs. In this way, the meta-learner (LogReg) can learn about the improper training of the base-learner.

Table 28. Correlation of probability distribution generated by six base-learners on rcp_ts169 dataset.

Pair-wise Correlations Among Six Classifies						
Classifiers	ET	RDF	SVM	GBC	KNN	BAG
ET	–	0.891	0.794	**0.652**	**0.600**	0.890
RDF	–	–	0.734	**0.676**	**0.610**	0.910
SVM	–	–	–	**0.627**	**0.556**	0.760
GBC	–	–	–	–	**0.558**	**0.693**
KNN	–	–	–	–	–	**0.603**

Correlation value less than 0.7 are marked by bold.

Table 28 shows that the output of a tree-based ET classifier is highly correlated with the other tree-based ensemble learners, RDF and BAG, whereas is less correlated with the outputs of GBC, SVM and KNN classifiers. Therefore, a potential set of complementary learners is, ET, SVM, GBC and KNN. On the other hand, SVM is found less-correlated with GBC and KNN classifiers with correlation values of 0.627 and 0.556, respectively. Note that, from the results reported in **Fig 49**, we found that SVM is the best representative classifier for this application and the GBC and KNN are less competitive. Therefore, another potential set is: SVM, GBC and KNN classifiers. We have further verified different sets of base learners in the section below.

5.4.4 Parameter Selection for the Stacked Models

Here, we evaluate different sets of base-learners and features, and select the best combination to be used to generalize the stacking in PBRpredict-Suite.

5.4.4.1 *Selection of the Base-Level Learners*

Here, we evaluated four different combination of base-learners for stacked models (sM):

- **sM1**: ET, SVM, GBC, KNN, RDF and BAG (all six learners).

- **sM2**: ET, SVM, GBC and KNN. The RDF and BAG, which were highly correlated with ET (**Table 28**) are discarded.
- **sM3**: ET, SVM and GBC. The least performing KNN classifier while tested as a sole model is not considered in this set.
- **sM4**: SVM, GBC and KNN. Here, we combined the best performing base-learner, SVM with two relatively less competitive classifiers, GBC and KNN.

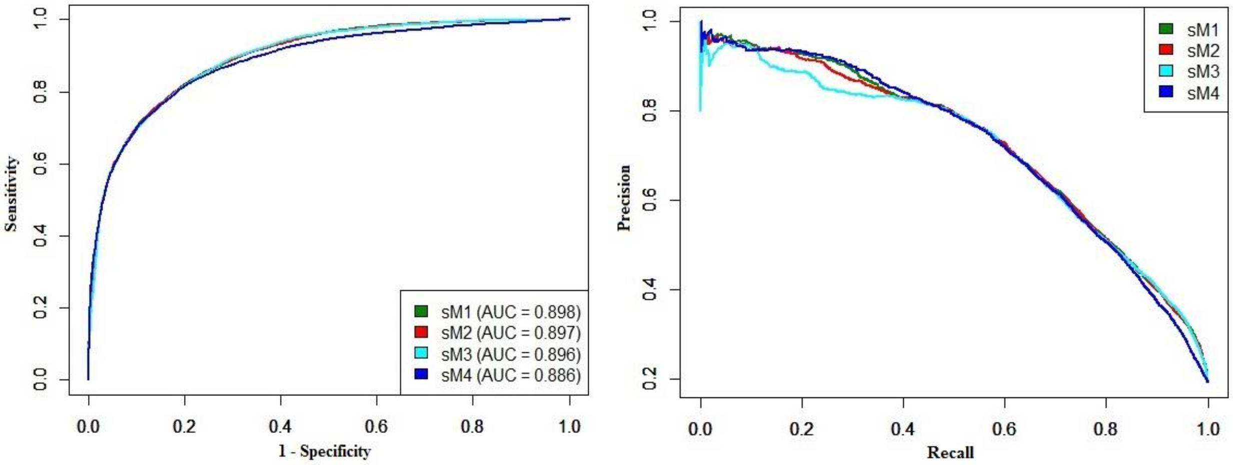
For all cases, the meta-level learner was the LogReg which was trained using rcp_tr475 dataset. Here, two different folds of the dataset were independently predicted by the base-learner models to generate probabilities while the models were trained on the other fold (**Fig 45**). Finally, the LogReg models were evaluated using independent test set, rcp_ts169.

Table 29. Comparison of different stacked models in peptide-binding residue prediction on rcp_ts169 dataset.

Score Types	**sM1**	**sM2**	**sM3**	**sM4**
Sensitivity (TPR)	0.551	0.551	0.546	**0.553**
Specificity (TNR)	0.959	0.959	**0.960**	0.959
Fall-out Rate (FPR)	0.041	0.041	**0.040**	0.041
Miss Rate (FNR)	0.449	0.449	0.454	**0.447**
ACC	0.755	0.755	0.753	**0.756**
Precision	0.762	0.762	**0.765**	0.760
F1 Score	0.639	0.640	0.638	**0.640**
MCC	0.581	**0.582**	0.580	0.581

sM1 uses ET, SVM, GBC, KNN, RDF and BAG as base-learners
sM2 uses ET, SVM, GBC and KNN as base-learners.
sM3 uses ET, SVM and GBC as base-learners.
sM4 uses SVM, GBC and KNN as base-learners.
Best values are marked in bold.

The performance comparison among four stacked models, shown in **Table 29**, clarifies that our assumption about the effective set of base-learners was reasonable. The model using all six base-learners (sM1) was outperformed by the stacked models with reduced number of complementary base-learners. After removing BAG and RDF classifiers from the set (sM2), we got a slight improvement in MCC score. The sM3 with ET, SVM and GBC only provided the highest specificity/TNR (0.96) and precision (0.765), and the lowest fall-out rate (0.04). On the other hand, the stacking of SVM, GBC and KNN in sM4 gave the highest recall (TPR), ACC and F1 score of values 0.553, 0.756 and 0.640, respectively.



（a） ROC Curves

（b） Precision-recall Curves

Fig 51. ROC and precision-recall curves given by 4 different stacked models on peptide-binding residue prediction. Comparison of （a） ROC curves and （b） precision-recall curves on rcp_ts169 dataset by three stacked models. The area under ROC curves （AUROC） are given in the plot （a）.

Fig 51 shows the ROC and precision-recall curve comparison among four stacked models, respectively in (a) and (b). The ROC curves for sM1, sM2 and sM3 were overlapping and the AUC values were also similar. However, the AUC of sM4 was slightly worse. On the other hand, the stacked model sM4 gave the best precision-recall curve with highest precision at the range of low recall (0.0 – 0.5).

We prioritized the balanced prediction capacity of a model in this classification task that can be measured by ACC and F1 score (**Table 29**). Therefore, we utilized the base-learner set of sM4 (SVM, GBC and KNN) to develop the predictor models in the PBRpredict-Suite which were combined using LogReg as meta-learner.

5.4.4.2 *Combination of Features for Meta-Learner*

During the selection of base-learners, results reported in **Table 30**, we used only the probability outputs generated from the base-learners as the features in the meta-level. Here, we further want to include additional features to boost up the capacity of meta-learner. We tested two different feature plans to train the meta-learner of sM4 stacked model that combines SVM, GBC and KNN that are given below.

(1) **Feature plan# 1** that contains the probability outputs generated by the base-learners only.

(2) **Feature plan# 1** contains the probability outputs generated by the base-learners and the 60 features per-residue, which are discussed in **Section 5.2.3**.

Table 30. Comparison of stacked model (sM4) with two different feature plans for meta-learner on rcp_ts169 test set.

Score Types	Feature Plan# 1	Feature Plan# 2
True Positive	2855	**2880**
True Negative	**20912**	20901
False Positive	**903**	914
False Negative	2307	**2282**
Sensitivity/Recall	0.553	**0.558**
Specificity (TNR)	**0.959**	0.958
Fall-out Rate (FPR)	**0.041**	0.042
Miss Rate (FNR)	0.447	**0.442**
ACC	0.756	**0.758**
Precision	**0.760**	0.759
F1 Score	0.640	**0.643**
MCC	0.581	**0.584**

Best values are marked in bold.

The outputs of feature plan# 1 and 2 were complementary, shown in **Table 30**. The meta-learner of plan# 1 gave better specificity, which emphasizes the predictors capacity to identify non-binding residues. In contrast, the meta-model of plan# 2 provided better recall that focuses the predictor's ability to accurately identify the binding residues. Moreover, the model with feature plan# 2 resulted in balanced prediction in terms of ACC, MCC and F1 score. Therefore, the final models in PBRpredict-Suite use SVM, GBC and KNN as the base-learners that were trained using (60 × 25) features and LogReg as meta-learner that was trained using 63 features.

5.4.5 Finalizing PBRpredict-Suite Models

In the proposed PBRpredict-Suite, we included three models to predict the protein's peptide-binding residues from sequence alone: PBRpredict-strict, PBRpredict-moderate and PBRpredict-flexible. In this section, we discuss the related results to support the development of these 3 different predictor models.

We named the stacked model sM4 with 63 features in the meta-level (**Section 5.4.4**) as PBRpredict-strict. This model provided a well-balanced performance when compared with the state-of-the-art predictor that is supported by both statistics (**Section 5.4.6**) and case-studies (**Section 5.4.7**). However, we called this model '*strict*' in predicting the positive class– peptide-binding residues as it resulted in fine false positive rate (fall-out rate/FPR) even at the cost of compromised recall score (TPR). Moreover, we intended to design models that can identify the peptide-binding sites in the structure-specific (relatively shorter) sequence as well as within the full-length protein sequence. Note that, we included only the structure-specific sequences from PDB in our training dataset as we needed the experimental structures to extract the interaction information and annotate the protein sequence. Therefore, the model was informed about the shorter sequence only. We observed that the PBRpredict-strict model provides conservative performance in identifying the binding residues in full-length sequence to avoid the false positive predictions or over-prediction (*see* **Figure 9**). These observations led us to tune our model further to improve the true positive rate (recall/TPR) or positive-class prediction accuracy of our model.

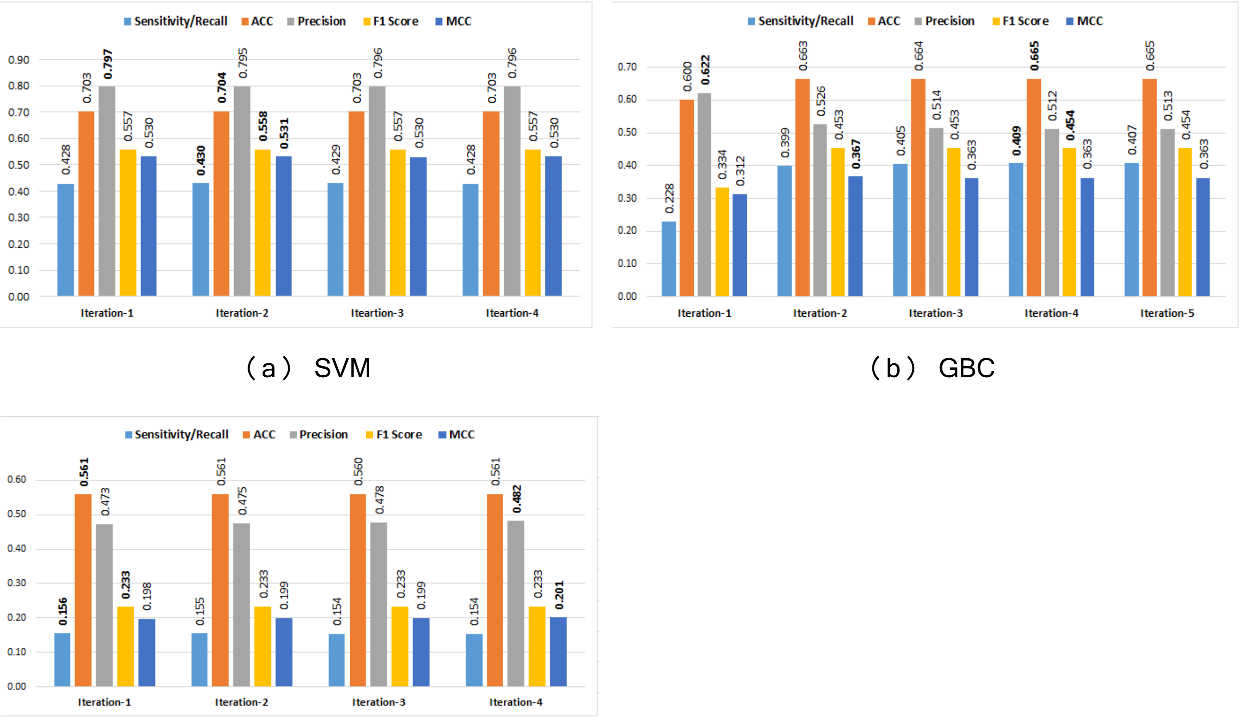
To increase the recall score of the model, we tried two techniques. As the first technique, we iteratively trained the base-learners to improve their performance and then combined their outputs with a meta-learner. Secondly, we modified the classification thresholds of both base-level and meta-level learners to trade-off between the recall (true positive rate/TPR) and fall-out rate (false positive rate/FPR). We discuss the results of the experiments below.

5.4.5.1 *Iterative Training of the Base-Learners*

To improve the recall, we tried to boost-up the performances of the individual base-learners by iterative training. Thus, we used the output probabilities generated by a learner as a feature in addition to the original feature set to train that learner in the second iteration. In this way, we continued to train each of the base-learners (SVM, GBC and KNN) until their performances degraded or reached a plateau, specifically up to 4th iterations for SVM and KNN, and 5th iterations for GBC. **Figure 52 (a) – (c)** show the outputs of iterative training and test of SVM, GBC and KNN. The scores correspond to the output of two-fold cross-validation on the training set (rcp_tr475), thus the average performance of testing with the one-fold while training with the other fold.

From **Figure 52(a)**, we observed that SVM resulted in the best recall, ACC, F1 score and MCC at the 2nd iteration, and after that the performance is deteriorated. In case of the gradient boosting classifier (GBC), the recall, accuracy and F1 scores continued to increase up to the 4th iteration and then got flat, while the best precision and MCC were achieved in the 1st and 2nd iterations, respectively. For KNN, the harmonic mean of the precision and recall (F1 score) and the balanced accuracy remained alike for all the iterations. On the other hand, the best recall value was scored at the 1st iteration.

To emphasize on the recall score, we chose the 1st iteration model for KNN along with the respective 2nd and 4th iteration models for SVM and GBC for combination using LogReg meta-level learner. We tested the meta-learner using rcp_ts169 test dataset. The output scores were 0.542, 0.638 and 0.582 in terms of recall, F1 score and MCC, respectively, which were 2.85%, 0.80% and 0.32% lower than those of the original PBRpredict-strict model (**Table 30**). Therefore, the iterative training improved the individual performance of the base-learners, however could not improve the performance of the stacked model. Moreover, the separately improved base-learners were not better than the PBRpredict-strict model on the rcp-ts169 dataset. Therefore, we did not consider this technique further into our predictor-models.



（a） SVM

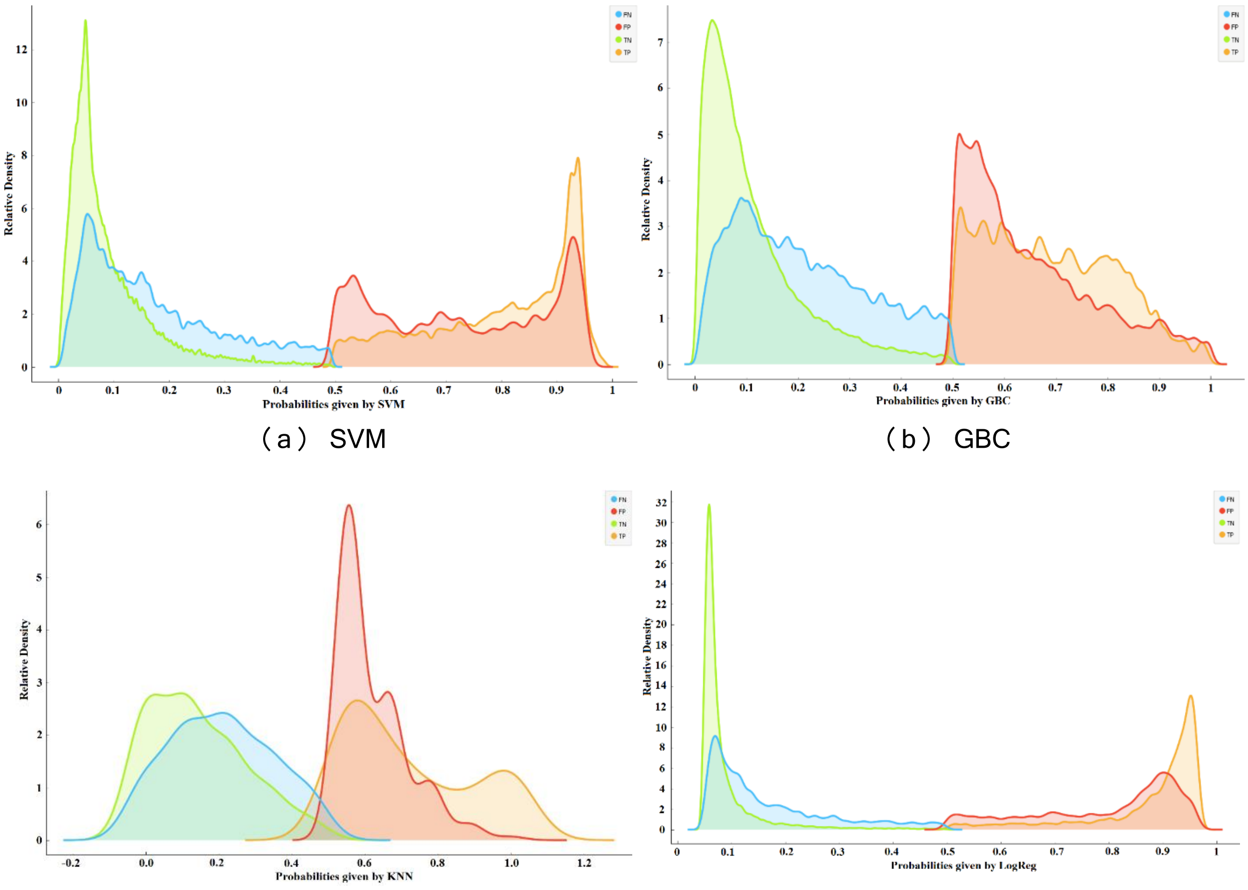
（b） GBC

（c） KNN

Fig 52. Performance of the iterative training and testing of the base-learners: (a) SVM, (b) GBC and (c) KNN. The training was carried-out using one-fold of the training dataset（rcp_tr475）, and the output model was tested on the other fold. This process was repeated for the two folds of rcp_tr475 set（Table 25）and the average score values are reported. The best values are bold-faced.

5.4.5.2 *Tuning of Thresholds for the Learners*

Next, we attempted to relax the classification threshold to recover the positive-class type (peptide-binding) residues that are falsely predicted as negative-class (non-binding). A classification threshold, which is traditionally kept as 0.5, is used to binarize the real-value probabilities generated by a classifier algorithm such as the samples with a probability output ≥ threshold is predicted as of positive-class, otherwise labeled as of negative-class. To understand the probabilistic behavior of the learners, we visualized the distributions of the probabilities generated by SVM, GBC, KNN for four different prediction types: true positives (TP), false positive (FP), true negative (TN) and false negative (FN) using the threshold value 0.5. **Fig 53** shows the distribution plots. The plots for SVM, GBC and KNN of **Fig 53(a)-(c)** were generated from the prediction outputs on the full rcp_tr475 dataset, where the each of the two folds of the dataset was independently predicted using the model trained on the other fold (**Table 25**). The plot for LogReg in **Fig 53(d)** was generated from the prediction outputs on rcp_ts169 dataset while trained on the full rcp_tr475.



（a） SVM

（b） GBC

（c） KNN

（d） LogReg

Fig 53. Probability distributions of different prediction types given by (a) SVM, (b) GBC, (c) KNN and (d) LogReg using the threshold value 0.5. The curves for true positives （TP），false positives （FP），true negatives （TN） and false negatives （FN） are drawn in *orange*, *red*, *green* and *blue* respectively. The *x*–axis and *y*–axis show the probabilities generated by the corresponding classifier and the relative density，respectively.

Note that, by tuning the threshold our purpose is to correct the false negative (FN) prediction outputs, represented by the blue curve in **Fig 53**. However, we must be careful as lowering the threshold from 0.5 will convert the corresponding true negatives (TN) under the green curve into false positives (FP), represented by the red curve. Therefore, we can only increase the accuracy of positive class (peptide-binding residue) prediction or decrease the miss rate at a cost of increased over-prediction rate (false positive rate). The plots of **Fig 53** again highlight the strength of SVM for this application. The SVM model correctly predicts the highest mass of binding (orange curve) and non-binding residues (green curve) with a high confidence, higher (0.85 – 1.0) and lower (0.0 – 0.15) probability values respectively. Moreover, **Fig 53(a)** shows that the SVM model provided the lowest overlap between TNs (correctly predicted non-binding residues) and FNs (incorrectly predicted binding residues) near the threshold margin compared to the other base-learners, GBC and KNN. Therefore, we can lower the threshold of SVM to gain an increase in recall score (TPR) at a cost of lower increase in the false positive rate (FPR). On the other hand, we noticed an opposite scenario from the outputs of KNN in **Fig 53(c)** with almost overlapped density curves for TNs and FNs. Therefore, we can only achieve an increase in TPR at a cost of high FPR. To mention, the curves for GBC in **Fig 53(b)** were better than those of KNN, however worse than those given by SVM. **Fig 53(d)** shows that density curves given by LogReg which were even better than SVM in terms of overlap between TNs and FNs near the margin (0.5). It suggests that the application of the meta-learner improved the performance over the base-learners and we can tune the threshold of the meta-learner as well to correct the FNs.

To search for appropriate thresholds, we checked 7 different values: 0.45, 0.4, 0.35, 0.3, 0.25, 0.2 and 0.15 other than the traditional value: 0.5. We evaluated the base-learners, SVM, GBC and KNN by two-fold cross-validation on the training set, and the average results are shown in **Table 31**. This experiment did not result any certain value of the threshold. For all classifiers, the recall and balanced accuracy continued to increase with the lower threshold values at a cost of very high over-prediction which is not desirable. Thus, we finally chose the thresholds according to certain statistics on the probabilities of false negatives (FNs) given by the classifiers as our aim is to correct FNs by assigning a different threshold to segregate the positive and negative class.

We quantified the mean probabilities of FNs (FN_{prob}) from the distribution of **Fig 53** along with the standard deviations (std) which are 0.172 ± 0.122 for SVM, 0.209 ± 0.130 for GBC, 0.208 ± 0.138 for KNN and 0.199 ± 0.105 for the LogReg. We checked the median values as well which are 0.139 for SVM, 0.187 for GBC, 0.222 for KNN and 0.191 for the LogReg. Then, we considered the *mean*(FN_{prob}) + *std*(FN_{prob}), *mean*(FN_{prob}) and *median*(FN_{prob}) values as different sets of thresholds.

Table 31. Cross-validation performance of SVM, GBC and KNN using 8 different thresholds on the training dataset (rcp_tr475).

Thresholds	Recall (TPR)	Specificity (TNR)	Fall-out Rate (FPR)	Miss Rate (FNR)	ACC	Precision	F1 Score	MCC
SVM								
0.5	0.428	**0.978**	**0.022**	0.572	0.703	**0.797**	0.557	0.530
0.45	0.447	0.974	0.026	0.553	0.710	0.774	0.566	**0.532**
0.4	0.470	0.967	0.033	0.530	0.718	0.741	0.575	0.529
0.35	0.496	0.957	0.043	0.504	0.727	0.699	0.580	0.523
0.3	0.528	0.943	0.057	0.472	0.735	0.650	**0.582**	0.513
0.25	0.568	0.921	0.079	0.432	0.745	0.589	0.579	0.496
0.2	0.620	0.885	0.115	0.380	0.753	0.519	0.565	0.472
0.15	**0.696**	0.812	0.188	**0.304**	**0.754**	0.425	0.527	0.425
GBC								
0.5	0.228	**0.972**	**0.028**	0.772	0.600	**0.622**	0.334	0.312
0.45	0.270	0.964	0.036	0.730	0.617	0.597	0.372	0.330
0.4	0.314	0.951	0.049	0.686	0.632	0.561	0.403	0.339
0.35	0.368	0.933	0.067	0.632	0.651	0.524	0.432	0.349
0.3	0.430	0.908	0.092	0.570	0.669	0.482	0.454	0.354
0.25	0.506	0.870	0.130	0.494	0.688	0.437	**0.469**	**0.355**
0.2	0.593	0.814	0.186	0.407	0.703	0.388	**0.469**	0.348
0.15	**0.694**	0.722	0.278	**0.306**	**0.708**	0.332	**0.449**	0.325
KNN								
0.5	0.156	**0.966**	**0.034**	0.844	0.561	**0.473**	0.233	0.198
0.45	0.156	**0.966**	**0.034**	0.844	0.561	**0.473**	0.233	0.198
0.4	0.260	0.916	0.084	0.740	0.588	0.381	0.308	**0.206**
0.35	0.260	0.916	0.084	0.740	0.588	0.381	0.308	**0.206**
0.3	0.421	0.807	0.193	0.579	0.614	0.303	0.353	0.202
0.25	0.421	0.807	0.193	0.579	0.614	0.303	0.353	0.202
0.2	**0.654**	0.605	0.395	**0.346**	**0.629**	0.248	**0.360**	0.194
0.15	**0.654**	0.605	0.395	**0.346**	**0.629**	0.248	**0.360**	0.194

Best score values for each classifier are bold faced.

Table 32 shows the performances of SVM, GBC and KNN on rcp_ts169 dataset using these modified threshold values, $mean(FN_{prob})$ + $std(FN_{prob})$, $mean(FN_{prob})$ and $median(FN_{prob})$ along with the traditional value of 0.5. The results show that for all the classifiers, the recall, miss rate and accuracy (ACC) scores were improved if the thresholds are relaxed and set to a lower value, however, with a higher false positive (fall-out) rate and lower precision. The models with traditional threshold (0.5) produced the most balanced performance for SVM and KNN with the highest MCC scores. On the other hand, the models with thresholds equal to mean+std(FN_{prob}) provided the best F1 scores for all the classifiers and best MCC for GBC. Moreover, the fall-out or over-prediction rates with these threshold values were reasonable, specifically no greater than 7.5%. On the other, the median(FN_{prob}) values were lower than the mean(FN_{prob}) values for the SVM and GBC. Therefore, the use of median(FN_{prob}) values as thresholds resulted in outstanding recall scores, however at a cost of very high fall-out rate which was not desirable. In addition, the performances of KNN models with mean(FN_{prob}) and median(FN_{prob}) as thresholds were similar. Therefore, we did not consider the median(FN_{prob}) value as threshold in the meta-level.

Table 32. Comparison of SVM, GBC and KNN using different thresholds (statistically derived) on rcp_ts169 dataset.

Thresholds	Recall (TPR)	Specificity (TNR)	Fall-out Rate (FPR)	Miss Rate (FNR)	ACC	Precision	F1 Score	MCC
SVM								
Traditional: 0.5	0.547	**0.959**	**0.041**	0.453	0.753	**0.762**	0.637	**0.579**
Mean + Std: 0.3	0.639	0.926	0.074	0.361	0.782	0.672	**0.655**	0.576
Mean: 0.17	0.747	0.854	0.146	0.253	0.800	0.547	0.632	0.538
Median: 0.14	**0.785**	0.821	0.179	**0.215**	**0.803**	0.509	0.618	0.523
GBC								
Traditional: 0.5	0.373	**0.977**	**0.023**	0.627	0.675	**0.791**	0.507	0.480
Mean + Std: 0.34	0.526	0.934	0.066	0.474	0.730	0.652	**0.582**	**0.500**
Mean: 0.21	0.692	0.827	0.173	0.308	0.759	0.486	0.571	0.459
Median: 0.19	**0.722**	0.800	0.200	**0.278**	**0.761**	0.460	0.562	0.448
KNN								
Traditional: 0.5	0.348	**0.965**	**0.035**	0.652	0.657	**0.701**	0.465	**0.420**
Mean + Std: 0.35	0.440	0.926	0.074	0.560	0.683	0.586	**0.502**	0.411
Mean: 0.21	0.744	0.687	0.313	**0.256**	**0.716**	0.360	0.485	0.347
Median: 0.22	**0.744**	0.687	0.313	**0.256**	**0.716**	0.360	0.485	0.347

Best score values for each classifier are bold faced.

In **Table 33**, we report the results of the stacked models with modified threshold values on rcp_ts169 dataset. In the meta-level 63 features were used as suggested by the results reported in **Section 5.4.4.2**. The stacked model for which the $mean(FN_{prob}) + std(FN_{prob})$ and the $mean(FN_{prob})$ are used as thresholds for all the base-level and meta-level learners are named as PBRpredict-moderate and PBRpredict-flexible, respectively. The actual threshold values are reported in the column-heads of **Table 33**.

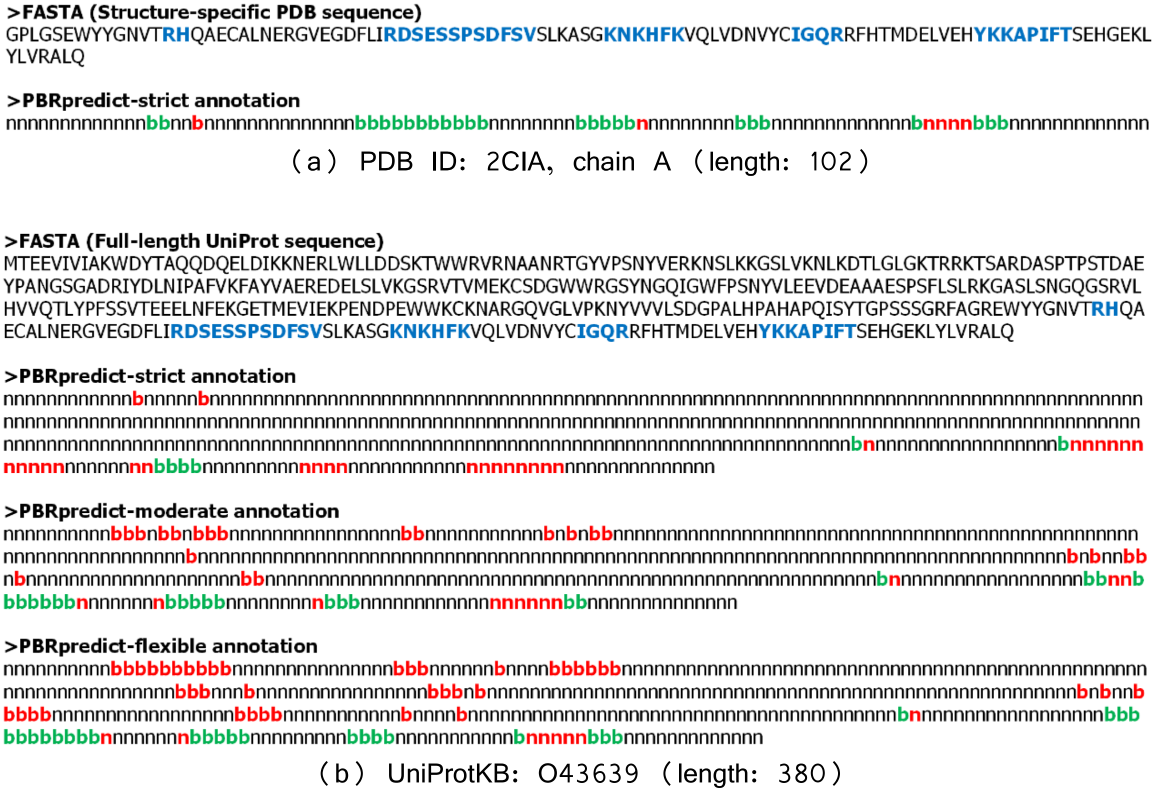
The outputs show that the PBRpredict-strict with threshold value of 0.5 resulted in the lowest fall-out rate with the highest MCC score (a balanced measure to assess a binary classifier), however the recall score was lower as well as the miss rate was higher than those of other models in the suite. In PBRpredict-moderate, the thresholds were relaxed and set to a relatively lower values, defined by the $mean(FN_{prob}) + std(FN_{prob})$. Subsequently, the true positive rate (TPR) was increased by 19.4% at a cost of 4.54% decrease in the true negative rate. In addition, the F1 score and ACC were also improved by 2.19% and 4.27% for the PBRpredict-moderate than those of PBRpredict-strict model. In the PBRpredict-flexible model, the thresholds were even more lowered and set to the $mean(FN_{prob})$. Therefore, all the false negative predictions (miss rate) of PBRpredict-strict with probability values greater than or equal to the $mean(FN_{prob})$ were corrected by the PBRpredict-flexible at a cost of high fall-out rate of around 16%.

Table 33. Comparison of the PBRpredict-strict, PBRpredict-moderate and PBRpredict-flexible models on rcp_ts169 dataset.

Performance metrics	PBRpredict-strict SVM (0.5), GBC (0.5), KNN (0.5), LogReg (0.5)	PBRpredict-moderate SVM (0.3), GBC (0.34), KNN (0.35), LogReg (0.3)	PBRpredict-flexible SVM (0.17), GBC (0.21), KNN (0.21), LogReg (0.2)
Recall (TPR)	0.558	0.666	**0.774**
Specificity (TNR)	**0.958**	0.915	0.841
Fall-out Rate (FPR)	**0.042**	0.085	0.159
Miss Rate (FNR)	0.442	0.334	**0.226**
Accuracy (balanced)	0.758	0.790	**0.808**
Precision	**0.759**	0.649	0.536
F1 Score	0.643	**0.657**	0.633
MCC	**0.584**	0.575	0.541

Best values are bold faced.
The threshold values for the classifiers in the PBRpredict-Suite models are reported in the column head.

In **Fig 54**, we illustrate the usefulness of these 3-different prediction using an example. PBD ID: 2CIA [391] stores the structure of a sequence (chain A) with SH2 domain bound to a phospho-peptide. In **Fig 54(a)**, we present structure-specific sequence of chain A (length: 102) and the predicted annotation produced by PBRpredict-strict. The peptide-binding residues are marked in *blue* on the amino acid sequence. The true and false predictions are marked respectively in *green* and *red* on the predicted annotations (‘b’ for peptide-binding and ‘n’ for non-binding). We observed that PBRpredict-strict could recognize most of the binding residues in the structure-specific sequence. However, the same model failed to recognize those residues when the input was the full-length sequence (UniProtKB: O43639, length: 380) containing the shorter structure-specific sequence (**Fig 54(b)**). On the other hand, the PBRpredict-moderate and flexible models could identify the binding residues on the full-length sequence, however with an increased number of false predictions of non-binding residues as binding residues. Therefore, the PBRpredict-Suite contains all these models that serve the purpose of recognizing peptide-binding residues under different scenarios.



>FASTA (Structure-specific PDB sequence)
GPLGSEWYYGNVT**RH**QAECALNERGVEGDFLI**RDSESSPSDFSV**SLKASG**KNKHFK**VQLVDNVYC**IGQR**RFHTMDELVEH**YKKAPIFT**SEHGEKL
YLVRALQ

>PBRpredict-strict annotation
nnnnnnnnnnnnn**bb**nn**b**nnnnnnnnnnnnnn**bbbbbbbbbbb**nnnnnnn**bbbbbn**nnnnnnnn**bbb**nnnnnnnnnnnnnn**bnnnnbbb**nnnnnnnnnnnnn

（a） PDB ID： 2CIA， chain A （length： 102）

>FASTA (Full-length UniProt sequence)
MTEEVIVIAKWDYTAQQDQELDIKKNERLWLLDDSKTWWRVRNAANRTGYVPSNYVERKNSLKKGSLVKNLKDTLGLGKTRRKTSARDASPTPSTDAE
YPANGSGADRIYDLNIPAFVKFAYVAEREDELSLVKGSRVTVMEKCSDGWWRGSYNGQIGWFPSNYVLEEVDEAAAESPSFLSLRKGASLSNGQGSRVL
HVVQTLYPFSSVTEEELNFEKGETMEVIEKPENDPEWWKCKNARGQVGLVPKNYVVVLSDGPALHPAHAPQISYTGPSSSGRFAGREWYYGNVT**RH**QA
ECALNERGVEGDFLI**RDSESSPSDFSV**SLKASG**KNKHFK**VQLVDNVYC**IGQR**RFHTMDELVEH**YKKAPIFT**SEHGEKLYLVRALQ

>PBRpredict-strict annotation
nnnnnnnnnnnn**b**nnnnn**b**nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn**bn**nnnnnnnnnnnnnnnnn**bnnnnnn**
nnnnnnnnnnn**nnbbbb**nnnnnnnnn**nnnn**nnnnnnnnnnn**nnnnnnnn**nnnnnnnnnnnnnn

>PBRpredict-moderate annotation
nnnnnnnnnn**bbb**n**bb**n**bbb**nnnnnnnnnnnnnnnn**bb**nnnnnnnnnnn**b**n**b**n**bb**nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnn**b**nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn**b**n**b**nn**bb**
n**b**nnnnnnnnnnnnnnnnnnnn**bb**nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn**bn**nnnnnnnnnnnnnnnnn**bbnnb**
bbbbbbnnnnnnnn**nbbbbb**nnnnnnnn**nbbb**nnnnnnnnnnnn**nnnnnnbb**nnnnnnnnnnnnnn

>PBRpredict-flexible annotation
nnnnnnnnnn**bbbbbbbbbb**nnnnnnnnnnnnnnn**bbb**nnnnnn**b**nnnn**bbbbbb**nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnn**bbb**nnn**b**nnnnnnnnnnnnnnnn**bbb**n**b**nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn**b**n**b**nn**b**
bbbbnnnnnnnnnnnnnnnnn**bbbb**nnnnnnnnnnn**b**nnnn**b**nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn**bn**nnnnnnnnnnnnnnnnn**bbb**
bbbbbbbbnnnnnnn**nbbbbb**nnnnnnnnn**bbbb**nnnnnnnnnnn**bnnnnnbbb**nnnnnnnnnnnnn

（b） UniProtKB： O43639 （length： 380）

Fig 54. The outputs of PBRpredict-Suite models on (a) the structure-specific and (b) the full-length sequence of the same protein. Fig 54（a） shows the protein sequence and predicted annotations given by PBRpredict-strict on PDB sequence （ID： 2CIA， chain A）． Fig 54（b） shows the protein sequence and predicted annotations given by all PBRpredict-Suite models on UniProt sequence （ID： O43696）． The peptide-binding residues are marked in *blue* on the amino acid sequence． The true and false predictions are marked respectively in *green* and *red* on the predicted annotations （‘b’ for peptide-binding and ‘n’ for non-binding）．

5.4.6 Performance Comparison with Other Predictors

In this section, we compare the performance of PBRpredict-Suite models with SPRINT [358]. SPRINT is a sequence-based predictor of protein-peptide binding residues that uses a SVM with optimized parameter set. Moreover, the dataset, model parameter set and feature set for SPRINT are different than those of PBRpredict. We ran SPRINT through its webserver on our test dataset, rcp_ts169. However, SPRINT server could generate prediction on 146 sequences out of 169, and failed for the rest. Thus, we compared the performance of the proposed models with that of SPRINT [358] on the 146 sequences only.

The comparison while evaluated against the synthetic annotation (with smoothing) is reported in **Fig 55**. We observed that SPRINT could result higher recall value than that of PBRpredict-strict model. Note that, we named this model 'strict' as it does not compromise the rate of false positive (fall-out rate) even at a cost of lower recall score. The recall score of PBRpredict-strict was found 10.69% lower than that of SPRINT. On the other hand, the fall-out rate of SPRINT, which defines the rate of miss-classification of non-binding residues as peptide-binding residues or tendency of over-prediction, was 86.52% higher than that of PBRpredict-strict. Moreover, the PBRpredict-strict gave more precise and balanced performance with 15.42%, 138.34%, 132.50% and 51.99% higher balanced accuracy (ACC), precision, F1 score and MCC, respectively than those given by SRINT. Further, the PBRpredict-moderate and flexible overcomes the shortcomings of the strict model. The PBRpredict-moderate and flexible provided 7.3% and 25.3% higher recall scores than that of SPRINT, respectively, while keeping the fall-out rate 72.3% and 48.3% lower than that of SPRINT. Thus, the three models in together made the PBRpredict-Suite comprehensive in identifying peptide-binding residues.

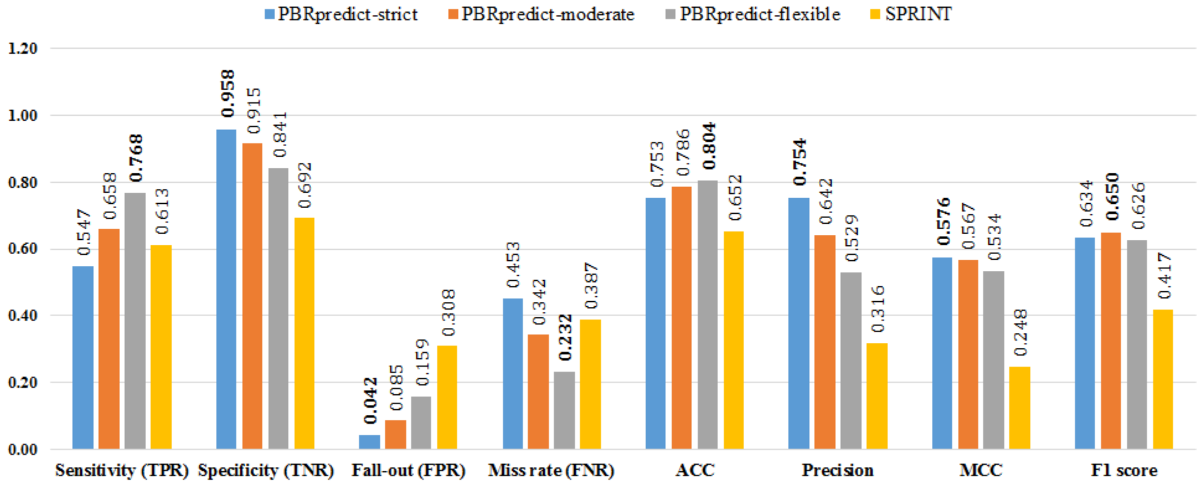


Fig 55. Performance Comparison of SPRINT and PBRpredict-Suite models in peptide-binding residue prediction, evaluated against synthetic annotation. The bars are grouped for the two predictors per metric. The best values in each metric type are marked in bold.

In **Fig 56**, we report the performance comparison while the predictions were evaluated against actual annotation (without smoothing). A similar result was obtained where SPRINT gave competitive sensitivity and miss-rate with PBRpredict-strict and moderate, however at a cost of higher fall-out rate, specifically 77% and 62.2% higher than that of PBRpredict-strict and moderate. Notwithstanding, PBRpredict-flexible resulted in 12.1% higher recall score than that of SPRINT even with 38.12% lower fall-out rate. In addition, PBRpredict-Suite models gave better balanced scores in case of assessing against actual annotation as well. Specifically, the ACC, precision, MCC and F1 score given by PBRpredict-strict were 6.79%, 109.82%, 74.72% and 43.99% higher than those of SPRINT, respectively. These differences in performance even increases for MCC and F1 scores when SPRINT was compared with PBRpredict-moderate as this model gave the best MCC and F1 score. The surprisingly superior performance given by SPRINT only in case of recall when compared to PBRpredict-strict, despite falling far behind it in terms of balanced measures such as MCC and F1 score provide us a clue that SPRINT suffers with over-prediction problem.

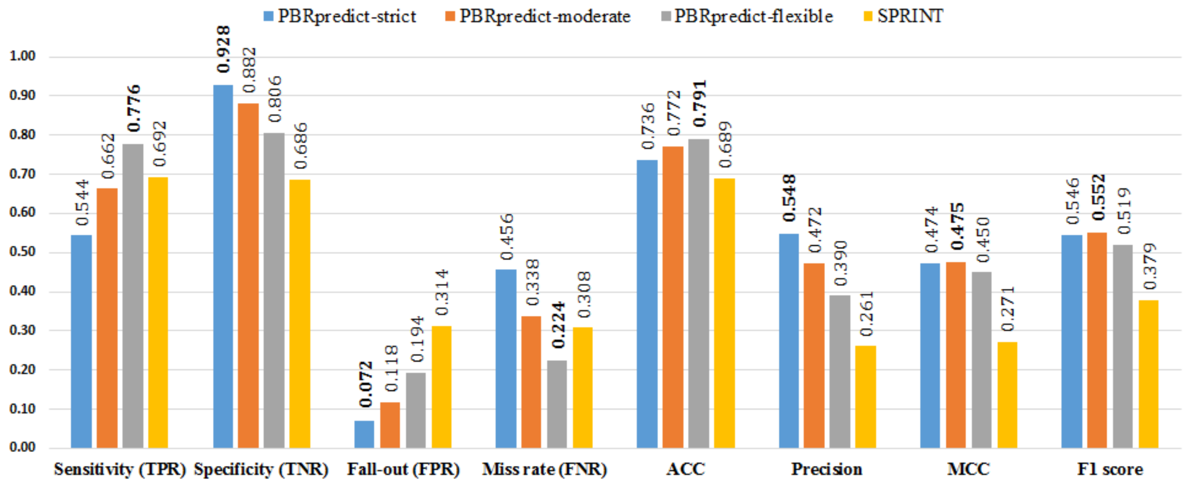
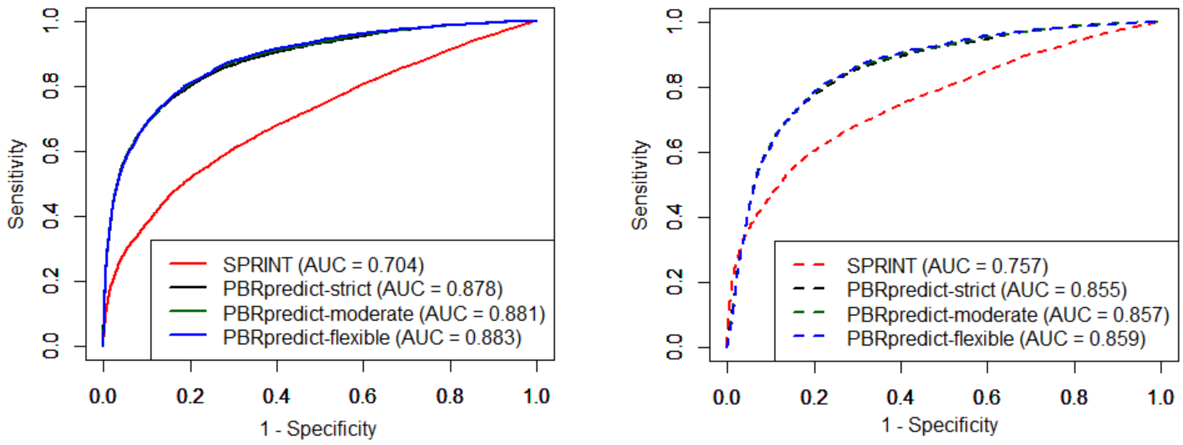


Fig 56. Performance Comparison of SPRINT and PBRpredict-Suite models in peptide-binding residue prediction, evaluated against actual annotation. The bars are grouped for the two predictors per metric. The best values in each metric type are marked in bold.

Fig 57 presents the ROC curves generated by SPRINT and PBRpredict-Suite models while the predictions are evaluated against both synthetic and actual annotations. The curves show the TPR (sensitivity)/FPR (1 – specificity) output pairs at different classification thresholds. The ROC curves given by different models of the PBRpredict-Suite nearly overlapped with each other. The curves highlight the strength of PBRpredict models in achieving a high true positive rate (TPR) of ≥ 80% (rate of correct prediction of peptide-binding residues) at a very low rate (20%) of false positive (FPR). On the other hand, SPRINT gave TPR ≥ 80% at a cost of high FPR ≥ 60% only. This performance gap persists when the predictions are compared against the actual annotation as well. Therefore, the synthetic annotation of the non-binding residues (negative class) as peptide-binding (positive class) in between disjoint peptide-binding regions did not contribute to over-prediction, rather better guided a machine learning technique to identify the binding residues from collective information of the residues at close vicinity. Moreover, the AUC scores given by PBRpredict-Suite models were at least 24.7% and 13% higher than those of SPRINT while evaluated against synthetic and actual annotation, respectively.



（a） ROC （synthetic annotation） （b） ROC （actual annotation）

Fig 57. Comparison of ROC curves and AUC values given by SPRINT and PBRpredict-Suite models on 146 chains. Evaluation against synthetic and actual annotations are indicated using solid and dotted lines in Figure （a） and （b）, respectively. The AUC values under the ROCs are reported in the legend.

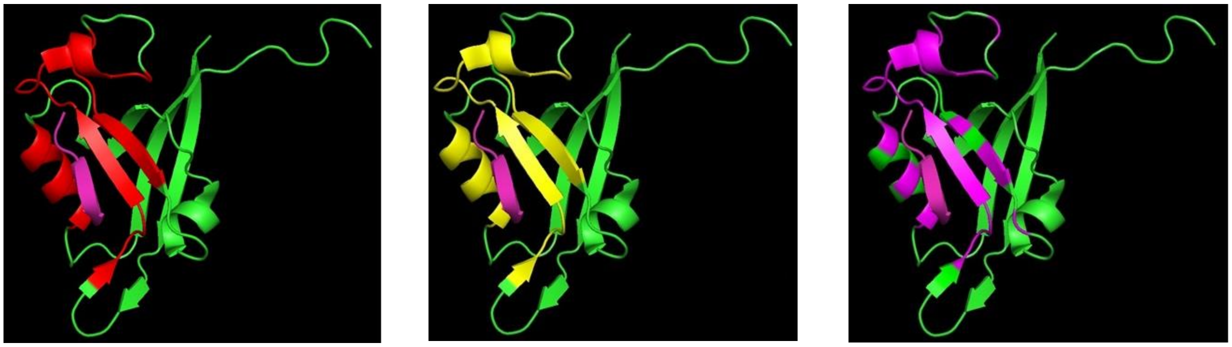
5.4.7 Case-Studies on Sequence with Known Domains

In this section, we perform case-studies with seven different proteins with different peptide-recognition domains (PRDs), discussed in **Section 5.2.1.1**. The structure-specific chains of these proteins were picked from the rcp_ts169 test set that share less than 40% similarity with any chain of the training set. However, chains with similar domain type were present in the training set. We applied the PBRpredict-strict that uses the traditional threshold and SPRINT to predict the peptide-binding residue in each protein and mapped the prediction outputs on the structure. We viewed the 3D structures using PyMOL [50] and the secondary structure was assigned using DSSP [51]. For a fair analysis and comparison on the structure-specific sequences, we applied the strictest model of the suite.

5.4.7.1 *PDZ Domain (PDB ID – 4NNM)*

We selected the crystal structure of Tax-Interacting Protein-1 (TIP-1) with PDZ domain[23] to analyze the performance of PBRpredict-strict and SPRINT in identifying PDZ domain, results shown in **Fig 58**. The structure is available as PDB ID: 4NNM where the PDZ domain of TIP-1 (*green*), having a disordered C-terminal, is bound to Y-iCAL36 (YPTSII) peptide (*pink*).

In **Fig 58 (a)**, the actual peptide-binding regions (by synthetic annotation) are shown in *red*. **Fig 58 (b)** and **(c)** show the predicted peptide-binding residues by PBRpredict-strict and SPRINT, highlighted in *yellow* and *pink*, along with the recall and MCC values. According to the statistics, the recall and MCC given by PBRpredict-strict (0.946 and 0.959) were better than those of SPRINT (0.729 and 0.755), however, SPRINT was comparable.



（a） Actual Annotation

（b） PBRpredict-strict Annotation
Recall = 0.946
MCC = 0.959

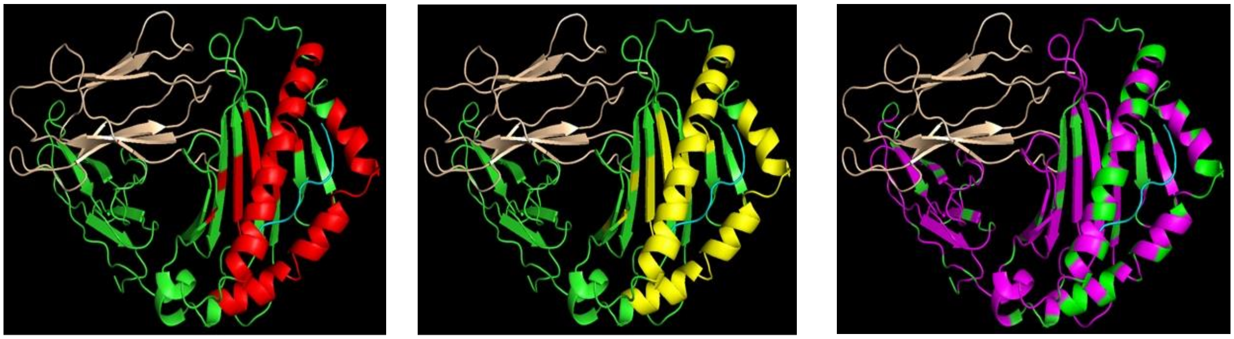
（c） SPRINT Annotation
Recall = 0.729
MCC = 0.755

Fig 58. Case study on PDZ domain (PDB ID: 4NNM). （a） Actual annotation of peptide-binding residues in the tax-intercation protein-1 （green） bound to peptide （*pink*）, （b） Prediction output of PBRpredict-strict （*yellow*） and （c） Prediction output of SPRINT （*pink*）, respectively. The figures in （b） and （c） are labeled by the corresponding prediction accuracies in terms of recall and MCC scores. We viewed the 3D structures using PyMOL[50] and the secondary structure was assigned using DSSP [51].

5.4.7.2 *MHC Domain (PDB ID – 1DL9)*

Here, we picked the three-dimensional structure of an H-2Ld protein interacting with a peptide, reported in PDB ID: 1LD9 [392]. The actual annotation is shown in **Fig 59(a)** where the MHC molecule is shown in *green*, the nine-residue long peptide is shown in *cyan* and the peptide-binding residues are marked in *red*. We predicted the peptide-binding residues of H-2Ld using PBRpredict-strict and SPRINT, shown in **Fig 59(b)** and **(c)**, respectively.

Prediction of PBRpredict-strict (*yellow*) for this case was very precise in terms of the statistical measures, recall value 1.0 and MCC value 0.99. The visual illustration of SPRINT prediction in **Fig 59(c)** clearly shows the over-predicted peptide-binding residues (*pink*) throughout the full chain with a MCC of -0.123 and recall of 0.59.



（a） Actual Annotation

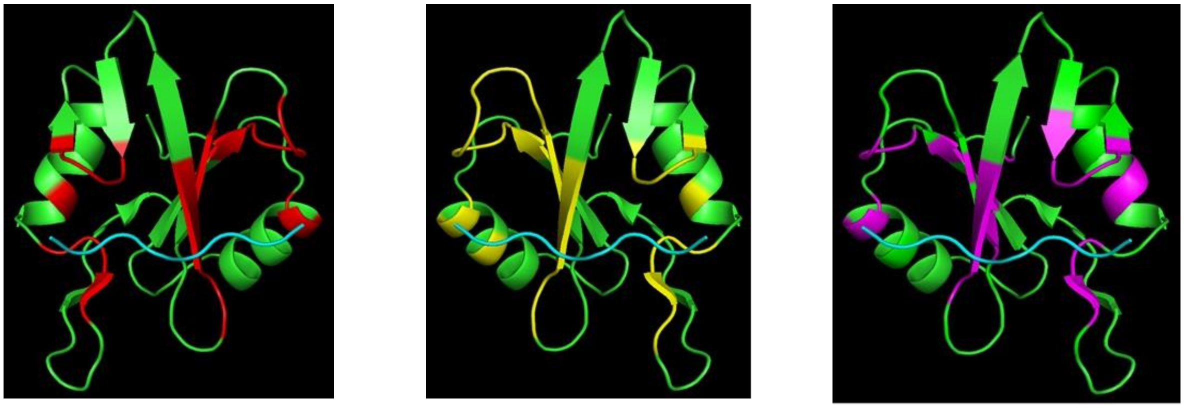
（b） PBRpredict-strict Annotation
Recall = 1.0
MCC = 0.99

（c）SPRINT Annotation
Recall = 0.59
MCC = -0.123

Fig 59. Case study with MHC molecule (PDB ID: 1DL9). （a） Actual annotation of peptide-binding residues （*red*） in the H-2Ld （green） bound to nine-residue long peptide （*cyan*），（b） Prediction output of PBRpredict-strict （*yellow*） and （c） Prediction output of SPRINT （*pink*），respectively. The figures in （b） and （c） are labeled by the corresponding prediction accuracies in terms of recall and MCC scores. We viewed the 3D structures using PyMOL [50] and the secondary structure was assigned using DSSP [51].

5.4.7.3 *SH2 Domain (PDB ID – 2CIA)*

To test the predictors on SH2 domain, we picked the structure of human Nck2 with SH2 domain (PDB ID: 2CIA [391]) in complex with a phosphotyrosine peptide. **Fig 60(a)** shows the actual peptide-binding residues (*red*) within the Nck2 protein (*green*) that recognizes the phosphopeptide (*cyan*).



（a）Actual Annotation

（b） PBRpredict-strict Annotation
Recall = 0.75
MCC = 0.77

（c）SPRINT Annotation
Recall = 0.63
MCC = 0.55

Fig 60. Case study with SH2 domain (PDB ID: 2CIA). （a） Actual annotation of peptide-binding residues （*red*） in Nck2 （green） bound to phosphotyrosine peptide （*cyan*），（b） Prediction output of PBRpredict-strict （*yellow*） and （c） Prediction output of SPRINT （*pink*），respectively. The figures in （b） and （c） are labeled by the corresponding prediction accuracies in terms of recall and MCC scores. We viewed the 3D structures using PyMOL [50] and the secondary structure was assigned by DSSP [51].

Fig 60(b) and **(c)** depicts the comparative performance of PBRpredcit-strict (*yellow*) and SPRINT (pink) for peptide-binding residue prediction within SH2 domain.We found that the most strict model of the PBRpredict-Suite resulted in recall and MCC values of 75% and 77%, respectively. These values were better than those given by SPRINT, recall: 63% and MCC: 55% respectively.

5.4.7.4 *Polo-Box Domain (PDB ID – 4LKL)*

Here, we picked the crystal structure of the polo-like kinase 1 (Plk1) with polo-box domain in bound with five-residue long PL-55, reported as PDB ID: 4LKL [393]. PBRpredict-strict and SPRINT were used to predict the peptide-binding residues of Plk1 polo-box domain, shown in **Fig. 61.** The actual annotation is shown in **Fig 61(a)** where the Plk1 molecule is shown in *green*, the five-residue long PL-55 is shown in *cyan* and the peptide-binding residues are marked in *red*.

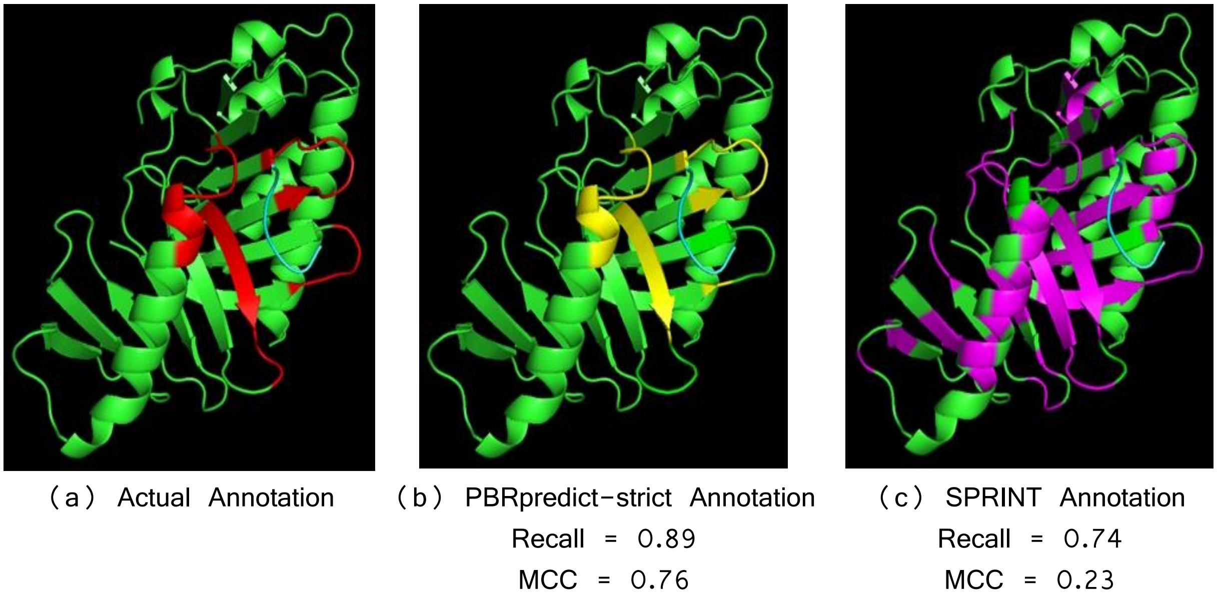


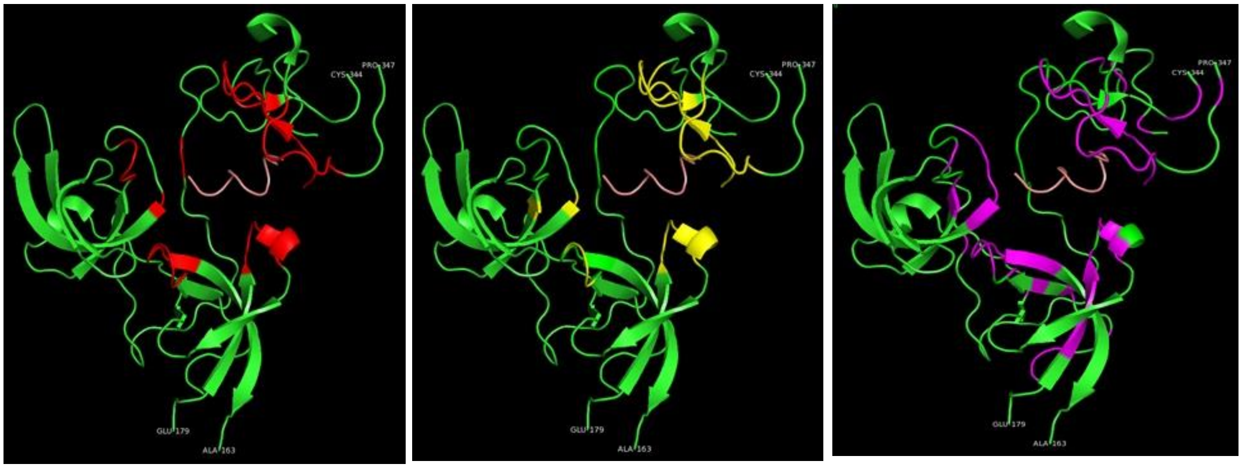
Fig 61. Case study with polo-box domain (PDB ID: 4LKL). (a) Actual annotation of peptide-binding residues (*red*) in Plk1 (green) bound to PL-55 (*cyan*), (b) Prediction output of PBRpredict-strict (*yellow*) and (c) Prediction output of SPRINT (*pink*), respectively. The figures in (b) and (c) are labeled by the corresponding prediction accuracies in terms of recall and MCC scores. We viewed the 3D structures using PyMOL[50] and the secondary structure was assigned using DSSP [51].

Fig 61(b) and **(c)** respectively show the predicted residues, generated by PBRpredict-strict (*yellow*) and SPRINT (*pink*). PBRpredict-strict correctly predicted 89% of the peptide-binding residues (recall) and gave a balanced MCC score of 0.76. On the other hand, SPRINT gave reasonable recall score of 74%, however, highly over-predicted the non-binding residues as peptide-binding (**Fig 61(c)**). Therefore, the MCC score of SPRINT was much lower (0.23) than that of PBRpredict-strict.

5.4.7.5 *Tudor Domain (PDB ID – 3ASK)*

To study with Tudor domain, we chose the crystal structure of UHRF1 [394], an essential factor for maintenance of DNA methylation, with a tandem Tudor domain and a PHD finger, in complex with the amino-terminal tail of histone H3. **Fig 62(a)** shows the peptide-binding residues (*red*) of UHRF1 (*green*), which are interacting with the histone tail (*tint*). The figure also shows the two disordered regions within UHRF1, residues 163 – 179 and 344 – 347.

Fig 62(b) and **(c)** show the predicted peptide-binding regions by PBRpredict-strict and SPRINT, highlighted in *yellow* and *pink* along with their recall and MCC values. We observed that for this PRD, the MCC and recall of PBRpredict were 74% and 77%. SPRINT gave reasonable recall score of 64%, however the MCC score was 44% only due to the over-prediction.



（a） Actual Annotation

（b）PBRpredict-strict Annotation
Recall = 0.74
MCC = 0.77

（c） SPRINT Annotation
Recall = 0.64
MCC = 0.44

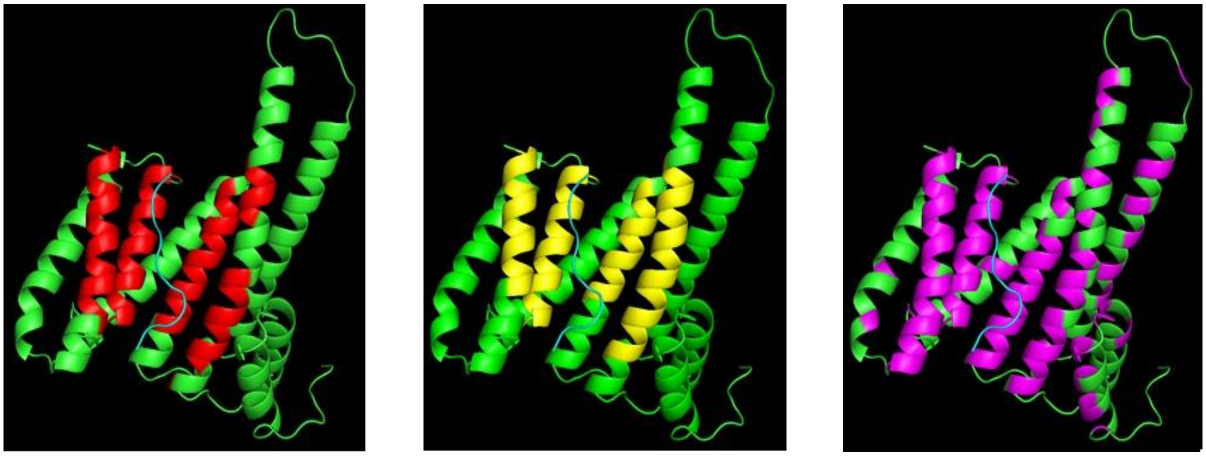
Fig 62. Case study with tudor domain (PDB ID: 3ASK). （a） Actual annotation of peptide-binding residues （*red*） in UHRF1 （green） bound to histone tail （*tint*）, （b） Prediction output of PBRpredict-strict （*yellow*） and （c） Prediction output of SPRINT （*pink*）, respectively. The figures in （b） and （c） are labeled by the corresponding prediction accuracies in terms of recall and MCC scores. We viewed the 3D structures using PyMOL[50] and the secondary structure was assigned using DSSP [51].

5.4.7.6 *14-3-3 Domain (PDB ID – 3MHR)*

Here, we picked the crystal structure of the 14-3-3 sigma in complex with phosphopeptide with phosphorylation of Ser127, reported as PDB ID: 3MHR [395]. BIRpredict and SPRINT were used to predict the peptide-binding residues of 14-3-3 domain, shown in **Fig 63.** The actual annotation is shown in **Fig**

63(a) where the 14-3-3 molecule is shown in *green*, the phosphopeptide is shown in *cyan* and the peptide-binding residues are marked in *red*.

Fig 63(b) and **(c)** respectively show the predicted residues, generated by PBRpredict-strict (*yellow*) and SPRINT (*pink*). PBRpredict-strict and SPRINT correctly identified 95% and 85%, respectively, of the peptide-binding residues (recall), which were comparable. On the other hand, the MCC scores for PBRpredict was much higher (0.93) than that of SPRINT (0.52).



(a) Actual Annotation

(b) PBRpredict-strict Annotation
Recall = 0.95
MCC = 0.93

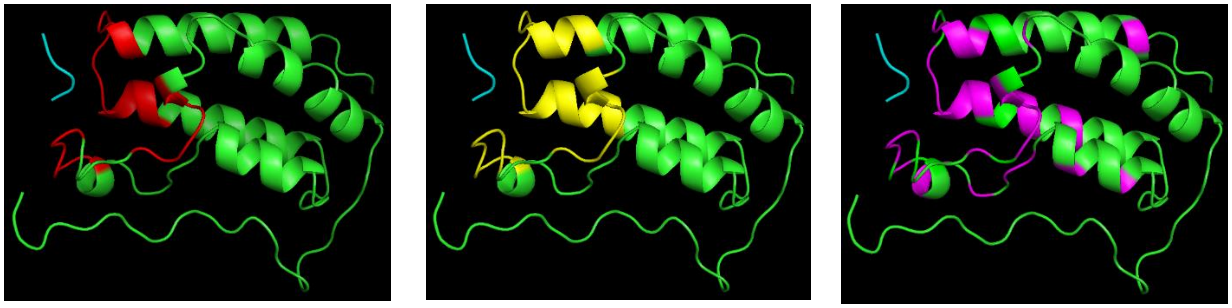
(c) SPRINT Annotation
Recall = 0.85
MCC = 0.52

Fig 63. Case study with 14-3-3 domain (PDB ID: 3MHR). (a) Actual annotation of peptide-binding residues (*red*) in 14-3-3 protein (green) bound to phosphoserine peptide (*cyan*), (b) Prediction output of PBRpredict-strict (*yellow*) and (c) Prediction output of SPRINT (*pink*), respectively. The figures in (b) and (c) are labeled by the corresponding prediction accuracies in terms of recall and MCC scores. We viewed the 3D structures using PyMOL[50] and the secondary structure was assigned by DSSP [51].

5.4.7.7 *Bromodomain (PDB ID – 3JVK)*

As To test the predictors on bromodomain, we picked the structure of mouse Brd4 (PDB ID: 3JVK[391]) in complex with histone H3-K(ac) peptide. **Fig 64(a)** shows the actual peptide-binding residues (*red*) within the Brd4 protein (*green*) that recognizes the acetylated-peptide (*cyan*).

Fig 64(b) and **(c)** depicts the performance of PBRpredcit-strict (*yellow*) and SPRINT (pink) for peptide-binding residue prediction within bromodomain. PBRpredict-strict correctly predicted all the peptide-binding residues (recall) and gave a balanced MCC score of 0.83. On the other hand, SPRINT gave a recall score of 0.75 and MCC score of 0.53.



（a） Actual Annotation

（b） PBRpredict-strict Annotation
Recall = 1.0
MCC = 0.83

（c） SPRINT Annotation
Recall = 0.75
MCC = 0.53

Fig 64. Case study with Bromodomain (PDB ID: 3JVK). （a） Actual annotation of peptide-binding residues （*red*） in Brd4 protein （green） bound to H3-K(ac) peptide （*cyan*）, （b） Prediction output of PBRpredict-strict （*yellow*） and （c） Prediction output of SPRINT （*pink*）, respectively. The figures in （b） and （c） are labeled by the corresponding prediction accuracies in terms of recall and MCC scores. We viewed the 3D structures using PyMOL[50] and the secondary structure was assigned using DSSP [51].

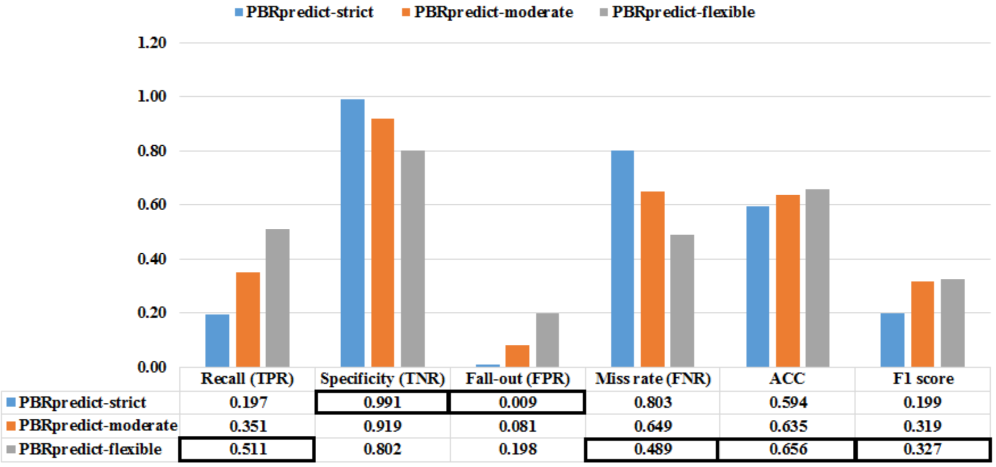
5.4.8 Case-Studies on sequences with Unknown Domains

In this section, we perform case-studies on structure-specific sequences with peptide-recognition domains that are not present in the dataset used to train PBRpredict-suite models. We picked 3 such domains: The MBT (**M**alignant **B**rain **T**umor) domain, VHS (**V**PS-27, **H**rs and **S**TAM) domain and CW domain. We collected the structures with these domains from PDB following similar steps described in **Section 2.1**. We respectively found 8, 9 and 10 structures of complexes in which chains with MBT, VHS and CW domains were bound to peptides. After filtering out the chains with a similar domain that shared greater than 40% sequence similarity, we had 6, 4 and 7 sequences with MBT, VHS and CW domains, respectively. Then, we extracted the interaction information from the structures and annotated the chains based on the atomic distance between the domain and peptide residue as described in **Section 5.2.2**. Below we discuss the performance of different PBRpredict-suite models in identifying peptide-binding residues on these sequences with domains that are not known to the models.

5.4.8.1 *MBT Domain*

The MBT domain recognizes the post-translational modifications, *i.e.*, methylation on lysine, on histone tails. The MBT domains are involved in transcriptional repression and have critical roles in diseases [396]. **Fig 65** shows the performances of the 3 models of PBRpredict-Suite in recognizing the residue patterns of this domain.

We can observe that the strict model identified only 19.7% of the peptide-binding residues, however, resulted in very low false positive rate (FPR). The moderate predictor could correct some of the incorrectly predicted binding residues, therefore the recall and accuracy scores were improved with a reasonable false prediction of the non-binding residues (FPR value of 8.1%). On the other hand, the model with the most flexible threshold values for the classification resulted in the highest recall, ACC and F1 scores.



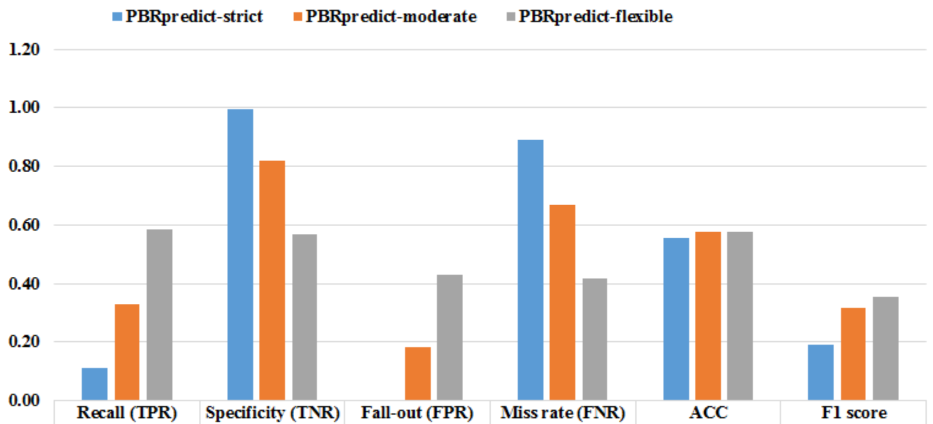
	Recall (TPR)	Specificity (TNR)	Fall-out (FPR)	Miss rate (FNR)	ACC	F1 score
■ PBRpredict-strict	0.197	0.991	0.009	0.803	0.594	0.199
■ PBRpredict-moderate	0.351	0.919	0.081	0.649	0.635	0.319
■ PBRpredict-flexible	0.511	0.802	0.198	0.489	0.656	0.327

Fig 65. Performance of the PBRpredict-Suite models on MBT domains. The baser are grouped for the three predictors per metric and the score values are reported in the data table below the plot. The best values for each metric are highlighted using black boxes.

5.4.8.2 *VHS Domain*

The VHS domains are mostly found in the N-terminal of many proteins and have crucial roles in membrane targeting [397]. VHS domain recognizes short peptide motifs, *i.e.*, D/ExxLL. **Fig 66** shows the average performances of the three PBRpredict-Suite models in recognizing the peptide-binding residues on 4 chains with this domain.

The results show that the PBRpredict-flexible model recognized the highest number of residues that were involved in the interactions with peptide residues with the highest recall (58.3%), accuracy (57.6%) and F1 score (35.4%). On the other hand, the strict model gave the lowest recall score, however, almost perfectly predicted the non-binding residues with only 2 false positives (FPR: 0.4%). The accuracy of the moderate model was in between the strict and flexible models.

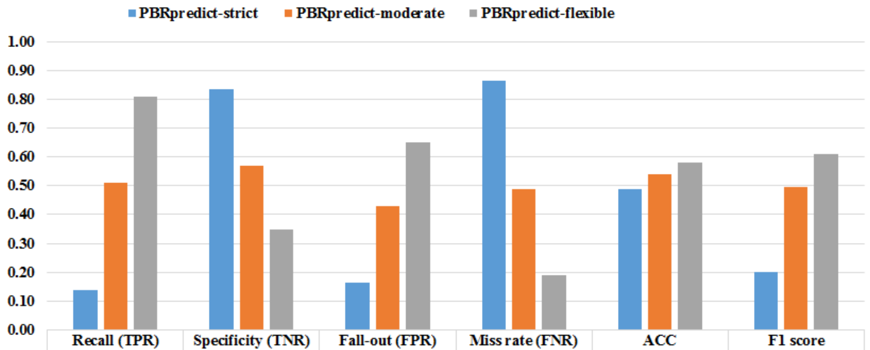


	Recall (TPR)	Specificity (TNR)	Fall-out (FPR)	Miss rate (FNR)	ACC	F1 score
▪ PBRpredict-strict	0.111	0.996	0.004	0.889	0.553	0.191
▪ PBRpredict-moderate	0.330	0.818	0.182	0.670	0.574	0.317
▪ PBRpredict-flexible	0.583	0.569	0.431	0.417	0.576	0.354

Fig 66. Performance of the PBRpredict-Suite models on VHS domains. The baser are grouped for the three predictors per metric and the score values are reported in the data table below the plot. The best values for each metric are highlighted using black boxes.

5.4.8.3 *CW Domain*

The CW domain recognizes the lysine methylation on the N-terminal histone tails. These post-translational modifications have key role in the tissue-specific gene expressions and chromatin regulations [398]. **Fig 67** shows the performances of the 3 models of PBRpredict-Suite in recognizing the residue patterns of this domain averaged over 7 chains.



	Recall (TPR)	Specificity (TNR)	Fall-out (FPR)	Miss rate (FNR)	ACC	F1 score
▪ PBRpredict-strict	0.137	0.836	0.164	0.863	0.487	0.201
▪ PBRpredict-moderate	0.511	0.570	0.430	0.489	0.541	0.497
▪ PBRpredict-flexible	0.810	0.349	0.651	0.190	0.580	0.610

Fig 67. Performance of the PBRpredict-Suite models on CW domains. The baser are grouped for the three predictors per metric and the score values are reported in the data table below the plot. The best values for each metric are highlighted using black boxes.

We observed a similar output for CW domain where the strict and the flexible models recognized the lowest and the highest percentage of the binding residues, respectively. On the other hand, the PBRpredict-moderate model resulted in a modest recall value.

The results of the above case-studies on the domains that were unseen by the PBRpredict-Suite models during training advocate the strength of the proposed models in locating potential peptide-binding sites within sequences for which the cognate domains are not known to the models. Therefore, the predictors, especially the moderate and the flexible models, can be useful in determining possible peptide-binding sites from protein sequence alone when no putative interaction information is known. The outputs then can be verified experimentally.

5.4.9 Case-Studies with PBRpredict-Suite models on Full-length Sequence

In this section, we study the full-length protein sequences with PBRpredict-Suite models. Here, we want to evaluate the ability of the proposed models in identifying potential peptide-binding residues in proteins for which no experimental or template structure is available. For this study, we chose the Gid4 protein. Recently, Chen et.al [399] discovered that the Gid4 subunit of the ubiquitin ligase GID in the yeast Saccharomyces cerevisiae targets the gluconeogenic enzymes, and recognizes the N-terminal proline (P) residue and the short 5-residue-long adjacent sequence motifs. The authors of the related study [399] identified such interactions through *in vitro* experiments with two-hybrid assays.

In this article, we computationally predict the potential residues in Gid4 protein that may mediate such interactions with gluconeogenic enzymes to degrade them and down-regulate the gluconeogenesis. We collected 3 Swiss-Prot reviewed proteins from UniProt, GID4_YEAST (ID: P38263), GID4_HUMAN (ID: Q8IVV7) and GID4_MOUSE (ID: Q9CPY6), and ran the PBRpredict-Suite models on these sequences to identify possible peptide-binding residues. The PBRpredict-strict model predicted only one residue as peptide-binding in GID4_YEAST and GID4_MOUSE, and found no binding residue in GID4_HUMAN. Therefore, we showed the predicted peptide-binding residues given by PBRpredict-moderate and flexible only.

5.4.9.1 *GID4_YEAST (UniProtKB – P38263)*

Fig 68(a) and **(b)** show the possible binding residues in *blue* identified by the PBRpredict-moderate and PBRpredict-flexible model in GID4_YEAST. The moderate and flexible model found 34 and 71 binding-residue respectively with a similar average confidence of 0.58 (mean probability values generated for the binding residues).

MINNPKVDSVAEKPKAVTSKQSEQAASPEPTPAPPVSRNQYPITFNLTSTAPFHLHDRHRYLQEQDLYKCASRDSLSSLQQLAHTPNGSTRKKYIVEDQSPYSSEN
PVIVTSSYNHTVCTNYLRPRMQFTGYQISGYKR**Y**QVTVNLKTVD**LP**K**KD**CTSLSPHLSGFLSIRGLTNQHPEIST**Y**FEAYAVNHKELG**FLSS**SWKDEPVLNEFKAT
D**Q**T**DLEHWIN**F**PSFRQLFLMSQ**K**N**G**L**NSTDDNGTTNAAKKLPPQQLPTTPSADAGNISRIFSQEKQFDNYLNERFIF**M**KWKEKFLVPDALLMEGVDGASYD
GFY**Y**IVHD**Q**VTGNIQGFYYHQDAEKFQQLELVPSLKNKVESSDCSFEFA

（a） PBRpredict−moderate annotation

MINNPKVDSVAEKPKAVTSKQSEQAASPEPTPAPPVSRNQYPITFNLTSTAPFHLHDRHRYLQEQDLYKCASRDSLSSLQQLAHTPNGSTRKKYIVEDQSPYSSE
NPVIVTSSYN**H**TVCTNYLRPRMQFTGYQISGYK**RY**QVTVNLKTV**DLPKKDC**TSLSPHLSG**F**LSIRGLTNQHPEIST**Y**F**EA**YAVN**HKEL**G**FLSSSW**KDEPVLNE
FKAT**DQTDLEHWINFPSFRQLFLMSQKNGL**NSTDDNGTTNAAKKLPPQQLPTTPSADAGNISRIFSQEKQFDNYLNER**FIFMKW**KEKFLVPDALLMEGVD
G**AS**Y**DG**F**YYI**VH**DQVTGN**IQGFYY**H**QDAEKFQQLELVPSLKNKVESSDC**S**FEFA

（b） PBRpredict−flexible annotation

Fig 68. GID4_YEAST protein annotated by PBRpredict moderate and flexible model. （a） and （b） show the prediction outputs of PBRpredict−moderate and flexible models that are mapped on to the sequence of GID4_YEAST. The predicted binding residues are marked in *blue*.

5.4.9.2 *GID4_HUMAN (UniProtKB – Q8IVV7)*

Fig 69(a) and **(b)** show the possible binding residues in *blue* identified by the PBRpredict-moderate and PBRpredict-flexible model in GID4_HUMAN. The moderate and flexible model found 8 and 39 binding-residue respectively with an average confidence of 0.58 and 0.55.

MCARGQVGRGTQLRTGRPCSQVPGSRWRPERLLRRQRAGGRPSRPHPARARPGLSLPATLLGSRAAAAVPLPLPPALAPGDPAMPVRTECPPPAGASAASAASLI
PPPPINTQQPGVATSLLYSGSKFRGHQKSKGNSYDVEVVLQHVD**T**GNS**Y**LCGYLKIKGLTEEYPTLTTFFEGEIISKK**H**PFLTRKWDADEDVD**R**KHWGKFLAFYQY
AKSFNSD**DF**DYEELKNGDYVFMRWKEQFLVPDHTIKDISGASFAGFYYICFQ**K**SAA**S**IEGYYYHRSSEWYQSLNLTHVPEHSAPIYEFR

（a） PBRpredict−moderate annotation

MCARGQVGRGTQLRTGRPCSQVPGSRWRPERLLRRQRAGGRPSRPHPARARPGLSLPATLLGSRAAAAVPLPLPPALAPGDPAMPVRTECPPPAGASAASAASLI
PPPPINTQQPGVATSLLYSGSKFRGHQKSKGNSYDVEVVLQ**H**VD**T**GNS**YLC**GYL**K**IKGLTEEYPTLTTF**F**EGEIISK**KHPF**LTRKWDADED**VDRK**H**WGK**F**L**A**F**
YQYAKSFNSD**DF**D**Y**EELKNGDYV**FM**RW**K**EQFLVPDHTIKDISGASFAGFY**Y**ICF**QKSA**A**SI**EGYYYH**R**SS**E**WYQSLNLT**H**VPEHSAPIYEFR

（b） PBRpredict−flexible annotation

Fig 69. GID4_HUMAN protein annotated by PBRpredict moderate and flexible model. （a） and （b） show the prediction outputs of PBRpredict−moderate and flexible models that are mapped on to the sequence of GID4_HUMAN. The predicted binding residues are marked in *blue*.

5.4.9.3 *GID4_MOUSE (UniProtKB – Q9CPY6)*

The potential binding residues in GID4_MOUSE, predicted by the PBRpredict-moderate and PBRpredict-flexible model, are shown in **Fig 70(a)** and **(b)**. The moderate and flexible model found 19 and 67 binding-residues respectively with an average confidence of 0.56 and 0.55.

MPVRTECPPPAGASTTSAASLIPPPPINTQQPGVATSLLYSGSKFRGHQKSKGNSYDVEVVLQ**H**VD**T**GNS**Y**LCGYLKIKGLTEEYPTLTTFFEGEIISKK**HP**FLTRK
WDADED**V**D**R**KHWGKFLA**F**Y**QY**AKSFNSD**DF**D**Y**EELKNGDYVF**M**RWKEQFLVPDHTIKDISGASFAGFYYICF**QK**SAA**SI**EGYYYHRSS**E**WYQSLNLTHVPEH
SAPIYEFR

（a） PBRpredict−moderate annotation

MPVRTECPPPAGASTTSAASLIPPPPIN**TQQ**PGVATSLLYSGSKFRGHQKSKGNSYDVEVVLQ**H**V**DT**G**N**S**YLCG**YL**K**IKGLTEEYPTLTTF**F**EGEIISK**KHPFL**T**R**
KW**DADE**D**VDRK**H**WGK**FL**AFYQYAK**SFNS**DDFDYE**ELKNGD**Y**V**FMRWKE**QFLVPDHTIKDISGAS**FAG**FYYIC**FQKSAASIE**G**Y**YYH**RS**S**E**W**Y**QSLNLT
HVPEHSAPIYEFR

（b） PBRpredict−flexible annotation

Fig 70. GID4_MOUSE protein annotated by PBRpredict moderate and flexible model. （a） and （b） show the prediction outputs of PBRpredict−moderate and flexible models that are mapped on to the sequence of GID4_MOUSE. The predicted binding residues are marked in *blue*.

The above case-studies show that the PBRpredict-Suite can be a useful tool in revealing the amino acid compositions that mediates the crucial interactions with peptide motifs from sequence alone when no structure is available. Such residue patterns can be further utilized for their cognate peptide identification. The above outcomes can further guide the experimental determination of the complex structure of these proteins by truncating the portion of the chain with potential peptide-binding sites.

5.5 Position Specific Binding Energy (PSBE)

While in Sections 5.2 to 5.4, we have elaborately discussed our contribution to identify the peptide-binding residues of proteins that play a major role in inducing a peptide-protein interaction. The short peptide motifs are usually part of disordered proteins or regions of proteins and undergo disorder-to-order transition only presence of an appropriate partner that can promote the binding [68, 129, 323]. In this section, we will focus on the other player of this peptide-protein interaction network, which are the residues on peptide surface that form the complex with peptide-recognition domains in partners. Specifically, we developed a residue-wise score to approximate the binding energy contribution ($\Delta\Delta G$), called position specific binding energy (PSBE) [14] to identify the hot spots on peptide surface that contribute most of the binding energy [321].

It is well-known that in protein-protein interfaces, the major contribution to binding energy is due to a small number of residues, which have been termed hot spot residues. This idea is established for protein-protein interactions as well [321]. Experimental identification of hot spot residues is primarily performed by alanine scanning. This process involves mutation of a target residue to alanine, and recording the resulting binding energy changes. If this mutation results in a marked drop in the binding energy, the residue is considered a hot spot [400]. Substitution with alanine removes all atoms in the side chain beyond the β-carbon. Furthermore, Alanine has relatively inert methyl functional group without contributing additional flexibility [401-403]. Mutation to glycine would also remove the side chain, but is not used since it can introduce unwanted conformational flexibility in the protein backbone [404]. The Binding free energy ($\Delta\Delta G$) [405] is computed using following equation:

$$\Delta\Delta G = \Delta G_{mutant} - \Delta G_{wild} \qquad (27)$$

Here, ΔG_{wild} and ΔG_{mutant} are the binding free energies upon complex formation of the wild-type and alanine-mutated proteins, respectively. We performed a similar computation of binding energy, however, the per-residues energy contribution was calculated as position specific estimated energy (PSEE) [11] from protein sequence only, a novel energy score proposed by us and discussed in **Chapter 4** of this thesis.

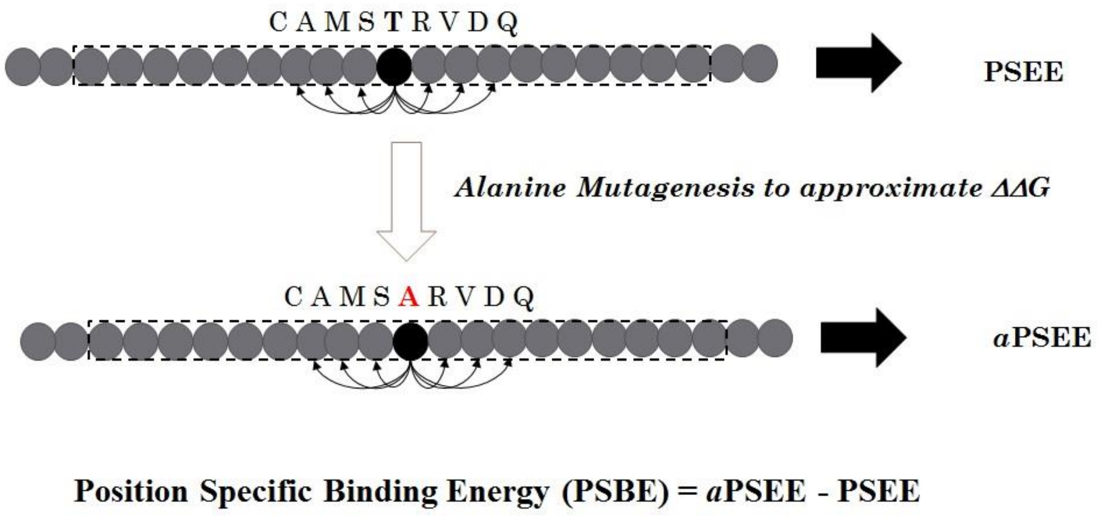

C A M S T R V D Q
PSEE
Alanine Mutagenesis to approximate ΔΔG
C A M S A R V D Q
aPSEE
Position Specific Binding Energy (PSBE) = aPSEE - PSEE

Fig 71. Definition of position specific binding energy (PSBE). PSEE is used to compute per-residue ΔG_{wild} and aPSEE, which is the recomputed PSEE value with target residue mutated using alanine, is used to compute per-residue ΔG_{mutant}. Then, the different between PSEE and aPSEE gives the PSBE, which can approximate $\Delta\Delta G$.

PSEE is an energy score that models the pairwise contact energy of amino acid residues within a neighborhood of the residue of interest to approximate its position specific energy contribution, however the contact energies are further weighted by the proportional burial of the neighborhood residues, which essentially captures the hydrophobic effect, the major force to determine the hydrophobic core of the 3D protein structure. The formulation of PSEE in given in Section 4.2, Equation 21. While per-residue PSEE can approximate the ΔG_{wild} of Equation 27, here, we recomputed PSEE with target residue mutated to alanine (A), shown in **Fig 71**. This modified value of PSEE is denoted as aPSEE, which approximates ΔG_{mutant}. After that, we quantify PSBE as the different between PSEE and aPSEE, illustrated in **Fig 59** to approximate $\Delta\Delta G$.

An analysis of amino acid propensities of peptide hot spot residues was carried out by London et.al [321] using 103 peptide-protein complex structures. In [321], the hot spot residues were identified by a computational alanine scan on each of the 103 complex structures using the Rosetta software [406]. Hot spots were defined as residues that upon mutation to alanine are identified to significantly decrease the binding energy, $\Delta\Delta G > 1 kcal/mol$ (Rosetta energy unit). Their observation suggested that the amino acids that are overrepresented in peptide hot spot residues are: Trp (W), Phe (F), Tyr (Y), Ile (I), and Leu (L).

We computed PSBE for the surface residues, interacting with a partner protein, of 724 peptide chains, culled from PDB. We generated the box-plots of binding-energies (PSBE) for 20 types of amino acids (AA) on the peptide interface while interacting with partners. The boxes are sorted according to the median of PSBE values, shown in **Fig 72**. The PSBE values Along with the average values highlighted in bigger fonts. We found that for five AAs (W, F, Y, I, L), the average PSBE values are higher than a particular threshold 0.5 within a range of -1.0 to $+1.0$, showing higher energy contribution in binding. This findings are consistent with the previous study [321], discussed above, which reports the AA overrepresentation in peptide hot-spot residues from structures of 103 peptide-protein complexes. This finding suggests that PSBE can approximate per-residue $\Delta\Delta G$ form sequence information only, thus, can recognize the peptide hot spots as well.

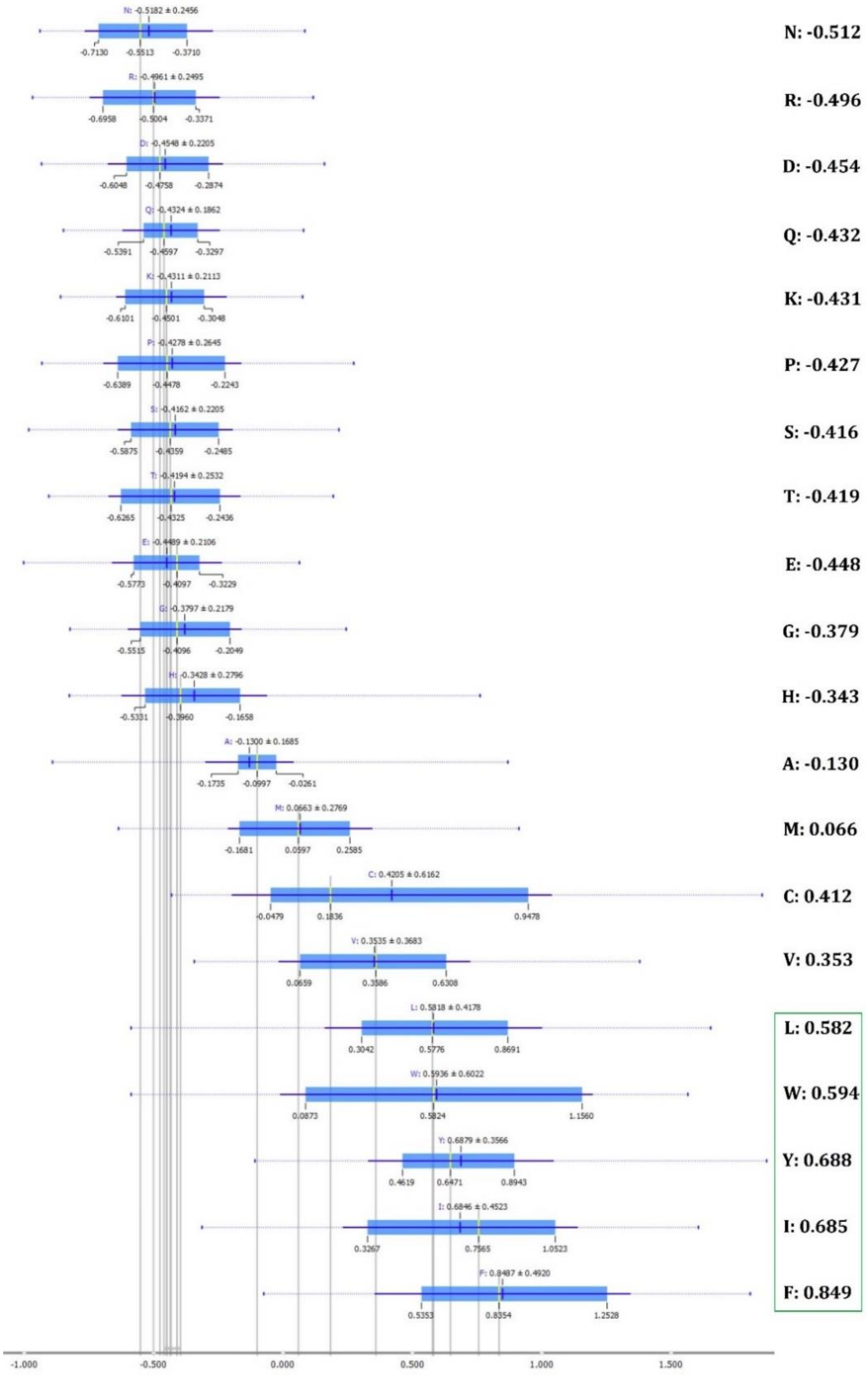


Fig 72. Box-plots of binding-energies (PSBE) for 20 types of amino acids (AA) on the peptide interface while interacting with partners. The boxes are sorted according to the median PSBE values. The boxes are sorted according to the median PSBE values and labeled by the average PSBE values. The AA with average PSBE > 0.5, likely to be in hot spots, are circled with a green box.

5.6 Summary and Conclusions

In this chapter, we describe the development of a suite of machine learning models to predict residues that can transiently interact with short flexible peptides in a complex and can result in induced-binding, using only protein sequence information. These residues are called peptide-binding residues and we call the proposed model, PBRpredict. For benchmarking purposes, we collected a new dataset of peptide-protein complexes. From that set, we extracted a non-redundant set of protein chains with wide range of peptide-recognition domains (PRDs), such as MHC, PDZ, WW, SH2, SH3, Polo-Box, Tudor, 14-3-3, PTB and others. Several of these PRDs have ability to recognize peptides with post-translational modifications (PTMs).

We labeled the residues that are in contact with peptide residues in a complex by measuring the distance between atomic coordinates. However, we carefully reduced the structural complexities involved in deciding the atomic association in a complex, like very short non-binding region of no greater than 3 residues can occur in between disjoint peptide-binding residues or regions due to the geometrical orientation of side chains and the associated steric clash. Specifically, we smoothed out those very short (1 – 3 residues long) non-binding residues in between binding residue stretches to generate a synthetic annotation of peptide-binding residues. Such synthetic annotation can effectively guide the predictor to learn about the local environment of the binding residues using only protein sequence information.

We provided a comprehensive set of residue-wise features to PBRpredict to characterize the inherent properties of the regions of interaction with peptides including chemical profile, evolutionary profile, local backbone profile and flexibility profile. In this work, we have integrated the other tools that we have developed and discussed in this thesis. We combined DisPredict (**Chapter 2**) that gave the disorder probabilities, REGAd^{3}p (**Chapter 3**) that provided the predicted accessible surface area and PSEE (**Chapter 4**) that scores the stability and energy contribution of the protein residues to generate features.

Moreover, to develop the predictor, we used stacked generalization or stacking, a popular method in modern machine learning community. Stacking operates combining the outputs by a set of base-learners of different types nonlinearly using a top-level meta-learner, unlike other ensemble techniques that use majority-voting (boosting) and weighted averaging (bagging). We investigated six different machine learning algorithms, support vector machine (SVM) with radial basis function as kernel, random forest (RDF), extremely-randomized tree (ET), gradient boosting classifier (GBC), k nearest neighbor (KNN) and bootstrap aggregation (BAG) to solve the problem of peptide-binding residue prediction. Through rigorous performance analysis, we found that SVM, GBC and KNN serve as an effective set of base-learners for this application. After that, we combined the predicted probabilities and the target residue features using logistic regression to build the final PBRpredict-Suite models. Therefore, this study can also be considered as a comprehensive review of machine learning algorithms to solve this challenging problem of proteomics.

After developing the initial PBRpredict model, called PBRpredict-strict, we carefully analyzed its applicability on full-length protein sequence for most of which the structures are unknown, and therefore, it is crucial to identify the potential peptide-binding sites from sequence alone. To make the predictor robust for recognizing peptide-binding residues on full-length sequence, unlike the structure-specific shorter sequence that were used to train the model, we developed two other predictors of similar framework using statistically derived relaxed threshold values, called PBRpredict-moderate and PBRpredict-flexible.

Due to the structural detail involved in protein-peptide interactions, most of the earlier predictions of peptide-binding regions in proteins have been done based on protein structure [353-355]. On the other hand, SPRINT [358] is a sequence-based method which is very recently introduced to predict protein-peptide binding sites using SVM, yields a MCC score of 0.248 on an independent test set of 146 chains. SPRINT is found to outperform two structure-based predictors [353, 354], as reported in [358]. To compare, the proposed predictor of this work, PBRpredict-strict scored 0.576 MCC in predicting peptide binding regions in proteins on the same dataset.

We further explored the biological relevance of the prediction output of SPRINT and PBRpredict-strict through case studies. We found that SPRINT overestimates the presence of peptide-binding residues throughout the full sequence of a receptor in complex, therefore results in a higher recall score. On the other hand, the outputs of PBRpredict seem to be biologically useful as it identifies few peptide-binding regions with contiguous residues, which is more relevant considering the intuitive number of regions that may possibly interact with a short peptide in a compact 3D structure.

Moreover, the PBRpredict-Suite models were found promising in locating peptide-binding sites in domains that were not seen by the models during training. In addition, the two relaxed models of the suite, PBRpredict-moderate and PBRpredict-flexible could detect the possible peptide-binding residues in GID4 protein which is recently found to bind to N-terminal peptide with proline (P) residue. To current date, no structure is available for GID4 protein, however, the proposed tool can guide the *in vitro* experiment with the potential sites only. Thus, PBRpredict can essentially be regarded an invaluable additional in the field of computational biology and worth further investigation in applications, like hot-spot region prediction and peptide binding-site prediction.

We have also studied short peptides involved in peptide-protein interaction. To characterize the hot spot residues on the peptide surface that mostly contribute to the binding energy, we extracted a per-residue energy score from protein sequence only using PSEE, discussed in Chapter 4. We performed a residue-wise alanine scanning within the protein sequence to recomputed PSEE and the induced gap in PSEE value after mutation is defined as position specific binding energy (PSBE). PSBE effectively identified the amino acids that are known as potential hot spots of peptide surface.

The ultimate goal of this study is to dig into the two-player complex process of induced folding between peptides and receptor proteins given sequence information only, of which, an accurate predictor of regions that induce such transient binding is a prerequisite. While the experimental screen is costly, computational methods with reasonable accuracy and relevance can be engaged to accelerate the process, thus increasing the productivity at reduced cost. With this predictor and PSBE score, a larger set of peptides and linear motifs of the human proteome can be scanned faster against the potential binding-inducing regions, and therefore can be linked to their potential binding partners. This study gives a step towards above-mentioned goal with the use of machine learning and proposes BIRpredict that can be potentially useful in understanding insights of peptide-protein binding.

Chapter 6

Conclusions

In this dissertation, we strived for the systematic discovery and characterization of new biological properties of proteomic data and computational modeling of several structural and interaction properties of proteins to better understand their roles in biological process. Our comprehensive research objective addressed applications in two disciplines:

(1) **Bioinformatics**, which includes development and implementation of tools using novel algorithms that enable efficient access and management of different types of biological information;

(2) **Computational Biology**, which involves analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures to learn new biology.

Besides development of machine learning based novel frameworks for protein sequence to (un)structure mapping and interaction prediction, we have devised new biological properties, such as Position Specific Estimated Energy (PSEE) and Binding Energy (PSBE). In our working procedure, we have established each of the tools or properties that we have developed as a software and have also applied them in exploring another challenge. In the last project of interaction prediction (described in **Chapter 5**), we have integrated all the tools that were developed earlier under this dissertation. Therefore, the overall flow of this thesis work and the outputs are interconnected yet each component are independently usable by the broader scientific community.

In this chapter, we first give a quick summary of the contributions and then present some directions for future research, and finish by some concluding remarks.

6.1 Summary

In the following, we summarize the contributions of this dissertation.

DisPredict: We have developed an optimized SVM based framework for predicting intrinsically disordered proteins (IDPs) or regions (IDRs) in proteins from sequence alone. In this research, we performed large scale proteomic data collection, purification and analysis from multiple sources such as PDB, DisProt and IDEAL. To develop the predictors, we implemented Support Vector Machine (SVM) with Radial Basis Function (RBF) kernel which is a well-known classifier to handle non-linearly separable classes. To best of our knowledge, we applied an optimized parameter set for in the first time in SVM-based disorder prediction. We carried out tuning of the cost of SVM and the mode of RBF, using Grid Search. Such optimized parameter set made the predictor competitive. Moreover, we applied two new features, Monogram and Bigram, in DisPredict to predict disorder for the first time.

In **DisPredict1.0** (version 1.0) [10], we directly used the probability values generated by SVM to classify ordered and disordered residue using a threshold of 0.5. In the next version (**DisPredict1.1**) [193], we have carried-out a post-processing of the probabilities generated by the SVM, which essentially averages the probability value of the target residue and those for a pre-defined number of residues on the either side to compute the final probability score. This final score was then binarized using a threshold value of 0.5. The development and benchmarking procedure of DisPredict1.0 and DisPredict1.1 are described in **Chapter 2**.

Through rigorous analysis, we have also investigated the correlation of ordered and disordered regions, reported in DisProt v6.02, in 2-dimensional feature space. This analysis highlighted the possible overlaps of the ordered or disordered regions in their feature space (relative exposure and coil probability). To meet the necessity of more efficient feature to characterize order versus disorder, we then developed a new residue-wise biological property, PSEE and used it in the feature space to predict intrinsically disordered proteins. The new predictor, **DisPredict2** [11] (discussed in **Chapter 4**) uses a similar framework that included optimized SVM with RBF kernel. Further, we quantified an optimized threshold value of 0.79 to finally segregate the two classes. DisPredict2 performed very well in comparison to several other state-of-the-art predictors including its predecessor. We have utilized the output probabilities of DisPredict1.0 and DisPredict2.0 as residue-wise feature, respectively, in predicting accessible surface area and peptide-binding residues of protein.

REGAd^{3}p: We developed a predictor of accessible surface area (ASA) of protein residues as real value from primary sequence. In this research work, we developed a new predictor paradigm, namely **REGAd^{3}p** [15], for real value prediction through Regularized Exact regression and Genetic Algorithm (GA). GA was used to optimize both Mean Absolute Error (MAE) and Pearson Correlation Coefficient (PCC). Further, the kernel of the exact regression was extended to 3-degree-polynomial as this kernel was found to be the best to predict ASA while testing with large datasets collected from PDB. However, the framework is general for a real-value prediction work and the kernel can be easily tuned for an application. Therefore, we believe this framework will be useful for similar prediction tasks.

We have applied my tool in several other applications of bioinformatics. We modeled the error between actual and predicted ASA in terms of Energy to discriminate native proteins from their decoys. We combined this ASA based energy linearly with the components of an existing energy function, 3DIGARS [16], using Genetic Algorithm to develop an improved versions, 3DIGARS2.0 [15]. The design process, evaluation and performance analysis of REGAd^{3}p for ASA prediction and the formulation of 3DIGARS2.0 energy function are discussed in **Chapter 3**.

We have further utilized the predicted ASA generated by REGAd^{3}p to quantify the relative exposure (or burial) of protein residue, which was used to devise a new property of protein residue, PSEE. Moreover, we have utilized the per-residue predicted ASA by REGAd^{3}p as residue-wise feature for predicting peptide-binding residues in protein sequence with peptide recognition domain.

PSEE and PSBE: We have devised a sequence-based feature for protein residues to characterize its ordered and disordered state, named Position Specific Estimated Energy (PSEE) [11]. Essentially, the PSEE resembles an energy-like quantity that scores the stability of a protein residue in its tertiary structure in terms of its contribution to the free-energy state of the full protein. The novel approach to quantify (PSEE) of a residue was based on two hypotheses: the contact of a target residue with different types of amino acid residues within a neighborhood region affects its tertiary structure, so as its energy contribution; and the pair-wise interaction between a target residue with its neighboring residues is further guided by the relative exposure (or burial) of the protein residues, which determines the hydrophobic effect, a major force that stabilizes protein fold. Therefore, we combined the pairwise interaction (or contact) energy between different types of amino acids with the residue's solvent accessibility to compute PSEE from protein sequence alone. Here, the pairwise interaction captures the sequential environment, whereas the predicted solvent accessibility, which is eventually used to compute relative burial of a residue, includes the hydrophobic effect and captures the respective structural environment in PSEE. The extraction of PSEE is described in **Chapter 4**.

We have performed a thorough analysis of PSEE values of ordered and disordered residues as well as regions of full DisProt v6.02 [100] database, which showed a reasonable gap of PSEE values in between these two classes. PSEE was also found effective in segregating different secondary structure type residues, beta (mostly stays in the core of protein), helix, and coil (mostly stays on the surface of protein), according to their stability. PSEE could further characterize hydrophobic and polar amino acid type residues by computing their constituent energies.

A feature that can be computed from the primary amino acid sequence of proteins is crucial in the process of inducing a machine learning model that is capable of accurately predicting 3D structural descriptor of protein. Computational tools for existing protein structure prediction problems require features, like PSEE that can capture the complexity of molecular level interactions. As an application, we have applied PSEE in improving DisPredict where DisPredict2 with PSEE outperformed the similar framework without PSEE. Development and benchmarking of DisPredict2 is also discussed in **Chapter 4**. Moreover, we have utilized the per-residue PSEE values as residue-wise feature for predicting peptide-binding residues from protein sequence.

The Position Specific Binding Energy (PSBE) [14] is a score that estimates the contribution of short peptide residues in binding with its partner in a complex. Using a sequence-based energy score PSEE, we have adopted a similar concept of alanine mutagenesis, which is usually performed with protein structure, to estimate binding energy. Specifically, we computed PSEE value of protein residue with its original amino acid at its position and after mutation of that amino acid to alanine. The induced gap in PSEE values before and after mutation is quantified as PSBE. The residues that contribute mostly in the binding energy are known as hot spot residues. When we computed PSBE values of the residues of a set of peptide sequences, we found that average PSBE values are greater than a specific threshold (0.5) for the experimentally identified hot spot prone amino acid residues. The extraction and analysis on PSBE are given in **Chapter 5**. Therefore, we believe PSBE has the potential to serve as a crucial sequence-based feature for peptide prediction, hot spot residue prediction and related tasks.

PBRpredict-Suite: Identification of peptide-binding residues in proteins with peptide-binding domain is the key for assembling peptide-protein interactomes and peptide-based therapeutic discovery. Under this dissertation, we have developed a framework, called PBRpredict [317] to predict peptide-binding residues of receptor proteins in peptide-protein complex from sequence alone. A dataset of protein complexes with wide range of peptide binding domains, like MHC I and II, PDZ, SH2, SH3, WW, 14-3-3, Chromo and Bromo, Polo-Box, PTB, enzyme inhibitor, was collected from PDB and mined to collect interaction information based on the atomic distances from peptide residues in the structure. To predict the peptide-binding residue, we encoded the protein sequence using a comprehensive set of sequence-based features including chemical and evolutionary profile, secondary structure, surface area and local backbone profile, flexibility and an energy based profile, we guide our predictor to learn about peptide-binding residues using model-stacking approach. In this step of this project, we have utilized the other tools that we have developed, DisPredict2, REGAd^{3}p and PSEE for feature generation.

To develop the predictor, we developed a stacking-based framework, a popular ensemble mechanism in modern machine learning community. In stacking, a set of base-learners are applied first and then the outputs are combined by a top-level meta-learner to generate final prediction, unlike other ensemble techniques that use majority-voting (boosting) and weighted averaging (bagging). We investigated six different machine learning algorithms, support vector machine (SVM) with radial basis function as kernel, random forest (RDF), extremely-randomized tree (ET), gradient boosting classifier (GBC), k nearest neighbor (KNN) and bootstrap aggregation (BAG) to solve the problem of peptide-binding residue prediction. Through rigorous performance analysis, we found that SVM, GBC and KNN serve as an effective set of base-learners for this application. After that, we combined the predicted probabilities and the target residue features using logistic regression to build the final PBRpredict model. This study can also be considered as a comprehensive review of machine learning algorithms to solve this challenging problem of proteomics.

Using three different sets of classification thresholds, that were statistically derived to trade-off between the true positive predictions and false positive predictions, for the base-level and meta-level learners, we established 3 different predictors under the PBRpredict-Suite (strict, moderate and flexible). We tested the models statistically and under biologically relevant case-studies, *i.e.*, with different length sequence and sequences with known and/or unknown domains. Altogether, the three models are found effective in different cases.

6.2 Future Scopes

Here, we briefly discuss the future scopes of the research that has been conducted under this dissertation. The possible future directions (but not limited to) are following.

In **Section 4.6** of Chapter 4, we have discussed about the possibility of noisy annotation of ordered and disordered regions in DisProt database. We have observed such possibility when we plotted the ordered and disordered regions in their two-dimensional feature space with features, like PSEE, relative exposure and coil probability (**Fig 39** and **40**). To explore further, we searched for possible structure of disordered proteins within Protein Data Bank and found that 155 IDPs or Proteins with IDRs of DisProt have structures in PDB. This finding complies with the existing phenomenon of disorder-to-order transition of short disordered regions through induced-binding in presence of an appropriate partner. Therefore, it will be interesting to extend our work to identify any class that may exist in between ordered and disordered state.

It is interesting to note that PSEE can also identify the existence of these short peptides within disordered proteins as the average PSEE values of short disordered regions were found to exist in between the PSEE values of ordered region and long disordered regions. Therefore, PSEE can potentially serve as a useful feature for short peptide region prediction. As an outcome of the discussion in **Section 4.3**, we can foresee that PSEE can also be utilized in predicting secondary structural features in disordered state which has not yet been robustly characterized.

The REGAd^{3}p framework, discussed in **Chapter 3**, is a generic real-value predictor that can be easily tuned for many future applications that involves real-value prediction. One of the major advantage of this framework comes from using exact regression technique, which can generate fast output. As another application, besides accessible surface area prediction, we have utilized a similar framework for backbone angle fluctuation prediction from protein sequence. We have performed preliminary simulations on tuning of kernel parameter and the results found were promising in comparison to an existing predictor [179] (results not shown in this thesis). This work is currently ongoing.

In our work of peptide-binding region prediction, discussed in **Chapter 5**, we have used stacking. To best of our knowledge, we have applied this ensemble technique for the first time in proteomics and the method was found promising in this challenging application. Therefore, it is worth trying this framework in other bioinformatics and computational biology applications instead of boosting and bagging, which are more popular in current days within the proteomics community because of their simplicity from the implementation point of view.

Moreover, the PBRpredict-Suite models provided prediction outputs that worth further *in vitro* experimentations. The models recognized potential peptide-binding sites in the Gid4 subunit of the ubiquitin ligase GID in the yeast *Saccharomyces cerevisiae* for which no structure is available to date. The corresponding subunit is experimentally found to interact with N-terminal Pro-peptide and degrade the gluconeogenesis-specific enzymes. We are currently collaborating the biology department of UNO to setup experiments on the potential binding sites and the cognate pro-peptides to understand more about the underlying mechanism of this interaction.

Finally, some possible future research directions using PSBE, discussed in **Chapter 5**, includes prediction of hot spot residue, short peptide region that are likely to undergo conformational transformation though coupled-binding with appropriate partners and participate in crucial signaling related activities within the cell.

6.3 Conclusions

Development of computational methods for large-scale as well as fast prediction, and analysis of biological data to study structure and functions of proteins from sequence only, such as the ones we developed in this dissertation, can help outgrow the ability of the experimental techniques. To keep pace with the current explosion of sequence-data, development of efficient and broadly applicable predictive algorithms with reasonable accuracy is critical to further progress. For some problems, the need of these computational efforts is essential. For instance, to understand the functions of proteins that are IDPs or have IDRs that do not adopt well-defined structure, however can change their states and fold through binding, and can perform important biological functions. Therefore, experimental investigation of IDPs/IDRs can reveal little information about their possible structures and functionalities. On the other hand, high throughput computational tool like DisPredict can provide a supplementary way for fast and large-scale IDPs/IDRs analysis.

To conclude, in order to pursue a predictive understanding of how structural information is encoded and evolved from sequence, development of computational frameworks based on solid mathematical foundations and algorithms, and statistical evaluation is imperative. Fast and efficient annotation of structural descriptors have significant implication to keep up with the rapid pace of biological research, and furthermore will contribute to applications in other science and engineering domains involving predictive understanding and reasoning. The published all the tools and datasets developed under this dissertation work are publicly available as open source. We hope that our contributions, e.g., DisPredict, REGAd^{3}p, PSEE, PBRpredict and PSBE will serve as useful tools for advancing the computing as well as biological sciences, particularly in proteomics research and application using machine learning.

Bibliography

[1] E. Fischer, "Einfluss der Configuration auf die Wirkung der Enzyme," *Berichte der deutschen chemischen Gesellschaft*, vol. 27, pp. 2985-2993, 1894.

[2] V. N. Uversky, J. R. Gillespie, and A. L. Fink, "Why are "natively unfolded" proteins unstructured under physiologic conditions?," *Proteins*, vol. 41, pp. 415 - 4127, 2000.

[3] C. B. Anfinsen, "The principles that govern the folding of protein chains," *Science*, vol. 181, pp. 223 - 230, 1973.

[4] G. A. Petsko and D. Ringe, *Protein structure and function*: New Science Press, 2004.

[5] P. E. Wright and H. J. Dyson, "Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm.," *Journal of Molecular Biology*, vol. 293, pp. 321 - 331, 1999.

[6] A. K. Dunker, J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, *et al.*, "Intrinsically disordered protein.," *J Mol Graph Model*, vol. 19, pp. 26 - 59, 2001.

[7] V. N. Uversky and A. K. Dunker, "Understanding protein non-folding.," *Biochimica Et Biophysica Acta (BBA) - Proteins And Proteomics*, vol. 1804, pp. 1231 - 1264, 2010.

[8] P. Radivojac, Z. Obradovic, D. K. Smith, G. Zhu, S. Vucetic, C. J. Brown, *et al.*, "Protein flexibility and intrinsic disorder.," *Protein Science*, vol. 10, pp. 71 - 80, 2004.

[9] S. Iqbal, D. Smith, and M. T. Hoque, "Accurate Identification of Disordered Protein Residues using Deep Neural Network.," in *4th Annual Conference on Computational Biology and Bioinformatics*, New Orleans, LA, 2016.

[10] S. Iqbal and M. T. Hoque, "DisPredict: A Predictor of Disordered Protein Using Optimized RBF Kernel.," *PloS One*, vol. 10, p. e0141551, 2015.

[11] S. Iqbal and M. T. Hoque, "Estimation of Position Specific Energy as a Feature of Protein Residues from Sequence Alone for Structural Classification," *PLoS ONE*, vol. 11, p. e0161452, 2016.

[12] S. Iqbal and M. T. Hoque. (2016). *DisPredict2*. Available: https://github.com/tamjidul/DisPredict2_PSEE

[13] S. Iqbal and M. T. Hoque, "Estimation of Free Energy Contribution of Protein Residues as Feature for Structure Prediction from Sequence," in *The Great Lakes Bioinformatics and the Canadian Computational Biology Conference (GLBIO/CCBC)*, Toronto, Canada, 2016.

[14] S. Iqbal and M. T. Hoque, "A Study of Disorder-to-Order Transition by Characterizing the Binding Partners using a Statistical Potential," *Biophysical Journal*, vol. 112, p. 209a, 2017.

[15] S. Iqbal, A. Mishra, and M. T. Hoque, "Improved prediction of accessible surface area results in efficient energy function application," *Journal of Theoretical Biology*, vol. 380, pp. 380-391, 2015.

[16] A. Mishra and M. T. Hoque, "Three-Dimensional Ideal Gas Reference State based Energy Function," *Current Bioinformatics, Bentham*, 2015.

[17] S. Iqbal and M. T. Hoque. (2015). *REGAd3p.* Available: http://cs.uno.edu/~tamjid/Software/REGAd3p/REGAd3p.tar.gz

[18] A. Mishra, S. Iqbal, and M. T. Hoque, "Discriminate protein decoys from native by using a scoring function based on ubiquitous Phi and Psi angles computed for all atom," *Journal of Theoretical Biology,* vol. 398, pp. 112 - 121, 2016.

[19] M. M. Babu, R. Lee, N. S. Groot, and J. Gsponer, "Intrinsically disordered proteins: regulation and disease.," *Current Opinion in Structural Biology,* vol. 21, pp. 432 - 440, 2011.

[20] P. Radivojac, L. M. Iakoucheva, C. J. Oldfield, Z. Obradovic, V. N. Uversky, and A. K. Dunker, "Intrinsic Disorder and Functional Proteomics.," *Biophysical Journal,* vol. 92, pp. 1493 - 2007, 2007.

[21] The UniProt Consortium, "UniProt: the universal protein knowledgebase," *Nucleic Acids Res.,* vol. 45, pp. D158 - D169, 2017.

[22] G. A. Petsko and D. Ringe, *Protein Structure and Function*, illustrated, reprint ed.: New Science Press, 2004.

[23] U. R. Lemieux and U. Spohr, "How Emil Fischer was led to the lock and key concept for enzyme specificity," *Adv. Carbohydr. Chem. Biochem.,* vol. 50, pp. 1 - 20, 1994.

[24] H. Wu, "Studies on denaturation of proteins. XIII. A theory of denaturation," *Chin J Physiol,* vol. 5, pp. 321-344, 1931.

[25] J. T. Edsall, "Hsien Wu and the first theory of protein denaturation (1931)," *Advances in protein chemistry,* vol. 46, pp. 1-5, 1995.

[26] L. Pauling and R. B. Corey, "Configurations of polypeptide chains with favored orientations around single bonds two new pleated sheets," *Proceedings of the National Academy of Sciences,* vol. 37, pp. 729-740, 1951.

[27] L. Pauling, R. B. Corey, and H. R. Branson, "The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain," *Proceedings of the National Academy of Sciences,* vol. 37, pp. 205-211, 1951.

[28] J. Kendrew, R. Dickerson, B. Strandberg, R. Hart, D. Davies, D. Phillips*, et al.*, "Structure of myoglobin: A three-dimensional Fourier synthesis at 2 A. resolution," *Nature,* vol. 185, pp. 422-427, 1960.

[29] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. Parrish, H. Wyckoff, and D. C. Phillips, "A three-dimensional model of the myoglobin molecule obtained by x-ray analysis," *Nature,* vol. 181, pp. 662-666, 1958.

[30] C. Blake, D. Koenig, G. Mair, A. North, D. Phillips, and V. Sarma, "Structure of hen egg-white lysozyme: a three-dimensional Fourier synthesis at 2 Å resolution," *Nature,* vol. 206, pp. 757-761, 1965.

[31] C. B. Anfinsen, E. Haber, M. Sela, and F. White, "The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain," *Proceedings of the National Academy of Sciences,* vol. 47, pp. 1309-1314, 1961.

[32] M. Anson and A. Mirsky, "On some general properties of proteins," *The Journal of general physiology,* vol. 9, p. 169, 1925.

[33] A. L. Fink, L. J. Calciano, Y. Goto, T. Kurotsu, and D. R. Palleros, "Classification of acid denaturation of proteins: intermediates and unfolded states," *Biochemistry,* vol. 33, pp. 12504-12511, 1994.

[34] Y. Goto, N. Takahashi, and A. L. Fink, "Mechanism of acid-induced folding of proteins," *Biochemistry,* vol. 29, pp. 3480-3488, 1990.

[35] V. N. Uversky, "Natively unfolded proteins: a point where biology waits for physics," *Protein science,* vol. 11, pp. 739-756, 2002.

[36] V. N. Uversky, "What does it mean to be natively unfolded?," *European Journal of Biochemistry,* vol. 269, pp. 2-12, 2002.

[37] N. R. Ziegler, "The Specificity of Serological Reactions," ed: American Public Health Association, 1936.

[38] L. Pauling, *A theory of the structure and process of formation of antibodies*: American Chemical Society, 1940.

[39] F. Karush, "Heterogeneity of the binding sites of bovine serum albumin1," *Journal of the American Chemical Society,* vol. 72, pp. 2705-2713, 1950.

[40] B. Jirgensons, "Classification of proteins according to conformation," *Die Makromolekulare Chemie,* vol. 91, pp. 74-86, 1966.

[41] R. Doolittle, "Structural aspects of the fibrinogen to fibrin conversion," *Advances in protein chemistry,* vol. 27, pp. 1-109, 1973.

[42] A. K. Dunker, M. M. Babu, E. Barbar, M. Blackledge, S. E. Bondos, Z. Dosztányi*, et al.*, "What's in a name? Why these proteins are intrinsically disordered: Why these proteins are intrinsically disordered," *Intrinsically Disordered Proteins,* vol. 1, p. e24157, 2013.

[43] V. N. Uversky, "Intrinsically disordered proteins from A to Z," *The international journal of biochemistry & cell biology,* vol. 43, pp. 1090-1103, 2011.

[44] V. N. Uversky, C. J. Oldfield, and A. K. Dunker, "Showing your ID : intrinsic disorder as an ID for recognition, regulation, and cell signaling.," *J. Mol. Recogn.,* vol. 18, pp. 343 - 348, 2005.

[45] R. Kriwacki, D. M. Mitrea, A. Follis, L. Iconaru, J. Cika, D. Ban*, et al.*, "Two Decades of IDPs; What have we Learned?," *Biophysical Journal,* vol. 112, pp. 12a – 13a, 2017.

[46] P. Tompa, "Intrinsically unstructured proteins.," *TRENDS in Biochemical Sciences,* vol. 10, pp. 527 - 533, 2002.

[47] J. Song, L.-W. Guo, H. Muradov, N. O. Artemyev, A. E. Ruoho, and J. L. Markley, "Intrinsically disordered γ-subunit of cGMP phosphodiesterase encodes functionally relevant transient secondary and tertiary structure," *Proceedings of the National Academy of Sciences,* vol. 105, pp. 1505-1510, 2008.

[48] F. D. Smith, S. L. Reichow, J. L. Esseltine, D. Shi, L. K. Langeberg, J. D. Scott*, et al.*, "Intrinsic disorder within an AKAP-protein kinase A complex guides local substrate phosphorylation," *Elife,* vol. 2, p. e01319, 2013.

[49] S. R. Sheftic, R. Page, and W. Peti, "Investigating the human Calcineurin Interaction Network using the πϕLxVP SLiM," *Scientific Reports,* vol. 6, 2016.

[50] *The PyMOL Molecular Graphics System, Version 1.2r3pre, Schrödinger, LLC*. Available: https://www.pymol.org/

[51] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features " *Biopolymers,* vol. 22, pp. 2577 - 2637, 1983.

[52] S. Vucetic, C. J. Brown, A. K. Dunker, and Z. Obradovic, "Flavors of protein disorder.," *Proteins: Structure, Function, and Bioinformatics,* vol. 52, pp. 573 - 584, 2003.

[53] A. K. Dunker and Z. Obradovic, "The protein trinity--linking function and disorder.," *Nat Biotechnol,* vol. 19, pp. 805 - 806, 2001.

[54] J. Gu and V. Hilser, "The significance and impacts of protein disorder and conformational variants," *Structural bioinformatics,* pp. 939-962, 2009.

[55] V. Uversky, "A multiparametric approach to studies of self-organization of globular proteins," *Biochemistry. Biokhimiia,* vol. 64, p. 250, 1999.

[56] V. Uversky, "Protein folding revisited. A polypeptide chain at the folding–misfolding–nonfolding cross-roads: which way to go?," *Cellular and Molecular Life Sciences CMLS,* vol. 60, pp. 1852-1871, 2003.

[57] G. W. Daughdrill, G. J. Pielak, V. N. Uversky, M. S. Cortese, and A. K. Dunker, "Natively disordered proteins," *Protein folding handbook,* pp. 275-357, 2005.

[58] S. L. Crick, M. Jayaraman, C. Frieden, R. Wetzel, and R. V. Pappu, "Fluorescence correlation spectroscopy shows that monomeric polyglutamine molecules form collapsed structures in aqueous solutions," *Proceedings of the National Academy of Sciences,* vol. 103, pp. 16764-16769, 2006.

[59] A. H. Mao, S. L. Crick, A. Vitalis, C. L. Chicoine, and R. V. Pappu, "Net charge per residue modulates conformational ensembles of intrinsically disordered proteins," *Proceedings of the National Academy of Sciences,* vol. 107, pp. 8183-8188, 2010.

[60] H. T. Tran, A. Mao, and R. V. Pappu, "Role of Backbone− Solvent Interactions in Determining Conformational Equilibria of Intrinsically Disordered Proteins," *Journal of the American Chemical Society,* vol. 130, pp. 7380-7392, 2008.

[61] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig*, et al.*, "The protein data bank," *Nucleic Acids Res,* vol. 28, pp. 235 - 242, 2000.

[62] Z. Obradovic, K. Peng, S. Vucetic, P. Radivojac, C. J. Brown, and A. K. Dunker, "Predicting intrinsic disorder from amino acid sequence.," *Proteins,* vol. 53, pp. 566 - 572, 2003.

[63] B. Xue, R. L. Dunbrack, R. W. Williams, A. K. Dunker, and V. N. Uversky, "PONDR-FIT: A Meta-Predictor of Intrinsically Disordered Amino Acids.," *Biochim Biophys Acta,* vol. 1804, pp. 996 - 1001, 2010.

[64] A. K. Dunker, I. Silman, V. N. Uversky, and J. L. Sussman, "Function and structure of inherently disordered proteins," *Current opinion in structural biology,* vol. 18, pp. 756-764, 2008.

[65] M. Marín, V. N. Uversky, and T. Ott, "Intrinsic disorder in pathogen effectors: protein flexibility as an evolutionary hallmark in a molecular arms race," *The Plant Cell,* vol. 25, pp. 3153-3157, 2013.

[66] X. Sun, E. H. Rikkerink, W. T. Jones, and V. N. Uversky, "Multifarious roles of intrinsic disorder in proteins illustrate its broad impact on plant biology," *The Plant Cell,* vol. 25, pp. 38-55, 2013.

[67] A. K. Dunker, P. Romero, Z. Obradovic, E. C. Garner, and C. J. Brown, "Intrinsic protein disorder in complete genomes," *Genome Informatics,* vol. 11, pp. 161-171, 2000.

[68] J. Yan, A. K. Dunker, V. N. Uversky, and L. Kurgan, "Molecular recognition features (MoRFs) in three domains of life," *Molecular BioSystems,* 2016.

[69] H. J. Dyson and P. E. Wright, "Coupling of folding and binding for unstructured proteins.," *Current opinion in structural biology,* vol. 12, pp. 54 - 60, 2002.

[70] A. K. Dunker, C. J. Brown, and Z. Obradovic, "Identification and functions of usefully disordered proteins.," *Adv. Protein Chem,* vol. 62, pp. 25 - 49, 2002.

[71] P. B. Sigler, "Transcriptional activation. Acid blobs and negative noodles," ed, 1988.

[72] L. M. Iakoucheva, C. J. Brown, J. D. Lawson, Z. Obradović, and A. K. Dunker, "Intrinsic disorder in cell-signaling and cancer-associated proteins," *Journal of molecular biology,* vol. 323, pp. 573-584, 2002.

[73] A. K. Dunker, M. S. Cortese, P. Romero, L. M. Iakoucheva, and V. N. Uversky, "Flexible nets. The roles of intrinsic disorder in protein interaction networks.," *Febs Journal,* vol. 272, pp. 5129-5148, 2005.

[74] R. W. Kriwacki, L. Hengst, L. Tennant, S. I. Reed, and P. E. Wright, "Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity," *Proceedings of the National Academy of Sciences,* vol. 93, pp. 11504-11509, 1996.

[75] A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva, and Z. Obradovic, "Intrinsic disorder and protein function.," *Biochemistry,* vol. 41, pp. 6573 - 6582, 2002.

[76] P. Tompa, "The interplay between structure and function in intrinsically unstructured proteins," *FEBS letters,* vol. 579, pp. 3346-3354, 2005.

[77] M. Fuxreiter, I. Simon, and S. Bondos, "Dynamic protein–DNA recognition: beyond what can be seen," *Trends in biochemical sciences,* vol. 36, pp. 415-423, 2011.

[78] S. Vucetic, H. Xie, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, Z. Obradovic*, et al.*, "Functional anthology of intrinsic disorder. 2. Cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions," *Journal of proteome research,* vol. 6, pp. 1899-1916, 2007.

[79] H. Xie, S. Vucetic, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, Z. Obradovic*, et al.*, "Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins," *Journal of proteome research,* vol. 6, pp. 1917-1932, 2007.

[80] H. Xie, S. Vucetic, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, V. N. Uversky*, et al.*, "Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions," *Journal of proteome research,* vol. 6, pp. 1882-1898, 2007.

[81] P. R. Romero, S. Zaidi, Y. Y. Fang, V. N. Uversky, P. Radivojac, C. J. Oldfield*, et al.*, "Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms," *Proceedings of the National Academy of Sciences,* vol. 103, pp. 8390-8395, 2006.

[82] B. Xue, A. K. Dunker, and V. N. Uversky, "The Roles of Intrinsic Disorder in Orchestrating the Wnt-Pathway.," *Journal of Biomolecular Structure and Dynamics,* vol. 29, pp. 843 - 861, 2012.

[83] V. N. Uversky, C. J. Oldfield, and A. K. Dunker, "Intrinsically disordered proteins in human diseases: introducing the D2 concept," *Annu. Rev. Biophys.,* vol. 37, pp. 215-246, 2008.

[84] P. Kulkarni, K. Rajagopalan, D. Yeater, and R. H. Getzenberg, "Protein folding and the order/disorder paradox.," *J Cell Biochem,* vol. 112, pp. 1949 - 1952, 2011.

[85] V. N. Uversky, C. J. Oldfield, U. Midic, H. Xie, B. Xue, S. Vucetic*, et al.*, "Unfoldomics of human diseases: linking protein intrinsic disorder with diseases.," *BMC Genomics,* vol. 10, pp. S1 - S7, 2009.

[86] Y. Cheng, T. LeGall, C. J. Oldfield, J. P. Mueller, Y.-Y. J. Van, P. Romero*, et al.*, "Rational drug design via intrinsically disordered protein.," *Trends Biotechnol.,* vol. 24, pp. 435 - 442, 2006.

[87] V. N. Uversky, "(Intrinsically disordered) splice variants in the proteome: implications for novel drug discovery," *Genes & Genomics,* vol. 38, pp. 577-594, 2016.

[88] H. J. Dyson, "Making sense of intrinsically disordered proteins," *Biophys J,* vol. 110, pp. 1013-1016, 2016.

[89] D. Ringe and G. A. Petsko, "Study of protein dynamics by X-ray diffraction," *Methods in enzymology,* vol. 131, pp. 389-433, 1986.

[90] H. J. Dyson and P. Ewright, "Insights into the structure and dynamics of unfolded proteins from nuclear magnetic resonance," *Advances in protein chemistry,* vol. 62, pp. 311-340, 2002.

[91] H. J. Dyson and P. E. Wright, "Unfolded proteins and protein folding studied by NMR," *Chemical reviews,* vol. 104, pp. 3607-3622, 2004.

[92] C. Bracken, L. M. Iakoucheva, P. R. Romero, and A. K. Dunker, "Combining prediction, computation and experiment for the characterization of protein disorder," *Current opinion in structural biology,* vol. 14, pp. 570-576, 2004.

[93] S. Kosol, S. Contreras-Martos, C. Cedeño, and P. Tompa, "Structural characterization of intrinsically disordered proteins by NMR spectroscopy," *Molecules,* vol. 18, pp. 10802-10828, 2013.

[94] G. D. Fasman, *Circular dichroism and the conformational analysis of biomolecules*: Springer Science & Business Media, 2013.

[95] K. Ikeda, K. HAMAGUCHI, M. YAMAMOTO, and T. IKENAKA, "Circular dichroism and optical rotatory dispersion of trypsin inhibitors," *The Journal of Biochemistry,* vol. 63, pp. 521-531, 1968.

[96] S. W. Provencher and J. Gloeckner, "Estimation of globular protein secondary structure from circular dichroism," *Biochemistry,* vol. 20, pp. 33-37, 1981.

[97] R. W. Woody, "Circular dichroism," *Methods in enzymology,* vol. 246, pp. 34-71, 1995.

[98] V. N. Uversky, "A multiparametric approach to studies of self-organization of globular proteins," *Biochemistry. Biokhimiia,* vol. 64, pp. 250-266, 1999.

[99] V. Receveur-Bréchot, J. M. Bourhis, V. N. Uversky, B. Canard, and S. Longhi, "Assessing protein disorder and induced folding," *Proteins: Structure, Function, and Bioinformatics,* vol. 62, pp. 24-45, 2006.

[100] M. Sickmeier, J. A. Hamilton, T. LeGall, V. Vacic, M. S. Cortese, A. Tantos*, et al.*, "DisProt: the Database of Disordered Proteins.," *Nucleic Acids Res,* vol. 35, pp. 786 - 793, 2007.

[101] D. Piovesan, F. Tabaro, I. Mičetić, M. Necci, F. Quaglia, C. J. Oldfield*, et al.*, "DisProt 7.0: a major update of the database of disordered proteins," *Nucleic acids research,* vol. 45, pp. D219-D227, 2017.

[102] S. Fukuchi, T. Amemiya, S. Sakamoto, Y. Nobe, K. Hosoda, Y. Kado, *et al.*, "IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners.," *Nucleic Acids Res.,* vol. 42, pp. D320 - D325, 2014.

[103] S. Fukuchi, S. Sakamoto, Y. Nobe, S. D. Murakami, T. Amemiya, K. Hosoda, *et al.*, "IDEAL: Intrinsically Disordered proteins with Extensive Annotations and Literature.," *Nucleic Acids Res.,* vol. 40, pp. D507 - D511, 2012.

[104] K. D. Pruitt, T. Tatusova, W. Klimke, and D. R. Maglott, "NCBI Reference Sequences: current status, policy and new initiatives.," *Nucleic Acids Res.,* vol. 37, pp. D32 - D35, 2009.

[105] T. D. Domenico, I. Walsh, A. J. M. Martin, and S. C. E. Tosatto, "MobiDB: a comprehensive database of intrinsic protein disorder annotations.," *Bioinformatics,* vol. 28, pp. 2080 - 2081, 2012.

[106] E. Potenza, T. D. Domenico, I. Walsh, and S. C. E. Tosatto, "MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins.," *Nucl. Acids Res.,* 2014.

[107] R. J. P. Williams, "The conformation properties of proteins in solution," *Biological Reviews,* vol. 54, pp. 389-437, 1979.

[108] B. He, K. Wang, Y. Liu, B. Xue, V. N. Uversky, and A. K. Dunker, "Predicting intrinsic disorder in proteins: an overview," *Cell research,* vol. 19, pp. 929-949, 2009.

[109] P. Romero, Z. Obradovic, C. Kissinger, J. Villafranca, and A. Dunker, "Identifying disordered regions in proteins from amino acid sequence," in *Neural Networks, 1997., International Conference on*, 1997, pp. 90-95.

[110] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics,* vol. 5, pp. 115-133, 1943.

[111] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning,* vol. 20, pp. 273-297, 1995.

[112] L. Breiman, "Random forests," *Machine learning,* vol. 45, pp. 5-32, 2001.

[113] P. Romero, Z. Obradovic, X. Li, E. C. Garner, C. J. Brown, and A. K. Dunker, "Sequence complexity of disordered protein," *Proteins: Structure, Function, and Bioinformatics,* vol. 42, pp. 38-48, 2001.

[114] Z. Obradovic, K. Peng, S. Vucetic, P. Radivojac, and A. K. Dunker, "Exploiting heterogeneous sequence properties improves prediction of protein disorder," *Proteins: Structure, Function, and Bioinformatics,* vol. 61, pp. 176-182, 2005.

[115] K. Peng, S. Vucetic, P. Radivojac, C. J. Brown, A. K. Dunker, and Z. Obradovic, "Optimizing long intrinsic disorder predictors with protein evolutionary information," *Journal of bioinformatics and computational biology,* vol. 3, pp. 35-60, 2005.

[116] K. Peng, P. Radivojac, S. Vucetic, A. K. Dunker, and Z. Obradovic, "Length-dependent prediction of protein intrinsic disorder.," *BMC Bioinformatics,* vol. 7, p. 208, 2006.

[117] X. Li, P. Romero, M. Rani, A. K. Dunker, and Z. Obradovic, "Predicting protein disorder for N-, C-and internal regions," *Genome Informatics,* vol. 10, pp. 30-40, 1999.

[118] R. Linding, L. J. Jensen, F. Diella, P. Bork, T. J. Gibson, and R. B. Russell, "Protein disorder prediction: implications for structural proteomics.," *Structure,* vol. 11, pp. 1453 - 1459, 2003.

[119] Z. R. Yang, R. Thomson, P. McNeil, and R. M. Esnouf, "RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins.," *Bioinformatics,* vol. 21 pp. 3369 - 3376, 2005.

[120] A. Schlessingera, J. Liu, and B. Rost, "Natively Unstructured Loops Differ from Other Loops.," *Bioinformatics,* vol. 3, pp. e140 - e151, 2007.

[121] J. Cheng, M. J. Sweredoski, and P. Baldi, "Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data.," *Data Mining and Knowledge Discovery,* vol. 11, pp. 213 - 222, 2005.

[122] I. Walsh, A. J. Martin, T. Di Domenico, and S. C. Tosatto, "ESpritz: accurate and fast prediction of protein disorder," *Bioinformatics,* vol. 28, pp. 503-509, 2012.

[123] T. Zhang, E. Faraggi, B. Xue, A. K. Dunker, V. N. Uversky, and Y. Zhou, "SPINE-D: accurate prediction of short and long disordered regions by a single neural-network based method.," *J Biomol Struct Dyn,* vol. 29, pp. 799 - 813, 2012.

[124] J. Eickholt and J. Cheng, "DNdisorder: predicting protein disorder using boosting and deep networks," *BMC bioinformatics,* vol. 14, p. 88, 2013.

[125] L. Deng and D. Yu, "Deep learning: methods and applications," *Foundations and Trends® in Signal Processing,* vol. 7, pp. 197-387, 2014.

[126] J. J. Ward, L. J. McGuffin, K. Bryson, B. F. Buxton, and D. T. Jones, "The DISOPRED server for the prediction of protein disorder.," *Bioinformatics,* vol. 20, pp. 2138 - 2139, 2004.

[127] D. T. Jones and J. J. Ward, "Prediction of disordered regions in proteins from position specific score matrices.," *Proteins,* vol. 53, pp. 573 - 578, 2003.

[128] J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton, and D. T. Jones, "Prediction and functional analysis of native disorder in proteins from the three kingdoms of life," *Journal of molecular biology,* vol. 337, pp. 635-645, 2004.

[129] D. T. Jones and D. Cozzetto, "DISOPRED3: precise disordered region predictions with annotated protein-binding activity," *Bioinformatics,* vol. 31, pp. 857-863, 2015.

[130] A. Vullo, O. Bortolami, G. Pollastri, and S. C. Tosatto, "Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines.," *Nucleic Acids Res.,* vol. 34, pp. W164 - W168, 2006.

[131] S. Hirose, K. Shimizu, S. Kanai, Y. Kuroda, and T. Noguchi, "POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions," *Bioinformatics,* vol. 23, pp. 2046-2053, 2007.

[132] K. Shimizu, S. Hirose, and T. Noguchi, "POODLE-S: web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a position-specific scoring matrix," *Bioinformatics,* vol. 23, pp. 2337-2338, 2007.

[133] K. Shimizu, Y. Muraoka, S. Hirose, K. Tomii, and T. Noguchi, "Predicting mostly disordered proteins by using structure-unknown protein data," *BMC bioinformatics,* vol. 8, p. 78, 2007.

[134] T. Joachims, "Transductive learning via spectral graph partitioning," in *ICML,* 2003, pp. 290-297.

[135] C. T. Su, C. Y. Chen, and Y. Y. Ou, "Protein disorder prediction by condensed PSSM considering propensity for order or disorder.," *BMC Bioinformatics,* vol. 7, pp. 319 - 334, 2006.

[136] C. T. Su, C. Y. Chen, and C. M. Hsu, "iPDA: integrated protein disorder analyzer.," *Nucleic Acids Res,* vol. 35, pp. W465 -- W472, 2007.

[137] E. A. Weathers, M. E. Paulaitis, T. B. Woolf, and J. H. Hoh, "Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein," *FEBS letters,* vol. 576, pp. 348-352, 2004.

[138] J. Yan, M. J. Mizianty, P. L. Filipow, V. N. Uversky, and L. Kurgan, "RAPID: fast and accurate sequence-based prediction of intrinsic disorder content on proteomic scale," *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics,* vol. 1834, pp. 1671-1680, 2013.

[139] A. Bulashevska and R. Eils, "Using Bayesian multinomial classifier to predict whether a given protein sequence is intrinsically disordered," *Journal of theoretical biology,* vol. 254, pp. 799-803, 2008.

[140] L. Wang and U. H. Sauer, "OnD-CRF: predicting order and disorder in proteins using [corrected] conditional random fields.," *Bioinformatics,* vol. 24, pp. 1401 - 1402, 2008.

[141] M. J. Mizianty, T. Zhang, B. Xue, Y. Zhou, A. K. Dunker, V. N. Uversky*, et al.*, "In-silico prediction of disorder content using hybrid sequence representation," *BMC bioinformatics,* vol. 12, p. 245, 2011.

[142] B. Xue, W. L. Hsu, J. H. Lee, H. Lu, A. K. Dunker, and V. N. Uversky, "SPA: Short peptide analyzer of intrinsic disorder status of short peptides," *Genes to Cells,* vol. 15, pp. 635-646, 2010.

[143] H. Ali, S. Urolagin, Ö. Gurarslan, and M. Vihinen, "Performance of protein disorder prediction programs on amino acid substitutions," *Human mutation,* vol. 35, pp. 794-804, 2014.

[144] R. Linding, R. B. Russell, V. Neduva, and T. J. Gibson, "GlobPlot: Exploring protein sequences for globularity and disorder.," *Nucleic Acids Res,* vol. 31, pp. 3701 - 3708, 2003.

[145] A. Campen, R. M. Williams, C. J. Brown, J. Meng, V. N. Uversky, and A. K. Dunker, "TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder," *Protein and peptide letters,* vol. 15, p. 956, 2008.

[146] O. V. Galzitskaya, S. O. Garbuzynskiy, and M. Y. Lobanov, "FoldUnfold: web server for the prediction of disordered regions in protein chain," *Bioinformatics,* vol. 22, pp. 2948-2949, 2006.

[147] K. Coeytaux and A. Poupon, "Prediction of unfolded segments in a protein sequence based on amino acid composition," *Bioinformatics,* vol. 21, pp. 1891-1900, 2005.

[148] Z. Dosztányi, V. Csizmok, P. T. and, and I. Simon, "IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content," *Bioinformatics,* vol. 21, pp. 3433-3434, 2005.

[149] Z. Dosztanyi, V. Csizmok, P. Tompa, and I. Simon, "The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins," *Journal of molecular biology,* vol. 347, pp. 827-839, 2005.

[150] J. Prilusky, C. E. Felder, T. Z.-B. -Mordehai, E. H. Rydberg, O. Man, J. S. Beckmann*, et al.*, "FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded.," *Bioinformatics,* vol. 21, pp. 3435 - 3438, 2005.

[151] J. C. Wootton, "Non-globular domains in protein sequences: automated segmentation using complexity measures," *Computers & chemistry,* vol. 18, pp. 269-285, 1994.

[152] A. Schlessinger, M. Punta, and B. Rost, "Natively unstructured regions in proteins identified from contact predictions.," *Bioinformatics,* vol. 23, pp. 2376 - 2384, 2007.

[153] L. J. McGuffin, "Intrinsic disorder prediction from the analysis of multiple protein fold recognition models.," *Bioinformatics,* vol. 24, pp. 1798 - 1804, 2008.

[154] M. Y. Lobanov, I. V. Sokolovskiy, and O. V. Galzitskaya, "IsUnstruct: prediction of the residue status to be ordered or disordered in the protein chain by a method based on the Ising model," *Journal of Biomolecular Structure and Dynamics,* vol. 31, pp. 1034-1043, 2013.

[155] X. Deng, J. Eickholt, and J. Cheng, "PreDisorder: ab initio sequence-based prediction of protein disordered regions.," *BMC Bioinformatics,* vol. 10, pp. 436 - 441, 2009.

[156] J. Cheng, J. Li, Z. Wang, J. Eickholt, and X. Deng, "The MULTICOM toolbox for protein structure prediction," *BMC bioinformatics,* vol. 13, p. 65, 2012.

[157] S. Hirose, K. Shimizu, and T. Noguchi, "POODLE-I: disordered region prediction by integrating POODLE series and structural information predictors based on a workflow approach," *In silico biology,* vol. 10, pp. 185-191, 2010.

[158] L. P. Kozlowski and J. M. Bujnicki, "MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins," *BMC bioinformatics,* vol. 13, p. 111, 2012.

[159] Y. J. Huang, T. B. Acton, and G. T. Montelione, "DisMeta: a meta server for construct design and optimization," *Structural Genomics: General Applications,* pp. 3-16, 2014.

[160] I. Walsh, A. J. M. Martin, T. D. Domenico, A. Vullo, G. Pollastri, and S. C. E. Tosatto, "CSpritz: accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs," *Nucleic Acids Res.,* vol. 39, pp. W190 - W196, 2011.

[161] T. Ishida and K. Kinoshita, "Prediction of disordered regions in proteins based on the meta approach," *Bioinformatics,* vol. 24, pp. 1344-1348, 2008.

[162] A. Schlessinger, M. Punta, G. Yachdav, L. Kajan, and B. Rost, "Improved Disorder Prediction by Combination of Orthogonal Approaches.," *PLOS ONE,* vol. 4, pp. e4433 - e4442, 2009.

[163] A. Schlessinger, G. Yachdav, and B. Rost, "PROFbval: predict flexible and rigid residues in proteins," *Bioinformatics,* vol. 22, pp. 891-893, 2006.

[164] M. J. Mizianty, W. Stach, K. Chen, K. D. Kedarisetti, F. M. Disfani, and L. Kurgan, "Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources.," *Bioinformatics,* vol. 26, pp. i489 - i496, 2010.

[165] M. J. Mizianty, Z. Peng, and L. Kurgan, "MFDp2: Accurate predictor of disorder in proteins by fusion of disorder probabilities, content and profiles.," *Intrinsically Disordered Proteins,* vol. 1, p. e24428, 2013.

[166] B. Monastyrskyy, A. Kryshtafovych, J. Moult, A. Tramontano, and K. Fidelis, "Assessment of protein disorder region predictions in CASP10.," *Proteins,* vol. 82, pp. 127 - 137, 2014.

[167] B. Monastyrskyy, K. Fidelis, J. Moult, A. Tramontano, and A. Kryshtafovych, "Evaluation of disorder predictions in CASP9," *Proteins: Structure, Function, and Bioinformatics,* vol. 79, pp. 107-118, 2011.

[168] O. Noivirt-Brik, J. Prilusky, and J. L. Sussman, "Assessment of disorder predictions in CASP8," *Proteins: Structure, Function, and Bioinformatics,* vol. 77, pp. 210-216, 2009.

[169] L. Bordoli, F. Kiefer, and T. Schwede, "Assessment of disorder predictions in CASP7," *Proteins: Structure, Function, and Bioinformatics,* vol. 69, pp. 129-136, 2007.

[170] Y. Jin and R. L. Dunbrack, "Assessment of disorder predictions in CASP6," *Proteins: Structure, Function, and Bioinformatics,* vol. 61, pp. 167-175, 2005.

[171] E. Melamud and J. Moult, "Evaluation of disorder predictions in CASP5," *Proteins: Structure, Function, and Bioinformatics,* vol. 53, pp. 561-565, 2003.

[172] F. L. Sirota, H.-S. Ooi, T. Gattermayer, G. Schneider, F. Eisenhaber, and S. Maurer-Stroh, "Parameterization of disorder predictors for large-scale applications requiring high specificity by using an extended benchmark dataset," *BMC genomics,* vol. 11, p. S15, 2010.

[173] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool.," *J Mol Biol.,* vol. 215, pp. 403 - 410, 1990.

[174] J. Meiler, M. Muller, A. Zeidler, and F. Schmäschke, "Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks," *J Mol Model,* vol. 7, pp. 360 - 369, 2001.

[175] E. Faraggi, T. Zhang, Y. Yang, L. Kurgan, and Y. Zhou, "SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles," *J Comput Chem,* vol. 33, pp. 259 - 267, 2012.

[176] A. Sharma, J. Lyons, A. Dehzangi, and K. K. Paliwal, "A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition," *Journal of theoretical biology,* vol. 320, pp. 41-46, 2013.

[177] A. Sharma, A. Dehzangi, J. Lyons, S. Imoto, S. Miyano, K. Nakai, *et al.*, "Evaluation of sequence features from intrinsically disordered regions for the estimation of protein function," *PloS one,* vol. 9, p. e89890, 2014.

[178] E. Faraggi, B. Xue, and Y. Zhou, "Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network," *Proteins,* vol. 74, pp. 847 - 856, 2009.

[179] T. Zhang, E. Faraggi, and Y. Zhou, "Fluctuations of backbone torsion angles obtained from NMR-determined structures and their prediction," *Proteins: Structure, Function, and Bioinformatics,* vol. 78, pp. 3353-3362, 2010.

[180] M. Vendruscolo, E. Paci, C. M. Dobson, and M. Karplus, "Three key residues form a critical contact network in a protein folding transition state," *Nature,* vol. 409, pp. 641-645, 2001.

[181] M. Y. Lobanov, E. I. Furletova, N. S. Bogatyreva, M. A. Roytberg, and O. V. Galzitskaya, "Library of disordered patterns in 3D protein structures," *PLoS Comput. Biol,* vol. 6, p. e1000958, 2010.

[182] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach.," *Biometrics,* vol. 44, pp. 837 - 845, 1988.

[183] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez*, et al.*, "pROC: an open-source package for R and S+ to analyze and compare ROC curves," *BMC bioinformatics,* vol. 12, p. 77, 2011.

[184] C.-C. Chang and C.-J. Lin, "LIBSVM : a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology,* vol. 2, pp. 27:1--27:27, 2011.

[185] L. J. McGuffin, "Intrinsic disorder prediction from the analysis of multiple protein fold recognition models," *Bioinformatics,* vol. 24, pp. 1798-1804, 2008.

[186] Z. Dosztányi, V. Csizmok, P. Tompa, and I. Simon, "IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content," *Bioinformatics,* vol. 21, pp. 3433-3434, 2005.

[187] C. M. Slupsky, L. N. Gentile, L. W. Donaldson, C. D. Mackereth, J. J. Seidel, B. J. Graves*, et al.*, "Structure of the Ets-1 pointed domain and mitogen-activated protein kinase phosphorylation site," *Proceedings of the National Academy of Sciences,* vol. 95, pp. 12129-12134, 1998.

[188] M. Baens, P. Peeters, C. Guo, J. Aerssens, and P. Marynen, "Genomic organization of TEL: the human ETS-variant gene 6," *Genome Research*, vol. 6, pp. 404-413, 1996.

[189] J. Colicelli, "Human RAS superfamily proteins and related GTPases," *Science's STKE: signal transduction knowledge environment*, vol. 2004, p. RE13, 2004.

[190] S. Piskacek, M. Gregor, M. Nemethova, M. Grabner, P. Kovarik, and M. Piskacek, "Nine-amino-acid transactivation domain: establishment and prediction utilities," *Genomics*, vol. 89, pp. 756-768, 2007.

[191] M. McCoy, E. S. Stavridi, J. L. Waterman, A. M. Wieczorek, S. J. Opella, and T. D. Halazonetis, "Hydrophobic side-chain size is a determinant of the three-dimensional structure of the p53 oligomerization domain," *The EMBO Journal,* vol. 16, pp. 6230-6236, 1997.

[192] K. Peng, S. Vucetic, P. Radivojac, C. J. Brown, A. K. Dunker, and Z. Obradovic, "Optimizing long intrinsic disorder predictors with protein evolutionary information.," *J Bioinform Comput Biol.,* vol. 3, pp. 35 - 60, 2005.

[193] S. Iqbal, M. N. Islam, and M. T. Hoque, "Improved Protein Disorder Predictor by Smoothing Output," presented at the Internation Conference on Computer & Information Technology (ICCIT), 2014.

[194] A. Schlessinger, G. Yachdav, and B. Rost, "PROFbval: predict flexible and rigid residues in proteins.," *Bioinformatics,* vol. 22, pp. 891 - 893, 2006.

[195] T. Ishida and K. Kinoshita, "PrDOS: prediction of disordered protein regions from amino acid sequence," *Nucleic acids research*, vol. 35, pp. W460-W464, 2007.

[196] M. N. Islam, S. Iqbal, A. R. Katebi, and M. T. Hoque, "A balanced secondary structure predictor," *Journal of Theoretical Biology,* vol. 389, pp. 60 - 71, 2016.

[197] B. Lee and F. M. Richards, "The interpretation of protein structures: estimation of static accessibility," *Journal of molecular biology,* vol. 55, pp. 379IN3-400IN4, 1971.

[198] A. Shrake and J. Rupley, "Environment and exposure to solvent of protein atoms. Lysozyme and insulin," *Journal of molecular biology,* vol. 79, pp. 351IN15365-364371, 1973.

[199] K. V. Klenin, F. Tristram, T. Strunk, and W. Wenzel, "Derivatives of molecular surface area and volume: Simple and exact analytical formulas," *Journal of computational chemistry,* vol. 32, pp. 2647-2653, 2011.

[200] J. Weiser, P. S. Shenkin, and W. C. Still, "Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO)," *Journal of Computational Chemistry,* vol. 20, pp. 217-230, 1999.

[201] A. Bondi, "van der Waals volumes and radii," *The Journal of physical chemistry,* vol. 68, pp. 441-451, 1964.

[202] C. Chothia, "Hydrophobic bonding and accessible surface area in proteins," *Nature,* vol. 248, pp. 338-339, 1974.

[203] J. A. Marsh, "Buried and Accessible Surface Area Control Intrinsic Protein Flexibility," *Journal of Molecular Biology,* vol. 425, pp. 3250 - 3263, 2013.

[204] H. Zhang, T. Zhang, K. Chen, S. Shen, J. Ruan, and L. Kurgan, "On the relation between residue flexibility and local solvent accessibility in proteins," *Proteins*, vol. 76, pp. 617 - 36, 2009.

[205] B. Lee and F. Richards, "The interpretation of protein structures: estimation of static accessibility," *J Mol Biol*, vol. 55, pp. 379-400, 1971.

[206] M. Connoly, "Solvent accessibility surfaces of protein and nucleic acids," *Science,* vol. 221, pp. 709 - 713, 1983.

[207] K. C. Chou and N. Y. Chen, "The biological functions of low-frequency phonons," *Scientia Sinica,* vol. 20, pp. 447 - 457, 1977.

[208] T. R. H. Raquel Requejo, Nikola J. Costa and Michael P. Murphy, "Cysteine residues exposed on protein surfaces are the dominant intramitochondrial thiol and may protect against oxidative damage," *The Febs Journal,* vol. 277, pp. 1465–1480, 2010.

[209] E. Butler, R. Davis, V. Bari, P. N. PA, and N. Ruiz, "Structure-Function Analysis of MurJ Reveals a Solvent-Exposed Cavity Containing Residues Essential for Peptidoglycan Biogenesis in Escherichia coli.," *Journal of Bacteriology,* vol. 195, pp. 4639-4649, 2013.

[210] M. Moret and G. Zebende, "Amino acid hydrophobicity and accessible surface area," *Physical Review E,* vol. 75, p. 011920, 2007.

[211] N. T. Southall, K. A. Dill, and A. Haymet, "A view of the hydrophobic effect," ed: ACS Publications, 2002.

[212] T. Zhou, J. R. Fleming, B. Franke, J. Bogomolovas, I. Barsukov, D. J. Rigden*, et al.*, "CARP interacts with titin at a unique helical N2A sequence and at the domain Ig81 to form a structured complex," *FEBS letters,* vol. 590, pp. 3098-3110, 2016.

[213] R. Fraczkiewicz and W. Braun, "Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules," *Journal of Computational Chemistry,* vol. 19, pp. 319-333, 1998.

[214] J. L. Battiste, T. V. Pestova, C. U. Hellen, and G. Wagner, "The eIF1A solution structure reveals a large RNA-binding surface important for scanning function," *Molecular cell,* vol. 5, pp. 109-119, 2000.

[215] K.-i. Cho, D. Kim, and D. Lee, "A feature-based approach to modeling protein–protein interaction hot spots," *Nucleic acids research,* vol. 37, pp. 2672-2687, 2009.

[216] J. Eickholt, X. Deng, and J. Cheng, "DoBo: Protein domain boundary prediction by integrating evolutionary signals and machine learning," *BMC bioinformatics,* vol. 12, p. 43, 2011.

[217] J. A. Marsh and S. A. Teichmann, "Relative Solvent Accessible Surface Area Predicts Protein Conformational Changes upon Binding," *Structure,* vol. 19, pp. 859–867, 2011.

[218] J. Cheng and P. Baldi, "A machine learning information retrieval approach to protein fold recognition," *Bioinformatics,* vol. 22, pp. 1456-1463, 2006.

[219] B. Rost, "TOPITS: Threading one-dimensional predictions into three-dimensional structures," *Third International Conference on Intelligent Systems for Molecular Biology,* pp. 314-312, 1995.

[220] D. Eisenberg and A. McLachlan, "Solvation energy in protein folding and binding," *Nature,* vol. 319, pp. 199 - 2013, 1986.

[221] S. Liu, C. Zhang, S. Liang, and Y. Zhou, "Fold recognition by concurrent use of solvent accessibility and residue depth," *Proteins,* vol. 68, pp. 636 - 664, 2007.

[222] D. Bonetti, H. Pérez-Sánchez, and A. Delbem, "An Efficient Solvent Accessible Surface Area calculation applied in Ab Initio Protein Structure Prediction."

[223] R. Khashan, W. Zheng, and A. Tropsha, "Scoring protein interaction decoys using exposed residues (SPIDER): a novel multibody interaction scoring function based on frequent geometric patterns of interfacial residues," *Proteins,* vol. 80, pp. 2207 - 2012, 2012.

[224] J. Wang and T. Hou, "Develop and Test a Solvent Accessible Surface Area-Based Model in Conformational Entropy Calculations," *Journal of Chemical Information and Modeling,* vol. 52, 2012.

[225] J. Moult, "Comparison of Database Potentials and Molecular Mechanics Force Fields.," *Curr Opin in Str Bio.,* vol. 7, pp. 194-199, April 1997 1997.

[226] S. Vajda, M. Sippl, and J. Novotny, "Empirical Potentials and Functions for Protein Folding and Binding.," *Curr Opin in Str Bio.,* vol. 7, pp. 222-228, 1997 1997.

[227] M.-H. Hao and H. A. Scheragat, "Designing Potential Energy Functions for Protein Folding.," *Curr Opin in Str Bio.,* vol. 9, pp. 184-188, April 1999 1999.

[228] S. Miyazawa and R. L. Jernigan, "An Empirical Energy Potential with a Reference State for Protein Fold and Sequence Recognition.," *Proteins: Struct., Funct., Genet.,* vol. 36, pp. 357-369, 8 April 1999 1999.

[229] T. Lazaridis and M. Karplus, "Effective Energy Functions for Protein Structure Prediction.," *Curr Opin in Str Bio.,* vol. 10, pp. 139-145, April 2000 2000.

[230] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson*, et al.*, "A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules.," *J. Am. Chem. Soc.,* vol. 117, pp. 5179-5197, 1995 1995.

[231] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, "CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations.," *J. Comput. Chem.,* vol. 4, pp. 187-217, 1 June 1983 1983.

[232] R. Samudrala and J. Moult, "An All-atom Distance-dependent Conditional Probability Discriminatory Function for Protein Structure Prediction.," *J. Mol. Biol.,* pp. 895-916, 1997.

[233] H. Zhou and Y. Zhou, "Distance-scaled, Finite Ideal-gas Reference State Improves Structure-derived Potentials of Mean Force for Structure Selection and Stability Prediction.," *Protein Sci.*, pp. 2714–2726, 2002.

[234] S. Tanaka and H. A. Scheraga, "Medium- and Long-Range Interaction Parameters between Amino Acids for Predicting Three-Dimensional Structures of Proteins.," *Macromolecules*, vol. 9, pp. 945-950, 1976 1976.

[235] R. L. Jernigan and I. Bahar, "Structure-Derived Potentials and Protein Simulations.," *Curr Opin in Str Bio.*, vol. 6, pp. 195-209, 1996 1996.

[236] K. K. Koretke, Z. Luthey-Schulten, and P. G. Wolynes, "Self-Consistently Optimized Statistical Mechanical Energy Functions for Sequence Structure Alignment.," *Protein Sci.*, vol. 5, pp. 1043-1059, 1996 1996.

[237] D. Tobi and R. Elber, "Distance-Dependent, Pair Potential for Protein Folding: Results From Linear Optimization.," *Proteins: Struct., Funct., Bioinf.*, vol. 41, pp. 40-46, 18 May 2000 2000.

[238] J. Skolnick, "In quest of an empirical potential for protein structure prediction.," *Curr Opin in Str Bio.*, vol. 16, pp. 166-171, 2006.

[239] A. Mishra and M. Hoque, "Three-Dimensional Ideal Gas Reference State based Energy Function," *Tech. Report TR-2014/2*, 2014.

[240] D. Frishman and P. Argos, "Knowledge-based protein secondary structure assignment," *Proteins: Structure, Function, and Bioinformatics*, vol. 23, pp. 566-579, 1995.

[241] J. Martin, G. Letellier, A. Marin, J.-F. Taly, A. G. de Brevern, and J.-F. Gibrat, "Protein secondary structure assignment revisited: a detailed analysis of different assignment methods," *BMC structural biology*, vol. 5, p. 17, 2005.

[242] L. Cavallo, J. Kleinjung, and F. Fraternali, "POPS: a fast algorithm for solvent accessible surface areas at atomic and residue level," *Nucleic acids research*, vol. 31, pp. 3364-3366, 2003.

[243] S. Mitternacht, "FreeSASA: An open source C library for solvent accessible surface area calculations," *F1000Research*, vol. 5, 2016.

[244] S. J. Hubbard and J. M. Thornton, "Naccess," *Computer Program, Department of Biochemistry and Molecular Biology, University College London*, vol. 2, 1993.

[245] K. D. Pruitt, T. Tatusova, and D. R. Maglott, "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins," *Nucleic acids research*, vol. 35, pp. D61-D65, 2007.

[246] B. Rost and C. Sander, "Conservation and prediction of solvent accessibility in protein families," *Proteins*, vol. 20, pp. 216 - 226, 1994.

[247] J. A. Cuff and G. J. Barton, "Application of multiple sequence alignment profiles to improve protein secondary structure prediction," *Proteins: Structure, Function, and Bioinformatics*, vol. 40, pp. 502-511, 2000.

[248] G. Pollastri, P. Baldi, P. Fariselli, and R. Casadio, "Prediction of coordination number and relative solvent accessibility in proteins," *Proteins: Structure, Function, and Bioinformatics*, vol. 47, pp. 142-153, 2002.

[249] S. Holbrook, S. Muskal, and S. Kim, "Predicting surface exposure of amino acids from protein sequence," *Protein Eng*, vol. 3, 1990.

[250] S. Ahmad and M. Gromiha, "NETASA: neural network based prediction of solvent accessibility," *Bioinformatics,* vol. 18, pp. 819 - 824, 2002.

[251] X. Li and X. Pan, "New method for accurate prediction of solvent accessibility from protein sequence," *Proteins,* vol. 42, pp. 1 - 5, 2001.

[252] Z. Yuan, K. Burrage, and J. Mattick, "Prediction of protein solvent accessibility using support vector machines," *Proteins,* vol. 48, pp. 566 - 570, 2002.

[253] H. Kim and H. Park, "Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local discriptor," *Proteins,* vol. 54, pp. 557 - 562, 2014.

[254] M. J. Thompson and R. A. Goldstein, "Predicting solvent accessibility: higher accuracy using Bayesian statistics and optimized residue substitution classes," 1996.

[255] J. Sim, S.-Y. Kim, and J. Lee, "Prediction of protein solvent accessibility using fuzzy k-nearest neighbor method," *Bioinformatics,* vol. 21, pp. 2844-2849, 2005.

[256] S. Ahmad, M. Gromiha, and A. Sarai, "Real value prediction of solvent accessibility from amino acid sequence," *Proteins,* vol. 50, pp. 629 - 635, 2013.

[257] O. Dor and Y. Zhou, "Real-SPINE: An integrated system of neural networks for real-value prediction of protein structural properties," *PROTEINS: Structure, Function, and Bioinformatics,* vol. 68, pp. 76-81, 2007.

[258] E. Faraggi, Y. Zhou, and A. Kloczkowski, "Accurate single-sequence prediction of solvent accessible surface area using local and global features," *Proteins: Structure, Function, and Bioinformatics,* vol. 82, pp. 3170-3176, 2014.

[259] R. Heffernan, K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J. Wang, *et al.*, "Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning," *Scientific reports,* vol. 5, 2015.

[260] R. Adamczak, A. Porollo, and J. Meller, "Accurate prediction of solvent accessibility using neural networks–based regression," *Proteins: Structure, Function, and Bioinformatics,* vol. 56, pp. 753-767, 2004.

[261] J.-Y. Wang, H.-M. Lee, and S. Ahmad, "Prediction and evolutionary information analysis of proteins solvent accessibility using multiple linear regression," *Proteins,* vol. 61, 2005.

[262] Z. Yuan and B. Huang, "Prediction of protein accessible surface areas by support vector regression," *Proteins,* vol. 57, pp. 558 - 564, 2014.

[263] J. Wang, H. Lee, and S. Ahmad, "SVM-cabins: prediction of solvent accessibility using accumulation cutoff set and support vector machine," *Proteins,* vol. 68, pp. 82 - 91, 2007.

[264] S. Altschul, T. Madden, A. Schäffer, J. Zhang, Z. Zhang, W. Miller, *et al.*, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Res,* vol. 25, pp. 3389-3402, 1997.

[265] A. Sharma, J. Lyons, A. Dehzangi, and K. Paliwal, "A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition," *J Theor Biol.,* vol. 320, pp. 41 - 46, 2013.

[266] A. Momen-Roknabadi, M. Sadeghi, H. Pezeshk, and S. A. Marashi, "Impact of residue accessible surface area on the prediction of protein secondary structures," *BMC Bioinformatics,* vol. 9, 2008.

[267] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.*: Springer, 2001.

[268] M. Kühn, T. Severin, and H. Salzwedel, "Variable Mutation Rate at Genetic Algorithms: Introduction of Chromosome Fitness in Connection with Multi-Chromosome Representation " *International Journal of Computer Applications* vol. 72, pp. 31 - 38, 2013.

[269] G. Ochoa, I. Harvey, and H. Buxton, "Optimal Mutation Rates and Selection Pressure in Genetic Algorithms," *Proc. Genetic and Evolutionary Computation Conference (GECCO},* 2000.

[270] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A Library for Large Linear Classification, Journal of Machine Learning Research," *Journal of Machine Learning Research,* vol. 9, pp. 1871--1874, 2008.

[271] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations,* vol. 11, 2009.

[272] W. Chen, P. M. Feng, H. Lin, and K. C. Chou, "iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition.," *Nucleic Acids Res. ,* vol. 41, p. e68, 2013.

[273] H. Lin, E. Z. Deng, H. Ding, W. Chen, and K. C. Chou, "iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition.," *Nucleic Acids Res. ,* vol. 42, pp. 12961 - 72, 2014.

[274] H. Ding, W. Z. Deng, L. F. Yuan, L. Liu, H. Lin, W. Chen, *et al.*, "iCTX-type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels.," *BioMed Research International,* vol. 2014, 2014.

[275] Y. Xu, X. Wen, L. S. Wen, L. Y. Wu, N. Y. Deng, and K. C. Chou, "iNitro-Tyr: prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition.," *PLoS One,* vol. 9, p. e105018, 2014.

[276] Z. Liu, X. Xiao, W. R. Qiu, and K. C. Chou, "iDNA-Methyl: identifying DNA methylation sites via pseudo trinucleotide composition," *Anal Biochem,* vol. 474, pp. 69 - 77, 2015.

[277] J. Jia, Z. Liu, X. Xiao, B. Liu, and K. C. Chou, "iPPI-Esml: An ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC.," *J Theor Biol,* vol. 377, pp. 47 - 56, 2015.

[278] K. C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition.," *J Theor Biol,* vol. 273, pp. 236 - 47, 2011.

[279] S. H. Guo, E. Z. Deng, L. Q. Xu, H. Ding, H. Lin, W. Chen*, et al.*, "iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition.," *Bioinformatics,* vol. 30, pp. 1522 - 9, 2014.

[280] C. KC, "Impacts of bioinformatics to medicinal chemistry," *Med Chem,* vol. 11, pp. 218 - 34, 2015.

[281] W. Revelle, "psych: Procedures for Psychological, Psychometric, and Personality Research," 2015.

[282] S. Iqbal and M. Hoque, "DisPredict: A Fine Disorder-Protein Predictor," *Tech. Report TR-2014/1,* 2014.

[283] A. Szilagyi, D. Györffy, and P. Zavodszky, "The Twilight Zone between Protein Order and Disorder," *Biophysical Journa,* vol. 95, pp. 1612 - 1626, 2008.

[284] J. Zhang and Y. Zhang, "A Novel Side-Chain Orientation Dependent Potential Derived from Random-Walk Reference State for Protein Fold Selection and Structure Prediction.," *Plos One,* vol. 5, 2010.

[285] Y. Yang and Y. Zhou, "Specific interactions for ab initio folding of protein terminal regions with secondary structures.," *Proteins,* vol. 72, pp. 793-803, 2008.

[286] H. Zhou and J. Skolnick, "GOAP: A Generalized Orientation-Dependent, All-Atom Statistical Potential for Protein Structure Prediction.," *Biophys. J* . vol. 101, pp. 2043-2052, 2011.

[287] J. Tsai, R. Bonneau, A. V. Morozov, B. Kuhlman, C. A. Rohl, and D. Baker, "An Improved Protein Decoy Set for Testing Energy Functions for Protein Structure Prediction.," *Proteins: Struct., Funct., Bioinf.,* vol. 53, pp. 76-87, 18 February 2003 2003.

[288] Z. T. Matthew, G. M. Austin, K. S. Dariya, J. S. Stephanie, and O. W. Claus, "Maximum Allowed Solvent Accissibilities of Residues in Proteins " *PLOS ONE,* vol. 8, p. e80635, 2013.

[289] P. E. Leopold, M. Montal, and J. N. Onuchic, "Protein folding funnels: a kinetic approach to the sequence-structure relationship," *Proceedings of the National Academy of Sciences,* vol. 89, pp. 8721-8725, 1992.

[290] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, "Funnels, pathways, and the energy landscape of protein folding: a synthesis," *Proteins: Structure, Function, and Bioinformatics,* vol. 21, pp. 167-195, 1995.

[291] K. A. Dill and J. L. MacCallum, "The protein-folding problem, 50 years on," *Science,* vol. 338, pp. 1042-1046, 2012.

[292] A. White, P. Handler, E. Smith, and D. Stetten Jr, "Principles of biochemistry," *Principles of Biochemistry.,* 1959.

[293] C. N. Pace, B. A. Shirley, M. McNutt, and K. Gajiwala, "Forces contributing to the conformational stability of proteins," *The FASEB journal,* vol. 10, pp. 75-83, 1996.

[294] D. Cui, S. Ou, and S. Patel, "Protein-spanning water networks and implications for prediction of protein–protein interactions mediated through hydrophobic effects," *Proteins: Structure, Function, and Bioinformatics,* vol. 82, pp. 3312-3326, 2014.

[295] S. Miyazawa and R. L. Jernigan, "Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation," *Macromolecules,* vol. 18, pp. 534-552, 1985.

[296] M. J. Sippl, "Calculation of conformational ensembles from potentials of mena force: an approach to the knowledge-based prediction of local structures in globular proteins," *Journal of molecular biology,* vol. 213, pp. 859-883, 1990.

[297] P. D. Thomas and K. A. Dill, "An iterative method for extracting energy-like quantities from protein structures," *Proceedings of the National Academy of Sciences,* vol. 93, pp. 11628-11633, 1996.

[298] D. Tobi, G. Shafran, N. Linial, and R. Elber, "On the design and analysis of protein folding potentials," *Proteins: Structure, Function, and Bioinformatics,* vol. 40, pp. 71-85, 2000.

[299] L. A. Mirny and E. I. Shakhnovich, "How to derive a protein folding potential? A new approach to an old problem," *Journal of molecular biology,* vol. 264, pp. 1164-1179, 1996.

[300] S. Miyazawa and R. L. Jernigan, "Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading," *Journal of molecular biology,* vol. 256, pp. 623-644, 1996.

[301] D. T. Jones, W. Taylort, and J. M. Thornton, "A new approach to protein fold recognition," *Nature,* vol. 358, pp. 86-89, 1992.

[302] A. E. Torda, "Perspectives in protein-fold recognition," *Current opinion in structural biology,* vol. 7, pp. 200-205, 1997.

[303] H. Gohlke, M. Hendlich, and G. Klebe, "Knowledge-based scoring function to predict protein-ligand interactions," *Journal of molecular biology,* vol. 295, pp. 337-356, 2000.

[304] M. A. Rashid, S. Iqbal, F. Khatib, M. T. Hoque, and A. Sattar, "Guided macro-mutation in a graded energy based genetic algorithm for protein structure prediction," *Computational Biology and Chemistry,* vol. 61, pp. 162-177, 2016.

[305] J. Khatun, S. D. Khare, and N. V. Dokholyan, "Can contact potentials reliably predict stability of proteins?," *Journal of molecular biology,* vol. 336, pp. 1223-1238, 2004.

[306] P. A. Alexander, Y. He, Y. Chen, J. Orban, and P. N. Bryan, "The design and characterization of two proteins with 88% sequence identity but different structure and function," *Proceedings of the National Academy of Sciences,* vol. 104, pp. 11963-11968, 2007.

[307] M. T. Hoque, Y. Yang, A. Mishra, and Y. Zhou, "sDFIRE: Sequence-specific statistical energy function for protein structure prediction by decoy selections," *Journal of Computational Chemistry,* 2015.

[308] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, *et al.*, "The Protein Data Bank.," *Nucleic Acids Res,* vol. 28, pp. 135 - 242, 1999.

[309] M. Z. Tien, A. G. Meyer, D. K. Sydykova, S. J. Spielman, and C. O. Wilke, "Maximum allowed solvent accessibilites of residues in proteins," *PLoS ONE,* vol. 8, p. e80635, 2013.

[310] M. N. Islam, S. Iqbal, and M. T. Hoque, "A balaced secondary structure predictor," *Journal of Theoretical Biology,* vol. 389, pp. 60 - 71, 2016.

[311] W. J. Youden, "Index for rating diagnostic tests," *Cancer,* vol. 3, pp. 32-35, 1950.

[312] T. Sing, O. Sander, N. Beerenwinkel, T. Lengauer, T. Sing, and O. Sander, "Visualizing the performance of scoring classifiers," *Package ROCR Version 1.0,* vol. 4, 2009.

[313] I. Walsh, A. Martin, T. D. Domenico, and S. Tosatto, "ESpritz: accurate and fast prediction of protein disorder.," *Bioinformatics,* vol. 28, pp. 503 - 9, 2012.

[314] L. McGuffin, "Intrinsic disorder prediction from the analysis of multiple protein fold recognition models," *Bioinformatics,* vol. 24, pp. 1798 - 804, 2008.

[315] S. Pawlicki, A. Le Béchec, and C. Delamarche, "AMYPdb: a database dedicated to amyloid precursor proteins," *BMC bioinformatics,* vol. 9, p. 273, 2008.

[316] B. Mészáros, I. Simon, and Z. Dosztányi, "Prediction of protein binding regions in disordered proteins," *PLoS Comput Biol.,* vol. 5, p. e1000376, 2009.

[317] S. Iqbal and M. T. Hoque, "Prediction of Peptide-Binding Residues of Receptor Proteins in a Complex," in *The 5th Annual Conference on Computational Biology and Bioinformatics*, New Orleans, LA, 2017.

[318] J. D. Scott and T. Pawson, "Cell signaling in space and time: where proteins come together and when they're apart," *Science,* vol. 326, pp. 1220-1224, 2009.

[319] P. L. Toogood, "Inhibition of protein− protein association by small molecules: Approaches and progress," *Journal of medicinal chemistry,* vol. 45, pp. 1543-1558, 2002.

[320] T. Pawson and P. Nash, "Assembly of cell regulatory systems through protein interaction domains," *science,* vol. 300, pp. 445-452, 2003.

[321] N. London, D. Movshovitz-Attias, and O. Schueler-Furman, "The structural basis of peptide-protein binding strategies," *Structure,* vol. 18, pp. 188-199, 2010.

[322] N. Malhis and J. Gsponer, "Computational identification of MoRFs in protein sequences," *Bioinformatics,* vol. 31, pp. 1738-1744, 2015.

[323] F. M. Disfani, W.-L. Hsu, M. J. Mizianty, C. J. Oldfield, B. Xue, A. K. Dunker*, et al.*, "MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins," *Bioinformatics,* vol. 28, pp. i75-i83, 2012.

[324] V. Neduva and R. B. Russell, "Peptides mediating interaction networks: new leads at last," *Current opinion in biotechnology,* vol. 17, pp. 465-471, 2006.

[325] C. Blikstad and Y. Ivarsson, "High-throughput methods for identification of protein-protein interactions involving short linear motifs," *Cell Communication and Signaling,* vol. 13, p. 38, 2015.

[326] B. A. Liu, B. W. Engelmann, and P. D. Nash, "High‐throughput analysis of peptide‐binding modules," *Proteomics,* vol. 12, pp. 1527-1546, 2012.

[327] A. Stein, R. Mosca, and P. Aloy, "Three-dimensional modeling of protein interactions and complexes is going 'omics," *Current opinion in structural biology,* vol. 21, pp. 200-208, 2011.

[328] A. Dömling, "Small molecular weight protein–protein interaction antagonists—an insurmountable challenge?," *Current opinion in chemical biology,* vol. 12, pp. 281-291, 2008.

[329] J. A. Wells and C. L. McClendon, "Reaching for high-hanging fruit in drug discovery at protein–protein interfaces," *Nature,* vol. 450, pp. 1001-1009, 2007.

[330] M. P. Stumpf, T. Thorne, E. de Silva, R. Stewart, H. J. An, M. Lappe*, et al.*, "Estimating the size of the human interactome," *Proceedings of the National Academy of Sciences,* vol. 105, pp. 6959-6964, 2008.

[331] P. Braun, "Interactome mapping for analysis of complex phenotypes: insights from benchmarking binary interaction assays," *Proteomics,* vol. 12, pp. 1499-1518, 2012.

[332] P.-O. Vidalain, M. Boxem, H. Ge, S. Li, and M. Vidal, "Increasing specificity in high-throughput yeast two-hybrid experiments," *Methods,* vol. 32, pp. 363-370, 2004.

[333] C. Von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields*, et al.*, "Comparative assessment of large-scale data sets of protein–protein interactions," *Nature,* vol. 417, pp. 399-403, 2002.

[334] E. Sprinzak, S. Sattath, and H. Margalit, "How reliable are experimental protein–protein interaction data?," *Journal of molecular biology,* vol. 327, pp. 919-923, 2003.

[335] B. A. Shoemaker and A. R. Panchenko, "Deciphering protein–protein interactions. Part II. Computational methods to predict protein and domain interaction partners," *PLoS Comput Biol,* vol. 3, p. e43, 2007.

[336] D. Petrey and B. Honig, "Structural bioinformatics of the interactome," *Annual review of biophysics,* vol. 43, pp. 193-210, 2014.

[337] C. Landgraf, S. Panni, L. Montecchi-Palazzi, L. Castagnoli, J. Schneider-Mergener, R. Volkmer-Engert, *et al.*, "Protein interaction networks by proteome peptide scanning," *PLoS Biol,* vol. 2, p. e14, 2004.

[338] R. Tonikian, X. Xin, C. P. Toret, D. Gfeller, C. Landgraf, S. Panni, *et al.*, "Bayesian modeling of the yeast SH3 domain interactome predicts spatiotemporal dynamics of endocytosis proteins," *PLoS Biol,* vol. 7, p. e1000218, 2009.

[339] P. Vanhee, F. Stricher, L. Baeten, E. Verschueren, T. Lenaerts, L. Serrano, *et al.*, "Protein-peptide interactions adopt the same structural motifs as monomeric protein folds," *Structure,* vol. 17, pp. 1128-1136, 2009.

[340] O. Keskin, B. Ma, and R. Nussinov, "Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues," *Journal of molecular biology,* vol. 345, pp. 1281-1294, 2005.

[341] S. Lise, C. Archambeau, M. Pontil, and D. T. Jones, "Prediction of hot spot residues at protein-protein interfaces by combining machine learning and energy-based methods," *BMC bioinformatics,* vol. 10, p. 365, 2009.

[342] S. Lise, D. Buchan, M. Pontil, and D. T. Jones, "Predictions of hot spot residues at protein-protein interfaces using support vector machines," *PLoS one,* vol. 6, p. e16774, 2011.

[343] E. Petsalaki and R. B. Russell, "Peptide-mediated interactions in biological systems: new discoveries and applications," *Current opinion in biotechnology,* vol. 19, pp. 344-350, 2008.

[344] J. R. Chen, B. H. Chang, J. E. Allen, M. A. Stiffler, and G. MacBeath, "Predicting PDZ domain–peptide interactions from primary sequences," *Nature biotechnology,* vol. 26, pp. 1041-1045, 2008.

[345] T. Hou, K. Chen, W. A. McLaughlin, B. Lu, and W. Wang, "Computational analysis and prediction of the binding motif and protein interacting partners of the Abl SH3 domain," *PLoS Comput Biol,* vol. 2, p. e1, 2006.

[346] S. Jain and G. D. Bader, "Predicting physiologically relevant SH3 domain mediated protein–protein interactions in yeast," *Bioinformatics,* vol. 32, pp. 1865-1872, 2016.

[347] I. E. Sánchez, P. Beltrao, F. Stricher, J. Schymkowitz, J. Ferkinghoff-Borg, F. Rousseau, *et al.*, "Genome-wide prediction of SH2 domain targets using structural information and the FoldX algorithm," *PLoS Comput Biol,* vol. 4, p. e1000052, 2008.

[348] N. London, C. L. Lamphear, J. L. Hougland, C. A. Fierke, and O. Schueler-Furman, "Identification of a novel class of farnesylation targets by structure-based modeling of binding specificity," *PLoS Comput Biol,* vol. 7, p. e1002170, 2011.

[349] J. DeBartolo, M. Taipale, and A. E. Keating, "Genome-wide prediction and validation of peptides that bind human prosurvival Bcl-2 proteins," *PLoS Comput Biol,* vol. 10, p. e1003693, 2014.

[350] H. Dinkel, K. Van Roey, S. Michael, N. E. Davey, R. J. Weatheritt, D. Born, *et al.*, "The eukaryotic linear motif resource ELM: 10 years and counting," *Nucleic acids research*, vol. 42, pp. D259-D266, 2014.

[351] A. Stein and P. Aloy, "Novel peptide-mediated interactions derived from high-resolution 3-dimensional structures," *PLoS Comput Biol,* vol. 6, p. e1000789, 2010.

[352] T. S. Chen, D. Petrey, J. I. Garzon, and B. Honig, "Predicting peptide-mediated interactions on a genome-wide scale," *PLoS Comput Biol,* vol. 11, p. e1004248, 2015.

[353] E. Petsalaki, A. Stark, E. García-Urdiales, and R. B. Russell, "Accurate prediction of peptide binding sites on protein surfaces," *PLoS Comput Biol*, vol. 5, p. e1000335, 2009.

[354] A. Lavi, C. H. Ngan, D. Movshovitz‐Attias, T. Bohnuud, C. Yueh, D. Beglov, *et al.*, "Detection of peptide‐binding sites on protein surfaces: The first step toward the modeling and targeting of peptide‐mediated interactions," *Proteins: Structure, Function, and Bioinformatics,* vol. 81, pp. 2096-2105, 2013.

[355] A. A. Das, O. P. Sharma, M. S. Kumar, R. Krishna, and P. P. Mathur, "PepBind: a comprehensive database and computational tool for analysis of protein–peptide interactions," *Genomics, proteomics & bioinformatics,* vol. 11, pp. 241-246, 2013.

[356] I. Hoof, B. Peters, J. Sidney, L. E. Pedersen, A. Sette, O. Lund, *et al.*, "NetMHCpan, a method for MHC class I binding prediction beyond humans," *Immunogenetics,* vol. 61, p. 1, 2009.

[357] M. Nielsen and M. Andreatta, "NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets," *Genome medicine,* vol. 8, p. 33, 2016.

[358] G. Taherzadeh, Y. Yang, T. Zhang, A. W. C. Liew, and Y. Zhou, "Sequence‐based prediction of protein–peptide binding sites using support vector machine," *Journal of computational chemistry,* 2016.

[359] *A Kaggler's Guide to Model Stacking in Practice*. Available: http://blog.kaggle.com/2016/12/27/a-kagglers-guide-to-model-stacking-in-practice/

[360] C. A. Janeway, P. Travers, M. Walport, and M. J. Shlomchik, *Immunobiology: the immune system in health and disease* vol. 1: Current Biology Singapore, 1997.

[361] B. Z. Harris and W. A. Lim, "Mechanism and role of PDZ domains in signaling complex assembly," *Journal of cell science,* vol. 114, pp. 3219-3231, 2001.

[362] B. A. Liu, K. Jablonowski, E. E. Shah, B. W. Engelmann, R. B. Jones, and P. D. Nash, "SH2 domains recognize contextual peptide sequence information to determine selectivity," *Molecular & Cellular Proteomics,* vol. 9, pp. 2391-2404, 2010.

[363] S. Zhou, S. E. Shoelson, M. Chaudhuri, G. Gish, T. Pawson, W. G. Haser*, et al.*, "SH2 domains recognize specific phosphopeptide sequences," *Cell,* vol. 72, pp. 767-778, 1993.

[364] A. Zarrinpar, R. P. Bhattacharyya, and W. A. Lim, "The structure and function of proline recognition domains," *Homo,* vol. 332, p. 20, 2003.

[365] D. M. Berry, P. Nash, S. K.-W. Liu, T. Pawson, and C. J. McGlade, "A high-affinity Arg-XX-Lys SH3 binding motif confers specificity for the interaction between Gads and SLP-76 in T cell signaling," *Current Biology,* vol. 12, pp. 1336-1341, 2002.

[366] S. D. Stamenova, M. E. French, Y. He, S. A. Francis, Z. B. Kramer, and L. Hicke, "Ubiquitin binds to and regulates a subset of SH3 domains," *Molecular cell,* vol. 25, pp. 273-284, 2007.

[367] A. J. Muslin, J. W. Tanner, P. M. Allen, and A. S. Shaw, "Interaction of 14-3-3 with signaling proteins is mediated by the recognition of phosphoserine," *Cell,* vol. 84, pp. 889-897, 1996.

[368] P.-J. Lu, X. Z. Zhou, M. Shen, and K. P. Lu, "Function of WW domains as phosphoserine-or phosphothreonine-binding modules," *Science,* vol. 283, pp. 1325-1328, 1999.

[369] W. M. Kavanaugh, C. W. Turck, and L. T. Williams, "PTB domain binding to signaling proteins through a sequence motif containing phosphotyrosine," *Science,* vol. 268, p. 1177, 1995.

[370] D. H. Mohammad and M. B. Yaffe, "14-3-3 proteins, FHA domains and BRCT domains in the DNA damage response," *DNA repair,* vol. 8, pp. 1009-1017, 2009.

[371] J. F. Amacher, P. R. Cushing, L. Brooks, P. Boisguerin, and D. R. Madden, "Stereochemical preferences modulate affinity and selectivity among five PDZ domains that bind CFTR: comparative structural and sequence analyses," *Structure,* vol. 22, pp. 82-93, 2014.

[372] P. Block, J. Paern, E. Huellermeier, P. Sanschagrin, C. A. Sotriffer, and G. Klebe, "Physicochemical descriptors to discriminate protein–protein interactions in permanent and transient complexes selected by means of machine learning algorithms," *PROTEINS: Structure, Function, and Bioinformatics,* vol. 65, pp. 607-622, 2006.

[373] M. Yeager, S. Kumar, and A. L. Hughes, "Sequence convergence in the peptide-binding region of primate and rodent MHC class Ib molecules," *Molecular biology and evolution,* vol. 14, pp. 1035-1041, 1997.

[374] K. K. Paliwal, A. Sharma, J. Lyons, and A. Dehzangi, "A tri-gram based feature extraction technique using linear probabilities of position specific scoring matrix for protein fold recognition," *IEEE Transactions on NanoBioscience,* vol. 13, pp. 44-50, 2014.

[375] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*: CRC press, 1984.

[376] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, *et al.*, "Package ‘pROC’," 2015.

[377] D. H. Wolpert, "Stacked generalization," *Neural networks,* vol. 5, pp. 241-259, 1992.

[378] E. Frank, M. Hall, L. Trigg, G. Holmes, and I. H. Witten, "Data mining in bioinformatics using Weka," *Bioinformatics,* vol. 20, pp. 2479-2481, 2004.

[379] "Kaggle Ensembling Guide."

[380] S. Nagi and D. K. Bhattacharyya, "Classification of microarray cancer data using ensemble approach," *Network Modeling Analysis in Health Informatics and Bioinformatics,* vol. 2, pp. 159-173, 2013.

[381] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *Evolutionary Computation, IEEE Transactions on,* vol. 1, pp. 67-82, 1997.

[382] T. K. Ho, "Random decision forests," in *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on,* 1995, pp. 278-282.

[383] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE transactions on pattern analysis and machine intelligence,* vol. 20, pp. 832-844, 1998.

[384] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research,* vol. 12, pp. 2825-2830, 2011.

[385] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning,* vol. 63, pp. 3-42, 2006.

[386] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis,* vol. 38, pp. 367-378, 2002.

[387] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician,* vol. 46, pp. 175-185, 1992.

[388] L. Breiman, "Bagging predictors," *Machine learning,* vol. 24, pp. 123-140, 1996.

[389] D. A. Freedman, *Statistical models: theory and practice*: cambridge university press, 2009.

[390] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, "Understanding variable importances in forests of randomized trees," in *Advances in Neural Information Processing Systems*, 2013, pp. 431-439.

[391] S. Frese, W.-D. Schubert, A. C. Findeis, T. Marquardt, Y. S. Roske, T. E. Stradal, *et al.*, "The phosphotyrosine peptide binding specificity of Nck1 and Nck2 Src homology 2 domains," *Journal of Biological Chemistry,* vol. 281, pp. 18236-18245, 2006.

[392] G. K. Balendiran, J. C. Solheim, A. C. Young, T. H. Hansen, S. G. Nathenson, and J. C. Sacchettini, "The three-dimensional structure of an H-2Ld-peptide complex explains the unique interaction of Ld with beta-2 microglobulin and peptide," *Proceedings of the National Academy of Sciences,* vol. 94, pp. 6880-6885, 1997.

[393] R. N. Murugan, M. Ahn, W. C. Lee, H.-Y. Kim, J. H. Song, C. Cheong, *et al.*, "Exploring the binding nature of pyrrolidine pocket-dependent interactions in the polo-box domain of polo-like kinase 1," *PloS one,* vol. 8, p. e80043, 2013.

[394] K. Arita, S. Isogai, T. Oda, M. Unoki, K. Sugita, N. Sekiyama, *et al.*, "Recognition of modification status on a histone H3 tail by linked histone reader modules of the epigenetic regulator UHRF1," *Proceedings of the National Academy of Sciences,* vol. 109, pp. 12950-12955, 2012.

[395] B. Schumacher, M. Skwarczynska, R. Rose, and C. Ottmann, "Structure of a 14-3-3σ–YAP phosphopeptide complex at 1.15 Å resolution," *Acta Crystallographica Section F: Structural Biology and Crystallization Communications,* vol. 66, pp. 978-984, 2010.

[396] R. Bonasio, E. Lecona, and D. Reinberg, "MBT domain proteins in development and disease," in *Seminars in cell & developmental biology,* 2010, pp. 221-230.

[397] O. Lohi, A. Poussu, Y. Mao, F. Quiocho, and V.-P. Lehto, "VHS domain–a longshoreman of vesicle lines," *FEBS letters,* vol. 513, pp. 19-23, 2002.

[398] V. Hoppmann, T. Thorstensen, P. E. Kristiansen, S. V. Veiseth, M. A. Rahman, K. Finne, *et al.*, "The CW domain, a new histone recognition module in chromatin proteins," *The EMBO journal,* vol. 30, pp. 1939-1952, 2011.

[399] S.-J. Chen, X. Wu, B. Wadas, J.-H. Oh, and A. Varshavsky, "An N-end rule pathway that recognizes proline and destroys gluconeogenic enzymes," *Science,* vol. 355, p. eaal3655, 2017.

[400] W. L. DeLano, "Unraveling hot spots in binding interfaces: progress and challenges," *Current opinion in structural biology,* vol. 12, pp. 14-20, 2002.

[401] J. A. Wells, "Additivity of mutational effects in proteins," *Biochemistry,* vol. 29, pp. 8509-8517, 1990.

[402] B. C. Cunningham and J. A. Wells, "High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis," *Science,* vol. 244, pp. 1081-1086, 1989.

[403] J. Skolnick, J. S. Fetrow, and A. Kolinski, "Structural genomics and its importance for gene function analysis," *Nature biotechnology,* vol. 18, pp. 283-287, 2000.

[404] K. L. Morrison and G. A. Weiss, "Combinatorial alanine-scanning," *Current opinion in chemical biology,* vol. 5, pp. 302-307, 2001.

[405] K. S. Thorn and A. A. Bogan, "ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions," *Bioinformatics,* vol. 17, pp. 284-285, 2001.

[406] T. Kortemme, D. E. Kim, and D. Baker, "Computational alanine scanning of protein-protein interfaces," *Sci STKE,* vol. 2004, p. pl2, 2004.

Vita

The author was born in Gopalganj, Bangladesh. She obtained her Bachelor’s degree in Computer Science and Engineering from Bangladesh University of Engineering and Technology (BUET) in 2009. In 2012, she obtained her Master’s degree in Computer Science and Engineering from BUET. She joined the University of New Orleans (UNO) computer science graduate program to pursue a PhD in Engineering and Applied Science (concentration: computer science) in 2013, and became a member of the Bioinformatics and Machine Learning (BML) laboratory of UNO computer science department. In BML lab, she carried out her dissertation work as a graduate research assistant under supervision of Dr. Md Tamjidul Hoque.