STRUCTURAL DYNAMICS

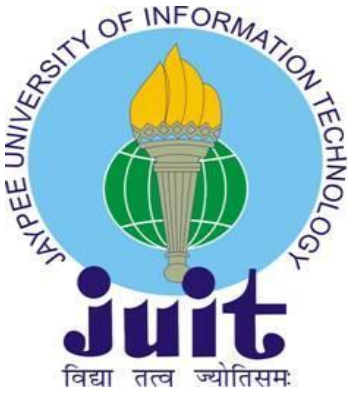
OF

LYTIC POLYSACCHARIDE MONOOXYGENASE

Enrolment No – 133812,131520

Name of Students – Priya Bharval, Sheena Sarswati

Name of Supervisor – Dr. Ragothaman M. Yennamalli



May 2017

Submitted in partial fulfillment of the requirement for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

BIOINFORMATICS

DEPARTMENT OF BIOTECHNOLOGY AND BIOINFORMATICS,

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY,

WAKNAGHAT, SOLAN 173234, HIMACHAL PRADESH, INDIA

ACKNOWLEDGEMENT

We would like to express our deepest appreciation to all those who provided us the possibility to complete this report. A special gratitude we give to our final year project supervisor, Dr. Ragothaman M. Yennamalli whose contribution in stimulating suggestions and encouragement helped us to coordinate our project especially in writing this report. Furthermore, we would also like to acknowledge with much appreciation the crucial role of the staff of Bioinformatics lab, who gave the permission to use all required equipment and the necessary material to complete the task. Last but not least, we have to appreciate the guidance given by other supervisors and the evaluation panels especially in our project presentation that has improved our communication skills.

CERTIFICATE

This is to certify that project report entitled "**Structural Dynamics of Lytic Polysaccharide Monooxygenase**", submitted by Priya Bharval and Sheena Sarswati is in its partial fulfillment for the award of degree of Bachelor of Technology in Bioinformatics Engineering to Jaypee University of Information Technology Waknaghat, Solan has been carried out under my supervision.

This work has not been submitted partially or fully to any other university or institution in order to achieve any award or any other degree.

Signature:

Supervisor Name: Dr. Ragothaman M. Yennamalli

Designation: Assistant Professor (Grade II),

Jaypee University of Information Technology,

Wakhnaghat Solan, Himachal Pradesh

CONTENTS

LIST OF FIGURES

Figure 14: Plots showing Kier molecular flexibility of 16 structures of AA9 calculated by Structuprint.

Figure 15: Plots showing fractional hydrophobic van der Waals surface area of 16 structures of AA9 calculated by Structuprint.

Figure 16: Plots showing total polar surface area of 16 structures of AA9 calculated by Structuprint.

Figure 17: Plots showing van der waals surface area of 16 structures of AA9 calculated by Structuprint.

Figure 18: Plots showing surface rugosity of 16 structures of AA9 calculated by Structuprint.

Figure 19: Plots showing Accessible Surface Area of 27 structures of AA10 calculated by Structuprint.

Figure 20: Plots showing electrostatic energy of 27 structures of AA10 calculated by Structuprint.

Figure 21: Plots showing solvation energy of 27 structures of AA10 calculated by Structuprint.

Figure 22: Plots showing potential energy of 27 structures of AA10 calculated by Structuprint.

Figure 23: Plots showing formal charge of 27 structures of AA10 calculated by Structuprint.

Figure 24: Plots showing globularity of 27 structures of AA10 calculated by Structuprint.

Figure 25: Plots showing Kier molecular flexibility of 27 structures of AA10 calculated by Structuprint.

Figure 26: Plots showing Fractional hydrophobic van der Waals surface area of 27 structures of AA10 calculated by Structuprint.

Figure 27: Plots showing total polar surface area of 27 structures of AA10 calculated by Structuprint.

Figure 28: Plots showing van der waals surface area of 27 structures of AA10 calculated by Structuprint.

Figure 29: Plots showing surface rugosity of 27 structures of AA10 calculated by Structuprint.

Figure 30: Plots of AA11 structure (pdb id: 4mah) for eleven properties calculated by Structuprint

Figure 31: Plots of AA11 structure (pdb id: 4mai) for eleven properties calculated by Structuprint

Figure 32: Sequence Correlations of AA10 A) histogram of the similarities between pairs of sequences B) The similarly matrix.

Figure 33: Positional Conservation in AA10.

Figure 34: (A) 3-D plots of the top three eigenvectors, (B) 2-D Plots of the top three eigenvectors, (C) Mapping Sequence correlation by positional correlation of AA10, (D) Defining Sectors according to the top Eigen mode vectors of AA10

Figure 35: (A) Side view by Statistical Coupling Analysis (SCA) and (B) Top view of residues that are predicted.

LIST OF TABLES

ABSTRACT

Lytic Polysaccharide Monooxygenases (LPMOs) found in fungi, bacteria, and viruses are redox enzymes that utilize copper to cleave glycosidic bonds in the recalcitrant crystalline form of polysaccharides. Cellulose and chitin are currently classified by CAZy under AA9, AA10, AA11, and AA13 families. LPMO's unusually "flat" and "rigid" active site framework for its catalytic activity has been the focus since their discovery in the early 2000s. LPMOs' molecular architecture to bind to cellulose and chitin (and other polymers) is most likely evolved due to the presence of a diverse substrate landscape. Here, using structural bioinformatics approach coupled with Elastic Network Modeling (ENM), we compare and contrast the structurally similar yet sequentially and functionally diverse polysaccharide monooxygenases. The structural dynamics studies of AA9, AA10, AA11, and AA13 families indicate that the "rigid" active site is highly flexible than previously hypothesized. Also, the loops on the substrate binding side are most mobile indicating their role in substrate binding. However, there are crucial dynamical and physicochemical differences between the four families that are responsible for their substrate specificity. The study also predicts key residues that are possibly responsible for substrate specificity and LPMO's function, in other words towards its allostery.

INTRODUCTION

General Background

Cellulose, the most abundant polymer found in plant biomass, presents itself as a unique and promising solution for solving the energy needs of the human race [1]. While, it is also the most recalcitrant polymer, nature has found numerous ways to break it down via cellulases, a broad class of enzymes that degrade cellulose [2].

Cellulose**,** a complex carbohydrate, or polysaccharide, consisting of 3,000 or more glucose units is the basic structural component of plant cell walls. Cellulose accounts for about 33%of all vegetable matter (90% of cotton and 50% of wood are cellulose) and is the most abundant of all naturally occurring organic compounds [3].The main ingredient of exoskeletons of arthropods is chitin and it is also the main component of the cell wall of fungi. Therefore, from beetles, butterflies and lobsters to spiders, crabs, and shrimp have chitin in their protective armors.

Lytic Polysaccharide Monooxygenases (LPMOs) are a recently discovered class of enzymes capable of oxidizing recalcitrant polysaccharides, such as cellulose and chitin. LPMOs found in fungi, bacteria, and viruses are redox enzymes that utilize copper to cleave glycosidic bonds in the recalcitrant crystalline form of polysaccharides [4]. They are considered to be important contributors for the conversion of polymers, where they synergistically act with other enzymes for the breakdown of recalcitrant polysaccharides, such as cellulose and chitin [5]. LPMOs have a planar binding site because of which their initial discovery in *Serratia marcesences* made them to be classified as chitin binding proteins [6]. However, in recent years LPMOs have been identified in various bacteria and other organisms including fungi and viruses. Previously, these enzymes were identified as glycoside hydrolase family 61 (GH61) and carbohydrate binding module family 33 (CBM33) and now are re-classified as AA9, AA10, AA11, AA13 [7]. As of writing this dissertation AA9 consist of 345 sequences, AA10 has 2558 sequences, AA11 has 68 sequences and AA13 has 16 sequences.

The core structure of LPMOs is an immunoglobulin like distorted β-sandwich fold [8,9]. It consists of antiparallel β-strands, which are connected by loops with different number of α-helix insertions. They have a flat surface which consists of a divalent copper centre that is

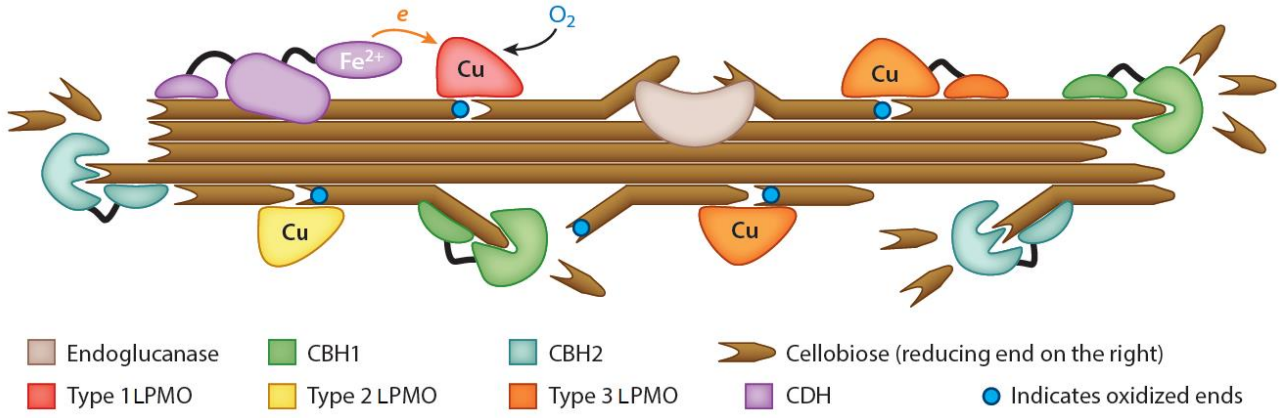


Figure 1: Schematic representation of LPMO's and other cellulolytic enzyme's synergistic activity on breaking down the crystalline cellulose. Image adapted from [12].

surrounded by three nitrogens in a T-shaped geometry [9]. The Imidazole side chain and the main chain amino group of the N terminal histidine contribute two of its nitrogens and the second conserved histidine contributes the third nitrogen. Copper ions reduce dioxygen, which requires electrons from an external electron donor, such as a reducing agent that is either provided by the substrate or by a cosecreted enzyme called cellobiose dehydrogenase [10]. Figure1 depicts the schematic representation of LPMO's synergism with other enzymes.

LPMO's unusually "flat" and "rigid" active site framework for its catalytic activity has been the focus since their discovery in the early 2000s [11]. LPMOs' molecular architecture to bind to cellulose and chitin (and other polymers) is most likely evolved due to the presence of a diverse substrate landscape [12].

Vaaje-Kolstad et al gave the first insight to the different mechanisms of LPMOs by using isotropically labeled dioxygen for the overall confirmation of monooxygenases activity on chitin [13]. Hemsworth et al and Beeson et al used β-amylase and phosphoric acid swollen cellulose to reveal action of different AA families, respectively [12, 13]. As LPMOs are reducing enzymes i.e. their main purpose is to abstract oxygen atoms from the substrate. In the first step, it removes the hydrogen atom (bound to either C1 or C4) thereby creating an electron imbalance in the glycosidic bond, which leads to release of oxygen and breaking of the glycosidic bond. There are majorly two type of reactions based on the hydrogen abstraction at the different carbon atom positions (either C1 or C4) followed by glycocidic (C-O) bond cleavage, where enzymes specifically attacking C1 are called as Type 1 LPMOs and enzymes specifically attacking C4 are called as Type 2 LPMOs. If some enzymes do not have a specifically to either C1 or C4, they are termed as type 3 LPMOs [14].

As shown in the Figure 2 if the hydrogen abstraction is from C1 site it is termed as Type 1 reaction and if the hydrogen is abstracted from C4 carbon atom it is called as Type 2 reaction, and if the enzyme abstracts oxygen from both the sites or is not specific of where the abstraction is being done, then it is called as Type 3 reaction.

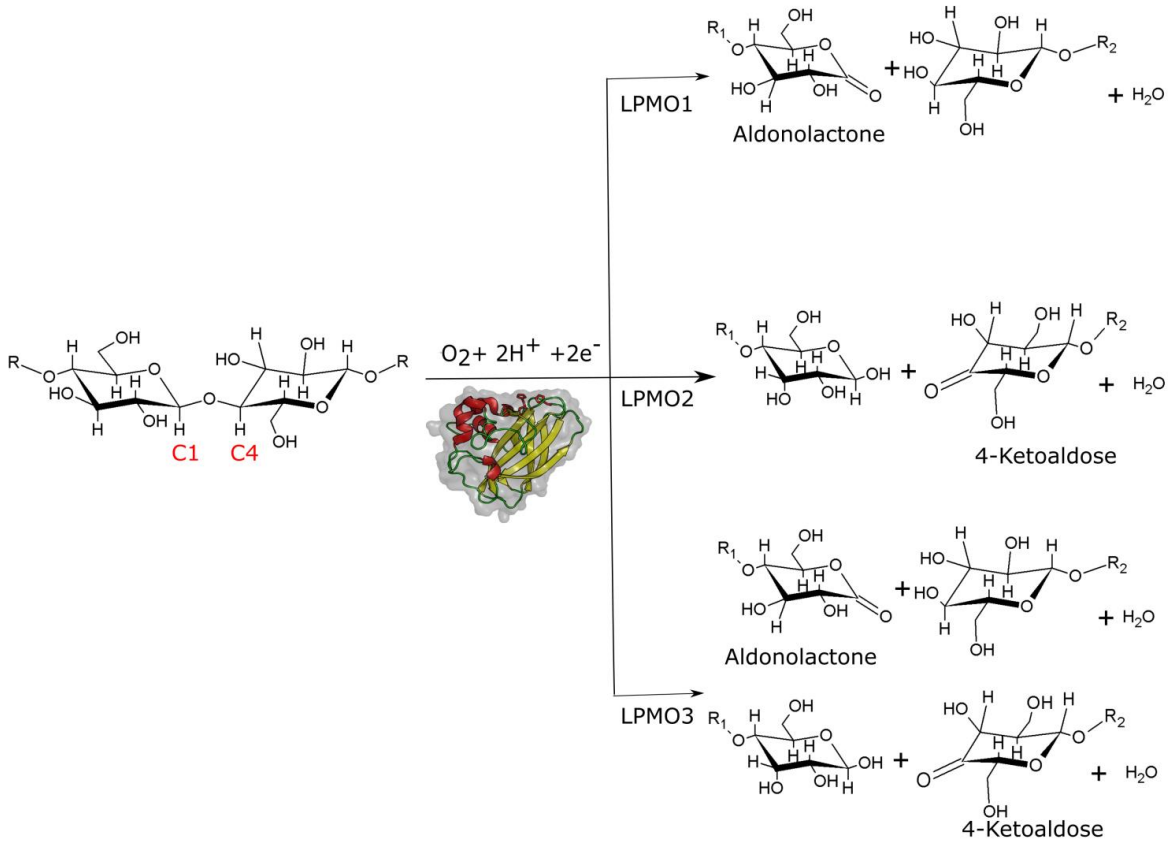


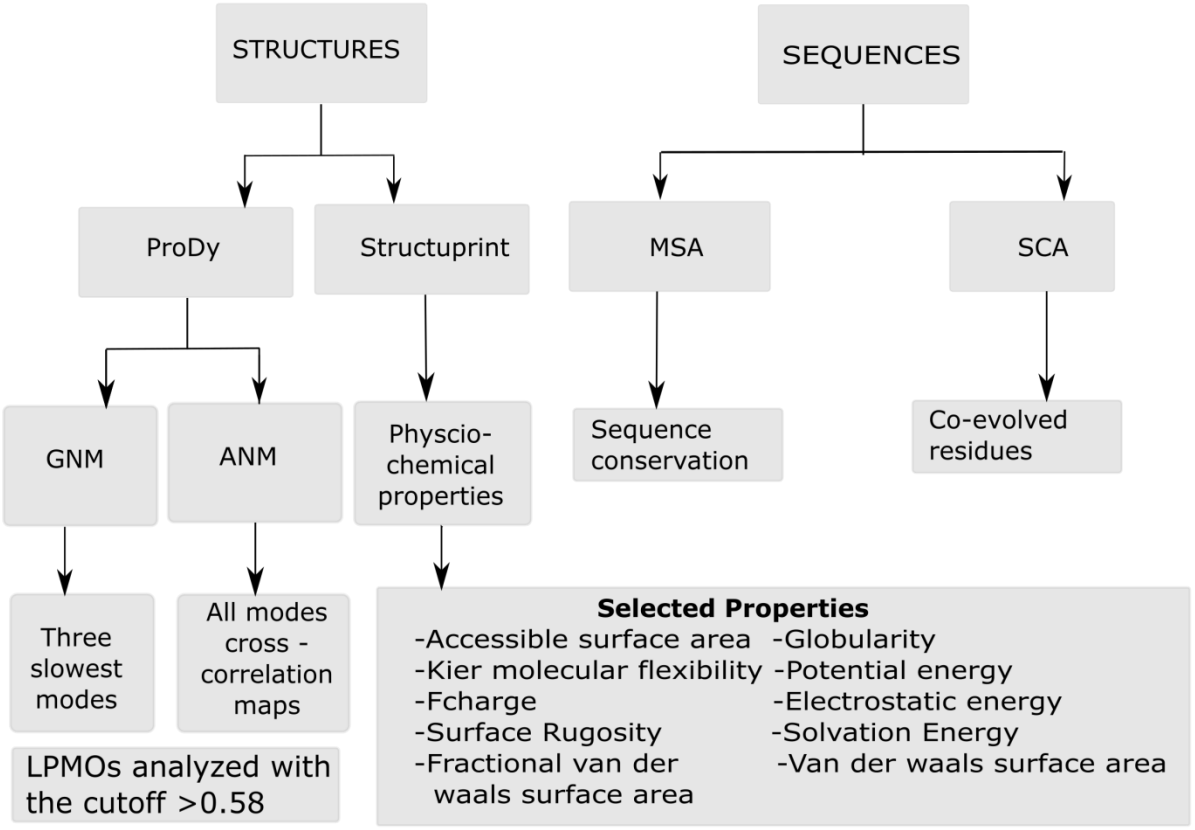
Figure 2: Three types of LPMOs classified based on the site of attack. Image adapted from [12].

HYPOTHESIS

In this study we hypothesize the following:

- As it has been observed, that all the four families of LPMOs have an unusually flat or planar surface to bind with crystalline polymers. It has been suggested that there is some degree of rigidity associated with the planar surface. We hypothesize that the planar surface is not rigid and has its inherent structural dynamics to bind optimally to the crystalline surface. In other words the question asked is whether the rigid planar substrate binding surface has any flexibility and can it be quantified.
- In terms of evolution, LPMOs are thought to have evolved in a divergent fashion, where the active site residues, specifically the three histidine residues bound to the divalent metal ion, are conserved (sequence and structure) across the four families. We hypothesize that there are other residues that have possibly coevolved leading to structural and functional conservation. In other words, we think that there are residues that play a role in allostery of LPMOs. In this study, we have attempted using statistical methods to identify some of the residues that influence allostery.
- While there are four families of LPMOs and currently 46 3D structures available, an exhaustive survey of physicochemical properties has not been conducted. Book et al in 2014 [9] and Mosses et al in 2016 [10] have shown that electrostatics and aromatic residue's distribution, respectively in LPMOs have a direct relation to substrate interaction. However, there are other physicochemical properties which have not been analyzed. Using structural bioinformatics approach we attempt to identify other physicochemical properties that may shed light on LPMO's structure and function.

MATERIALS AND METHODS



Data Retrieval:

CAZy:

It is a database of Carbohydrate Active enzymes (CAZy) i.e. information about the enzymes involved in the synthesis, transport and metabolism of carbohydrates[17].The database mainly includes glycoside hydrolases (assist in hydrolysis of glycosidic bonds in complex sugars), glycosyl transferases (catalyse the transfer of saccharide moieties from an activated nucleotide sugar to a nucleophilic glycosyl acceptor molecule), polysaccharide lyases (enzyme that catalyses the breaking of polysaccharide),carbohydrate esterase (enzymes that split esters into an acid and an alcohol) and carbohydrate binding families. As LPMOs are also carbohydrate active enzymes, it also comes under CAZy and we have used this database to retrieve the information of the four

different families of LPMOs i.e. AA9, AA10, AA11, AA13, which are tabulated in Table 1-4 [16].

Table 1: AA9 structures that were used in this study.

PDB ID	Uniprot Accession	Organism	AA9 PMO type	Citation
2vtc	Q7Z9M7	*Trichoderma reesei* QM6A	ND	[17]
2yet		*Thermoascus aurantiacus*	3	[17]
3eii		*Thielavia terrestris* NRRL 8126	1	[18]
3eja		*Thielavia terrestris* NRRL 8126	1	[16]
3zud		*Thermoascus aurantiacus*	3	[16]
4b5q		*Phanerochaete chrysosporium* K-3	1	[10]
4d7u	Q7SHI8	*Neurospora crassa* OR74A	2	[19]
4d7v		*Neurospora crassa* OR74A	2	[20]
4eir		*Neurospora crassa* OR74A	1	[20].
4eis		*Neurospora crassa* OR74A	3	[20]
4qi8	Q1K4Q1, Q873G1	*Neurospora crassa* OR74A	1	[21]
5acf		*Lentinus similis*	2	[21]
5acg		*Lentinus similis*	2	[22]
5ach		*Lentinus similis*	2	[22]
5aci		*Lentinus similis*	2	[22]
5acj		*Lentinus similis*	2	[22]

Table 2: AA10 structures that were used in this study.

PDB ID	Uniprot Accession	Organism	AA10 PMO type	Citation
4x29		*Unidentified entomopoxvirus/Melolontha melolontha entomopoxvirus*(MMEV)	ND	[22]
4x27		*Unidentified entomopoxvirus/Melolontha melolontha entomopoxvirus*(MMEV)	ND	[22]
4ow5	Q83389	*Unidentified entomopoxvirus/Melolontha melolontha entomopoxvirus*(MMEV)	ND	[23]
4yn2		*Unidentified entomopoxvirus*	ND	[23]
4yn1		*Anomala cuprea entomopoxvirus* CV6M	ND	[22]
2xwx	Q9KLDS	*Vibrio cholera O1 biovar EI Tor* str.N16961	ND	[23]
4gbo	Q47QG3	*Thermobifida fusca* YX	3	[23]
5ftz		*Streptomyces lividans* 1326	3	[24]
4oy7	Q9RJY2	*Streptomyces coelicolor* A3(2)	1	[25]
4oy8		*Streptomyces coelicolor* A3(2)	3	[26]
4oy6	Q9RJC1	*Streptomyces coelicolor* A3(2)	3	[27]
2lhs		*Serratia marcescens* BJL200	1	[27]

2ben		*Serratia marcescens* BJL200	1	[27]
2bem	083009	*Serratia marcescens* BJL200	1	[28]
5aa7	C7R4I0	*Jonesia denitrificans* DSM 20603	1	[29]
4alt		*Enterococcus faecalls* v583	1	[29]
4als		*Enterococcus faecalls* v583	1	[30]
4alr		*Enterococcus faecalls* v583	1	[31]
4alq		*Enterococcus faecalls* v583	1	[31]
4alc		*Enterococcus faecalls* v583	1	[31]
4ale		*Enterococcus faecalls* v583	1	[31]
4a02	Q838S1	*Enterococcus faecalls* v583	1	[31]
5fjq	B3PJ79	*Cellvibrio japonicas* Ueda107	1	[31]
3uam	Q3JY22	*Burkholderia pseudomallei*	ND	[32]
2yoy		*Bacillus amyloliquefaciens* DSM7	ND	[27]
2yox		*Bacillus amyloliquefaciens* DSM7	ND	[21]
2yow	E1UUV3	*Bacillus amyloliquefaciens* DSM7	ND	[33]
5IJU	Q9F9Q5	*Bacillus amyloliquefacens* ALKO 2718	ND	[33]

Table 3: AA11 structures that were used in this study.

PDB ID	**Uniprot Accession**	**Organism**	**AA11 PMO type**	**Citation**
4mah		*Aspergillus oryzae* RIB40	1	[33]
4mai		*Aspergillus oryzae* RIB40	1	[33]

Table 4: AA13 that were used in this study.

PDB ID	**Uniprot Accession**	**Organism**	**AA13 PMO type**	**Citation**
4opb	Q2U8Y3	*Aspergillus oryzae* RIB40	1	[33]

Protein Data Bank (PDB):

The respective LPMO structures we have downloaded from Protein Data Bank [34].

ProDy:

It is an open-source and free Python package [5]. It is used for protein structural dynamics analysis. It is suitable for development of various applications and for the interactive sessions. It performs structural and dynamics analysis anisotropic network model (ANM) and gaussian network model (GNM) were performed using ProDy. We generated different cross correlation maps and different output files using ProDy which were further visualized in VMD.

Normal mode analysis (NMA) is used for the representation of both the fast and slowest modes of a highly complex wired structure, such as a protein or any biomolecules [35]. Previously, such simulation was performed by Molecular Dynamics (MD) simulation. Due to high computation cost and time involved in a MD simulation, the coarse grained methods are getting more popular. While, the MD simulations focus more on exact forces between the atoms and then solve the equations in an appropriate manner, NMA approximate the equation for motion of the molecule which can be solved in a more exact manner. NMA can be simply described as study of harmonic potential wells by using analytical means and within a short time can provide the insight to more important dynamics of protein structures [36].

Tirion's "single parameter model" for NMA made the energy potential more simplified which explained NMA with uniform harmonic motion [37]. Later, Bahar and co-workers produced a much simpler version of NMA i.e. Elastic Network Model (ENM). Gaussian Network Model (GNM) was the important part of this model which was developed to study the contribution of topological constraints on the collective protein dynamics [38]. In GNM the Cα carbon residues were represents as nodes in a polymer network connected by springs, undergoing Gaussian distributed fluctuations. According to GNM, these fluctuations were assumed to be influenced by neighbouring atoms and their influence can be measured by the local packing density of residues around every single Cα residue [39]. In this, the contacts are represented as springs with uniform force constant which has been established by fitting the expression data. In general, the springs between the nodes (Cα residue) are connected if they are within 7Å distance [39].

Anisotropic Network Model (ANM) is a modification of GNM, the only difference is that the distance in GNM is in the form of vectors whereas the distance in ANM is in scalar form and the product of these scalar values results in anisotropic fluctuations by taking the second derivative of the potentials with respect to the displacement along any axis in a 3D space. ANM is better than GNM in terms of providing dimensionality, and also gives rise to excessively high fluctuations because GNM is penalized against any inter-residue fluctuation [40]. Now, as the accuracy level of ANM reduces the distance range for residues influencing motions increases and is often taken as 13.0Å. Despite of this less accuracy in local relative degrees of flexibility ANM is superior for accessing directional mechanism of motion as it has a $3N \times 3N$ Kirchhoff's matrix whereas GNM has $N \times N$ matrix [41].

Structuprint:

Structuprint is a software tool for two-dimensional representation of protein structures' surfaces. It is capable of generating animations or still images. It's free standalone software which is fully automated. The tool comes with a default database of 328 physico-chemical descriptors, which can be extended or substituted by user-provided ones [6]. Out of the 328 descriptors, we selected based on its ability to distinguish similar structures.

- **ASA-** Accessible surface area, first described by Lee & Richards in 1971, is the surface area of the biomolecules that is accessible to the solvent. Its measurement is usually described in units of square angstrom i.e. standard unit of measurement in microbiology. It uses the rolling ball algorithm for the purpose of calculations. This algorithm uses a sphere (of solvent) of a particular radius to 'probe' the surface of the molecule. Water accessible surface area calculated using a radius of 1.4 ˚A for the water molecule. A polyhedral representation is used for each atom in calculating the surface area.

- **E_ ele-** Electrostatic component of the potential energy. It is a potential energy measured in joules resulting from conservative Coulomb forces. An object may have electric potential energy due to two key elements: its relative position to other electrically charged objects and its own electric charge .The term "electric potential energy" is used to describe the potential energy in systems with time variant electric fields while the term "electrostatic potential energy" is used to describe the potential energy in systems with time variant electric fields

- **E_sol -** Solvation energy is the amount of energy linked with dissolving a solute in a solvent. It has two processes endothermic and exothermic i.e. having positive and negative numbers respectively. The energy of solvation is sometimes found by comparing the hydration energy i.e. the amount of energy released when the solute particles bond with the solvent and the lattice energy i.e. the amount of energy needed to break the bonds of the solute.

- **E-** Potential energy. It is a potential energy measured in joules resulting from conservative Coulomb forces. An *object* may have electric potential energy due to two key elements: its relative position to other electrically charged *objects* and its own electric charge. The term "electric potential energy" is used to describe the potential energy in systems with time variant electric fields while the term "electrostatic potential energy" is used to describe the potential energy in systems with time variant electric fields

- **FCharge-** Total charge of the molecule (sum of formal charges) is the charge assigned to an atom in a molecule. It assumes a chemical bond where electrons are shared equally between atoms, regardless of relative electronegativity. The formal charge of any atom in a molecule can be calculated by the following equation

$$FC=V-N\frac{B}{2}$$

  where, v= number of valence electrons of the neutral atom isolation.

  N=number of nonbonding valence electrons on this atom in the molecule.

  B=total number of electrons shared in bonds with other atoms in molecule.

- **Glob-** Globularity or inverse condition number (smallest eigenvalue divided by the largest eigenvalue) of the covariance matrix of atomic coordinates. A value of 1 indicates a perfect sphere while a value of 0 indicates a two- or one-dimensional object.

- **KierFlex-**Kier molecular flexibility index: (KierA1)(KierA2) it is the measure of molecular flexibility which is derived from the Kier alpha modified shape descriptors ${}^{1}_{\alpha}\kappa$ and ${}^{2}_{\alpha}\kappa$:

$$\Phi = \frac{({}^{1}\kappa \cdot {}^{2}\kappa)}{N_{SA}}$$

  ${}^{1}\boldsymbol{k}$ and ${}^{2}\boldsymbol{k}$ - Kier shape indices

N_{SA} - the number of non-hydrogen atom in the molecule

- **VSA _ FHYD-** Fractional hydrophobic van der Waals surface area. This is the sum of the vi such that |qi| is less than or equal to 0.2 divided by the total surface area. The vi are calculated using a connection table approximation.
- **TPSA-** Polar surface area calculated using group contributions to approximate the polar surface area from connection table information only.
- **VSA -**Van der Waals surface area. A polyhedral representation is used for each atom in calculating the surface area
- **vsurf R-** Surface rugosity. is a measure of small-scale variations of amplitude in the height of a surface.

Statistical Coupling Analysis:

Coevolved residues

The correlated mutations refer to the pairs of position with clear pattern of co variation i.e. in any particular multiple sequence alignment. The mutation of one residue will be compensated along the evolution by the mutation of a neighboring residue. The protein structure can also be predicted by the detection of correlated mutations in multiple sequence alignment.

The method to detect correlated mutations was a weak predictor of proximity between protein structures as described by Göbel et al in 1994. Later, the method was improved by combining it with other sequence based features like conservation and hydrophobicity. Though, these contacts are not very accurate but are very useful in filtering structural models and generating structures for *ab initio* simulation [42].

Allostery

The term allostery can be explained by splitting it in its components, first is "allos" that means others and second is "stereos" that refers to solid or any three dimensional structure. Allostery is generally defined as any change in the conformation or shape of the molecule. According to this definition of allostery there are two central dogmas to it. First, that in the absence of any ligand there exist two conformations, which are influenced by the ligand's equilibrium constant. Second, that allostery always means change of shape. As allostery mainly focuses on structure of the

molecule and altered by conformational changes but it has seen that dynamic fluctuations also play a huge role in allostery. So, currently allostery is considered as thermodynamic phenomena and it may or may not be a result of any conformational change and in particular the absence of marked shape changes does not imply any allostery.

The current definition of allostery is broader than the earlier one. This definition of allostery classifies the allosteric proteins into three types, Type I: consists of the proteins that the allosteric changes are governed by entropy, Type II: includes the allosteric proteins whose allostery is governed by enthalpy as well as entropy, Type III: includes allosteric proteins that show allostery strongly influenced by enthalpy only. It has been seen that under suitable environment conditions the increase or decrease in catalytic activities controls the proteins and ligand transport and also coordinates enzymatic and signaling pathways. This current view of allostery has vast implications in identifying new allosteric switches and drug targets [44].

Statistical Coupling analysis (SCA)

It is a method for the analysis of coevolution of amino acids explaining the structural basis for allostery. The basic concept of SCA is that if there are some relevant relationships between two amino acids, then they must have coevolved irrespective of the mechanism of their evolution. SCA contributes to two main findings about proteins: 1) most of the amino acids in protein molecules have evolved almost independently, showing a very weak coupling among them, 2) a small amount of amino acids have strongly coevolved and termed as sectors. These sectors are associated with conserved functional properties like allosteric regulation, catalysis, signal transduction etc. In a single protein more than single sector can also occur, which shows that the different phenotypes have evolved independently. Hence, sectors represent the fundamental units that have conserved structure. These proteins display the capacity for evolving novel allosteric regulation and communication. Sectors wire the active site to the multiple distant surface positions, representing “hot-spots” for the emergence of allosteric control in proteins [44].

In case of LPMOs the active site residues specifically histidines are conserved across the four families. We have used SCA to see whether, there are other residues (apart from the active site) that have possibly coevolved leading to structural and functional conservation? If so, are there residues that play a role in allostery of LPMOs?

In SCA, we consider four positions as *i, j, k, l* of a hypothetical protein and corresponding multiple sequence alignment of the protein family having sufficiently large and diverse protein molecules. From the analysis we can easily derive two things. First, if out of the four considered sites one site, *i* neither contributes to the folding nor to the function of the protein, then the corresponding amino acid frequencies in the MSA should be unconstrained so should use their mean values in all proteins. On the other hand, if sites *j, k,* and *l* are making some contribution, these sites should have a deviated distribution of amino acids from the mean values and the amount of their deviation will provide a quantitative measure of the evolutionary conservation. Second, if two sites, *i* and *j* are functionally related to each other then, they must exhibit mutual evolutionary constraint between these sites, which means that the distribution of residues at site *j* should depend on those at site *i* [45].

In SCA, first we perform multiple sequence alignment of the protein sequences from AA10 family (2558 sequences) with an identity cut-off of 90%, which resulted in 143 sequences. Then, we converted the MSA file in .free format and loaded it into the MATLAB. The .free format was checked to include the sequence corresponding to *Serratia marcescens*' AA10 structure (pdb id: 2bem). To ensure better results we truncated off the residues with gap frequency greater than 20% first, then we look at the uniform or homogeneous distribution of residues among proteins. Then we do the sector identification, calculating the correlations among the top eigenmodes of SCA position correlation matrix. If the sequences are highly clustered then, the sector identification can also be guided by the patterns divergence. Then, we compute the similarity matrix using *sim_seq(algn)* function generating a correlation matrix and a histogram inferring the same results. We can also generate the matrix of covariance, it will predict the same information but scaling of both will be different. After that, we measure the degree of positional conservation by an information theoretic quantity called as Global Kullback-Lieber relative entropy which can be given as $D(f(a,i)||q(a))$, capturing the divergence among the observed frequency of sequence amino acids and frequency given in non-redundant database. The function *cons.m* calculates this frequency and gives us a $20\times N$ position matrix of relative entropies. Then we perform the SCA which gives us positional correlation matrix and a sequence correlation matrix and shows the similarity among the AA10 family, showing that the proteins of AA10 family are biased towards more conserved positions. After completing the SCA calculations and generating different variables we do spectral decomposition, in which we analyze the C_p matrix generated

in SCA. This analysis is carried out to check the existence of non-trivial correlations between positions indicates that treating the amino acids as the basic units of proteins is not the most informative representation. We should rather use reparameterization of the protein in which the units of proteins are the collective groups of amino acids that coevolving as per the positional correlation matrix and are called "sectors". In this reparameterization, first we do the eigenvalue decomposition which mathematically transforms a current representation of a system in which variables are correlated into new variables having the property of being uncorrelated to each other, and are more informative.

The original matrix is written as a product of three matrices: X=VDV', where D is a diagonal matrix of eigenvalues and columns of V contain the associated eigenvectors. For the SCA positional correlation matrix, each eigenvector represents a weighted combination of sequence positions i.e. an eigenmode and the associated eigenvalue indicates the statistical importance of that mode. Then, we see that how many of the derived eigenvalues are statistically significant. To ensure this, we compare the spectral decomposition for the actual alignment with that for many instances of randomized alignments, where the amino acids are scrambled independently down each column. This manipulation removes all functional correlations and retains only the spurious correlations that are possible due to finite sampling. The function spectral_decomp.m carries out this calculation and returns a structure with the eigenvalue decomposition of actual and randomized alignments and makes a plot of the eigenspectra. Then we mapped the retrieved sectors to the structure of *Serratia marcescens*' (pdb id: 2bem) in PyMOL to understand the spatial arrangement of the residues predicted from SCA.

RESULTS AND DISCUSSION

Multiple structure alignment

We have superposed structures for the three families of LPMO in PyMOL software using the command “align all” and calculated RMSD the values. It is the measure of the average distance between the atoms. RMSD is inversely proportional to the structural similarity of the proteins i.e., greater the RMSD less similar the structures are vice versa.

Table 5: Root Mean Square Deviation values of structural superposition. (a) For AA9 the reference structure is 2bem, (b)For AA10 the reference structure is 2bem (c) For AA11 the reference structure is 4MAH.

PDB ID	RMSD
2YET	0.546
3EII	1.024
3EJA	0.975
3ZUD	0.584
4B5Q	1.031
4D7U	0.791
4D7V	0.812
4EIR	0.829
4EIS	0.765
4QI8	1.010
5ACF	1.065
5ACG	1.057
5ACH	1.014
5ACI	0.993
5ACJ	1.043

PDB ID	RMSD
2BEN	0.303
2LHS	1.273
2XWX	0.491
2YOW	0.445
2YOX	0.474
2YOY	0.440
3UAM	0.700
4A02	0.880
4ALC	0.673
4ALE	0.688
4ALQ	0.686
4ALR	0.686
4ALS	0.688
4ALT	0.674
4GBO	1.636
4OW5	16.995
4OY6	4.726
4OY7	1.859
4OY8	4.778
4X27	16.990
4X29	16.973
4YN1	7.545
4YN2	7.482
5AA7	0.737
5FJQ	0.837
5FTZ	0.796

PDB ID	RMSD
4MAI	0.083

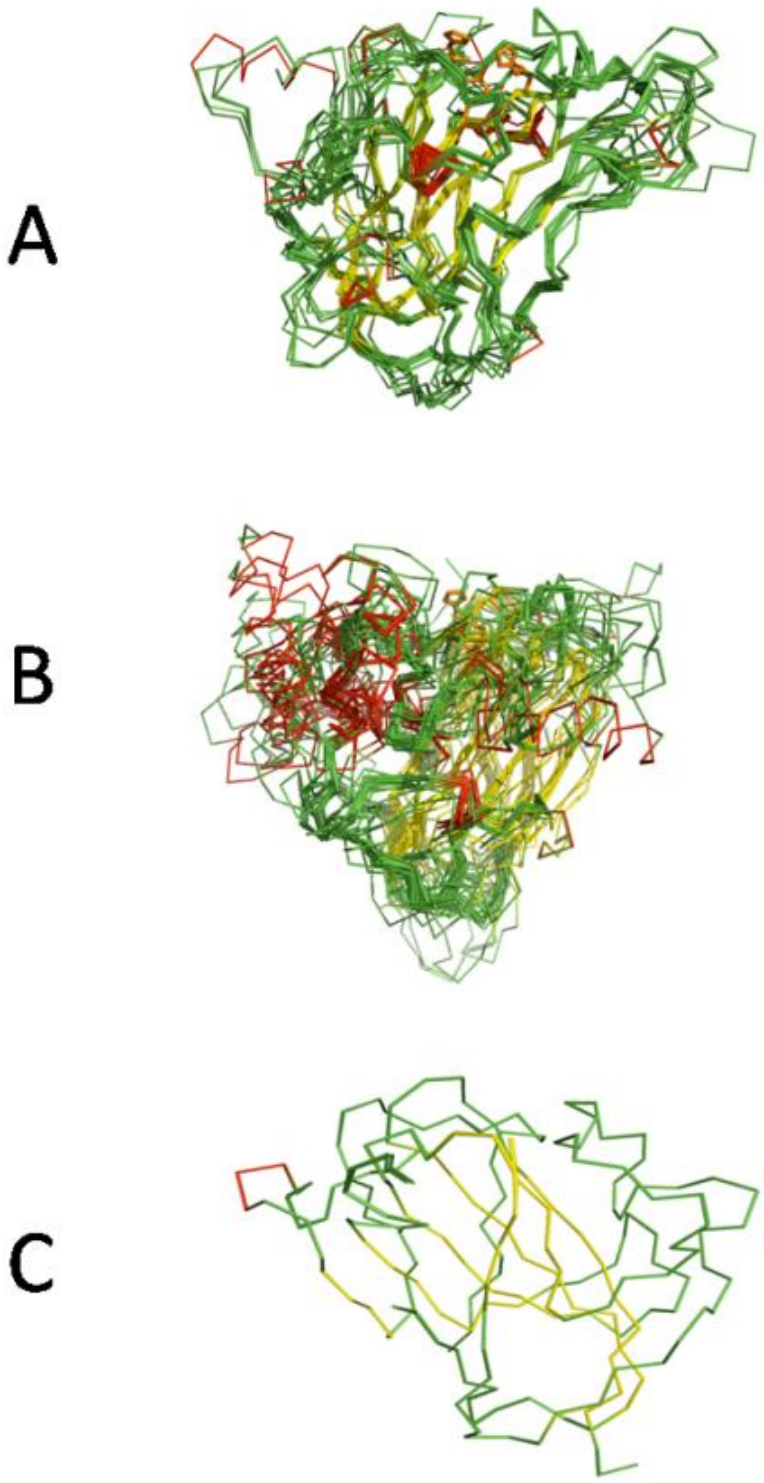


Figure 3: Structural superimposition of (a) AA9, (b) AA10, and (c) AA11 structures.

Elastic Network Models

Gaussian Network Model (GNM) is a model that stores the normal mode data describing the intrinsic dynamics of the protein structure and Kirchoff matrix that will be obtained. It is used for the observation of global dynamic behavior. The parameter cut off distance is generally used to build the network to determine whether the residues are in contact or not. The fluctuations obtained from GNM are illustrated in a single dimension space (N).The graph retrieved has node index on the x-axis and B-factor values on the y-axis. The information retrieved from the graph is the correlation of the theoretical and experimental B-factors.

Table 6: GNM B-factor correlation values for (a)AA9, (b)AA10, (c)AA11, and (d)AA13

PDB ID	B-FACTOR
2VTC	0.68
2YET	0.67
3EII	0.57
3EJA	0.49
3ZUD	0.70
4B59	0.83
4D7U	0.42
4D7V	0.35
4EIR	0.75
4EIS	0.57
4QI8	0.64
5ACF	0.61
5ACG	0.54
5ACH	0.45
5ACI	0.63
5ACJ	0.59

(a)

PDB ID	B-FACTOR
2BEM	0.63
2BEN	0.27
2LHS	NAN
2XWX	0.32
2YOY	0.40
2YOW	0.40
2YOX	0.28
3UAM	0.67
4A02	0.32

4ALC	0.52
4ALE	0.52
4ALQ	0.52
4ALR	0.53
4ALS	0.52
4ALT	0.52
4GBO	0.36
4OW5	0.70
4OY6	0.62
4OY7	0.70
4OY8	0.60
4X27	0.67
4X29	0.62
4YN1	0.57
4YN2	0.64
5AA7	0.57
5FJQ	0.52
5FTZ	0.50
5IJU	0.55

(b)

PDB ID	B - FACTOR
4MAH	0.72
4MAI	0.49

(c)

PDB ID	B-FACTOR
4OPB	0.52

(d)

Kundu et al have showed that GNM models are more reliable if they have a B-factor correlation of theoretical vs. experimental of 0.58 [41]. From the GNM results, we identified nine structures in AA9 (pdb id: 2vtc, 2yet, 3zud, 4b5q, 4eir, 4cib, 5acf, 5aci, and 5acj) and 9 structures of AA10 (pdb id: 2bem, 3uam, 4ow5, 4oy6, 4oy7, 4oy8, 4x29, 4yn2, and 4yn1) which has a B-factor correlation value of 0.58 or above. In the case of AA11 there is only one structure (pdb id: 4mah). However, the structure in AA13 had a β-factor correlation of 0.52. The selected 19 structures were analyzed filter by building ANM models.

Anisotropic Network Model (ANM) is a model for normal mode analysis of proteins. It is used for exploring the relation between function and dynamics for many proteins. It is essentially an ENM for the Cα atoms with a step function for the dependence of the force constants on the inter-particle distance. To explain the internal motions of a protein subject to a harmonic potential it

represents the biological macromolecule as an elastic mass-and-spring network ANM is an extension of the GNM to three coordinates per atom, thus accounting for directionality. ANM calculates the direction of fluctuations and magnitudes in 3N space. It provides the directional motions collectively that are useful in a protein both functionally and biologically. The graphical result obtained from ANM is different for each structure which represents residues on the x-axis and y –axis which illustrates the red regions as positive correlation and blue region shows negative cross correlation. The correlation values within each mode ranges from −1.0 to 1.0 corresponding to positive and negative correlation, respectively. Figure 4 shows the cross correlation information of nine AA9 structures. These 9 structures were shortlisted as the GNM B-factor correlation had a value of 0.58 or higher, as suggested by Kundu et al [41]. Figure 5 shows the cross correlation information of nine AA10 structures. These 9 structures were shortlisted on the basis of GNM B-factor correlation, where they had a value of 0.58 or higher. Figure 6 shows the cross correlation map of 4mah as it is the side structure but had a higher B-factor cross correlation.

Based on the ANM results depicted for 9 AA9 structures (Figure 7), in the case of 2yet and 2vtc the planar surface shows flexibility at the loop regions where the tyrosine residues are present that aid in substrate binding. Looking at the first three slowest modes the loops seem to be in a push-pull motion helps the protein to move across the crystalline surface of polymers.

In comparison the other AA9 structures (3zud, 4eir, 4b5q, 4qi8, 5acf, 5aci, and 5acj) have surface flexibility concentrated on one “side” of the protein.

In all the nine AA9 proteins the active site area consisting of His-His-Tyr/Phe residues show least flexibility. The results indicate that in AA9 the planar surface interacting with the substrate is not rigid and the loop regions have coordinated motions for LPMO to bind on a flat crystalline surface.

In the 9 structures of the aa10 the loop regions on the substrate binding site are most mobile in all the three slowest modes. Interestingly, these motions are relatively closer to the active site region compared to AA9. In 4oy6, 4x27, 4oy7, 4x29, and 4yn2 show slightly different motions in the loop regions away from the active site. It is highly possible that the loops away from the active site incorporate allosteric behavior of LPMOs while binding to the substrate.

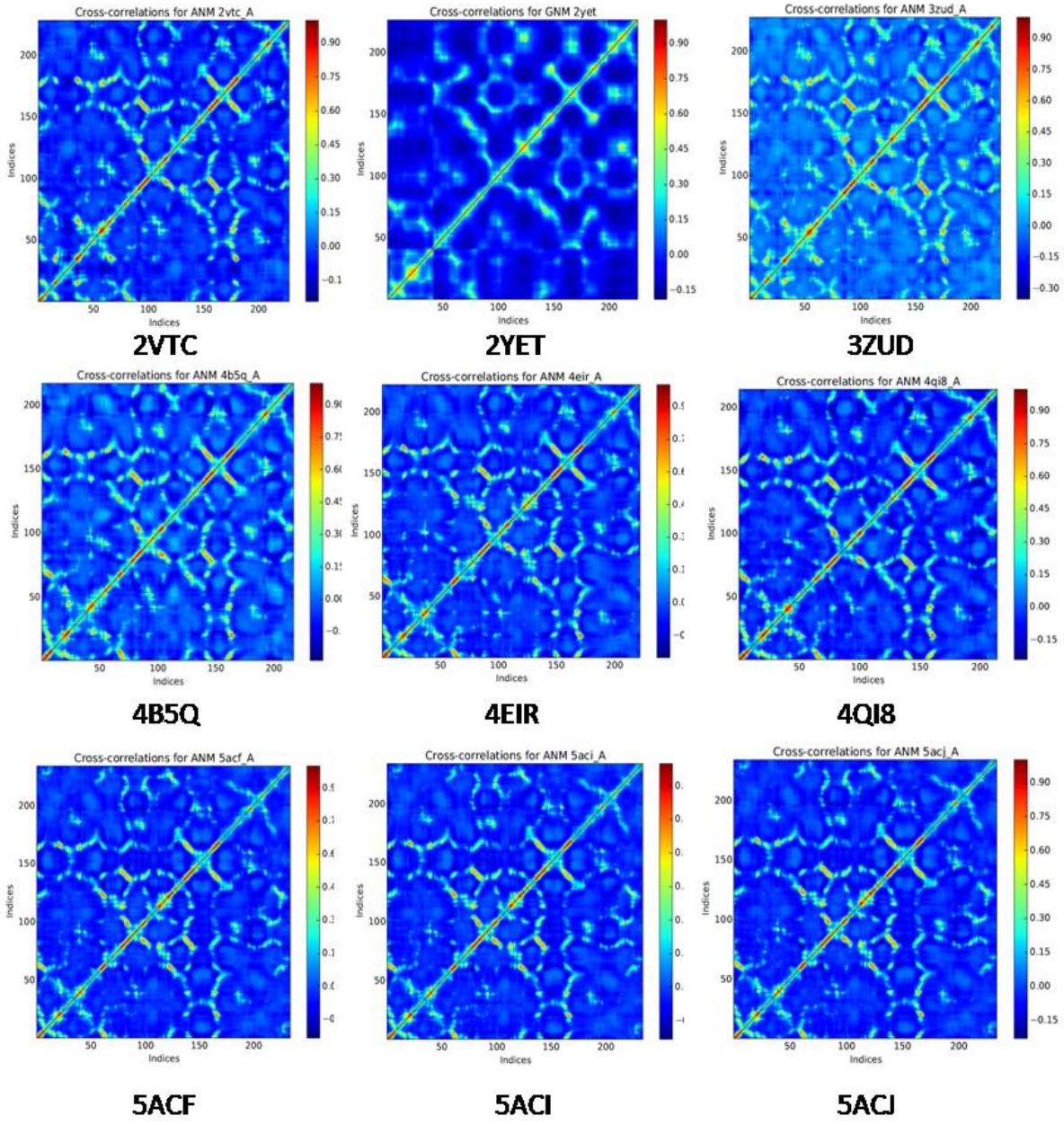


Figure 4: Cross correlation map for AA9 in all modes having B-factor greater than 0.58. These values are in accordance with the average correlation coefficient value for GNM, i.e. 0.58 [41].

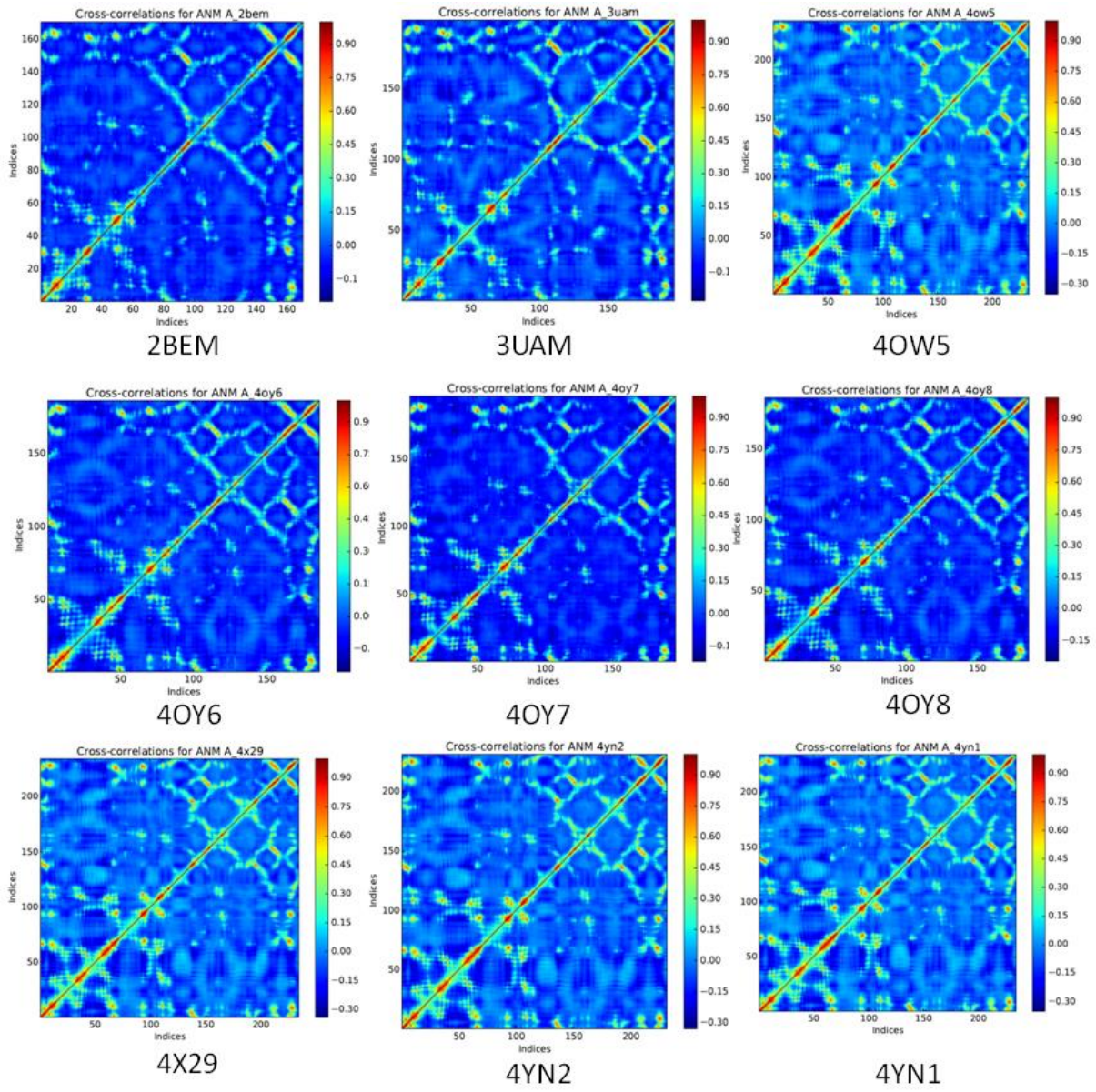


Figure 5: Cross correlation map for A10 having B-factor greater than 0.5. These values are in accordance with the average correlation coefficient value for GNM, i.e. 0.58 [41].

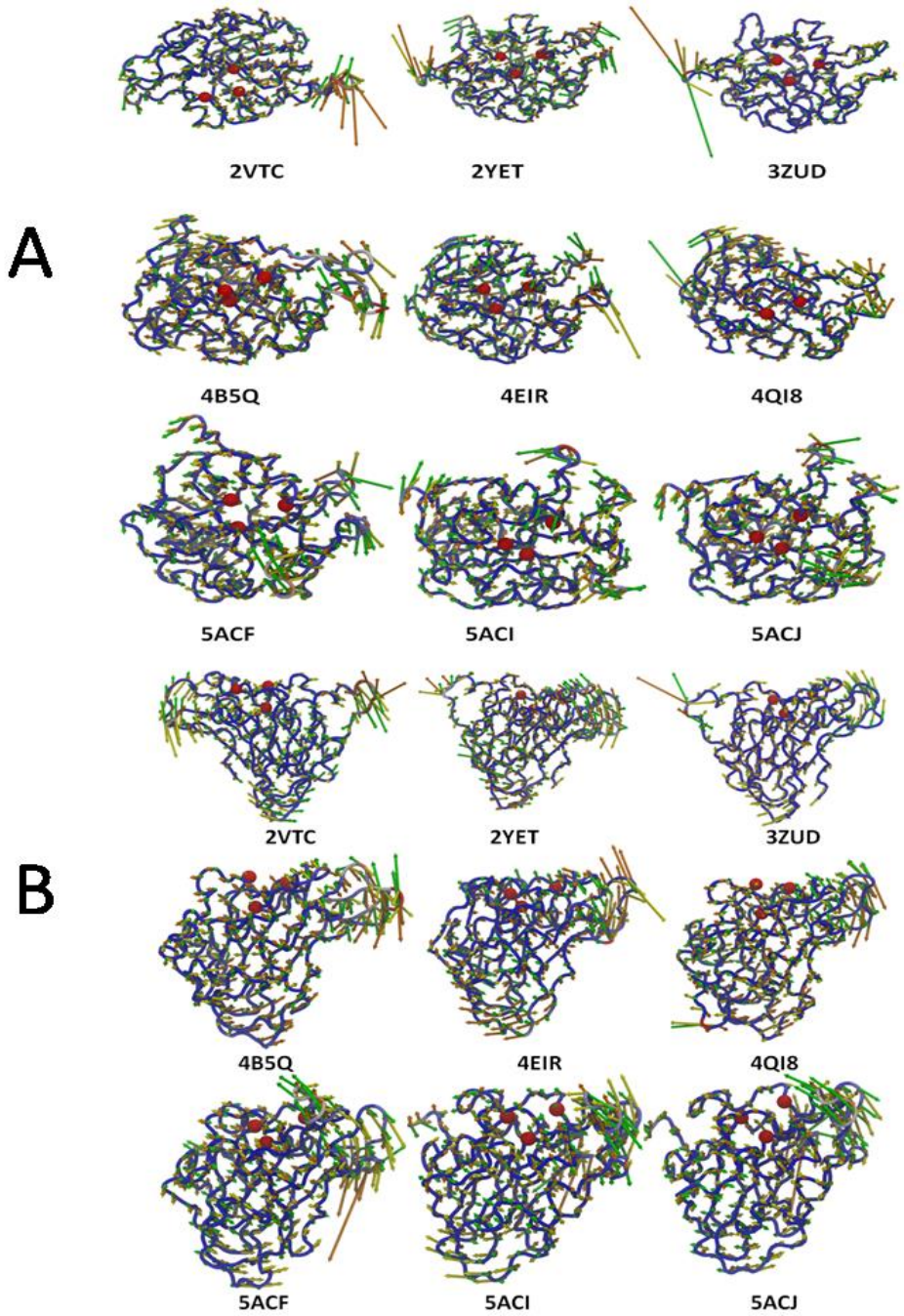


Figure 6: First three slowest modes of AA9 (A) Top view showing the active site residues as spheres with planar surface (B) Side view. The first three slowest modes are plotted using NMwiz of VMD visualization tool. The protein is shown as Cα trace, where the region colored red are most mobile and the regions colored blue are least mobile. The first slowest mode is represented with orange arrows, the second slowest mode is represented with yellow arrows, and the third slowest mode is represented with green arrows. The histidines and tyrosine/phenylalanine in the active site are shown as red spheres.

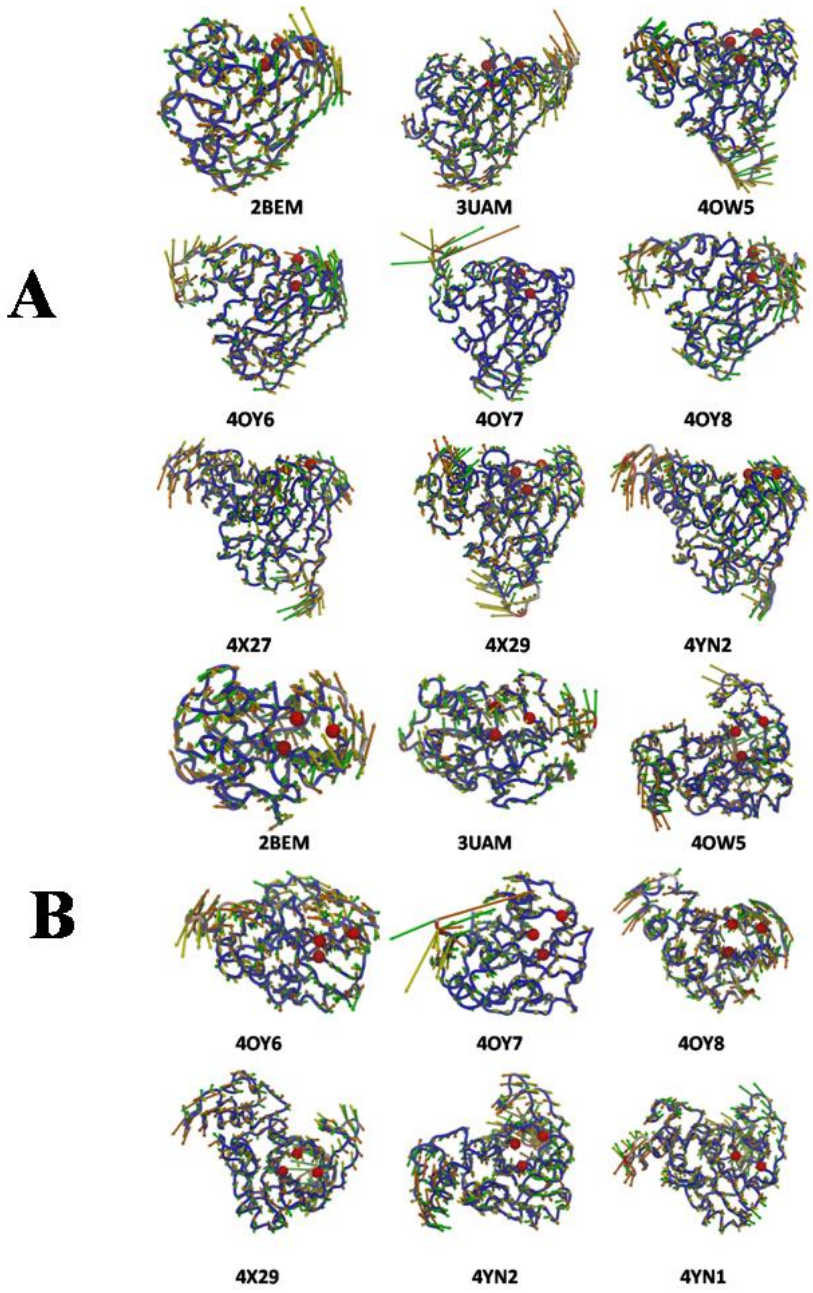


Figure 7: First three slowest modes of AA10 (A) Top view showing the active site residues as spheres with planar surface (B) Side view. The first three slowest modes are plotted using NMwiz of VMD visualization tool. The protein is shown as Cα trace, where the region colored red are most mobile and the regions colored blue are least mobile. The first slowest mode is represented with orange arrows, the second slowest mode is represented with yellow arrows, and the third slowest mode is represented with green arrows. The histidine and tyrosine/phenylalanine in the active site are shown as red spheres.

Physico-chemical properties

Structuprint can describe total of 328 properties and out of these properties we have considered only 11 properties. Figure 8 shows the variations on the basis of water **accessible surface area** for AA9 structures. It is calculated using a radius of 1.4Å for water molecule. The results depict that most of the structures of AA9 were lying between the range 217-306 which infers that the residues have less accessibility to the water molecule whereas some of them showed higher values ranging upto 395 which infers that some of the structures were having more accessibility to the water molecule. Figure 9 shows the **electrostatic potential energy** as calculated by Structuprint for AA9 structures. Electrostatic potential energy used to describe the potential energy in systems with time variant electric fields. Its scale reads −38 to 38 with a red to blue scale representing the negative and positive electrostatic potential respectively. The graphs have varying abundance of negative electrostatic potential energy.

Figure 10 shows the plots for **solvation energy** i.e. the process of attraction and association of molecules of a solvent with molecules or ions of a solute or the amount of energy linked with dissolving a solute in a solvent. The results depict that some of the structures of AA9 have more solvation energy ranging from -110 to -56 and most of them had large negative values ranging from -164 to -110. Figure11 shows the **potential energy** calculated for AA9 structures. The potential energy is measured in joules resulting from conservative Coulomb forces. Its scale ranges from −64 to 64. All structures have a net negative potential energy. Figure 12 shows the **total charge of the molecule** (sum of formal charges). It is the charge assigned to an atom in a molecule. The red dots in the graph illustrate the presence of the total negative charge and the blue dots illustrate the presence of the total positive charge. Its scale ranges from -1 to 1. Some of the protein structures have more total negative charge than that total positive charge and vice versa. Figure13 shows the **globularity** i.e. inverse condition number (smallest eigenvalue divided by the largest eigenvalue) of the covariance matrix of atomic coordinates. A value of 1 indicates a perfect sphere while a value of 0 indicates a two- or one-dimensional object. It tells us about the degree of spherical or globe nature of the protein surface. Most of the values were ranging between 0 and 0.5 on a scale of 0 to 1.0.Figure 14 shows the **Kier-Flex molecular flexibility index** which is derived from the Kier alpha modified shape descriptors. Most of the values were ranging between 1.0 and 3.5 only few of them were in the range of 3.5 and 6.0.

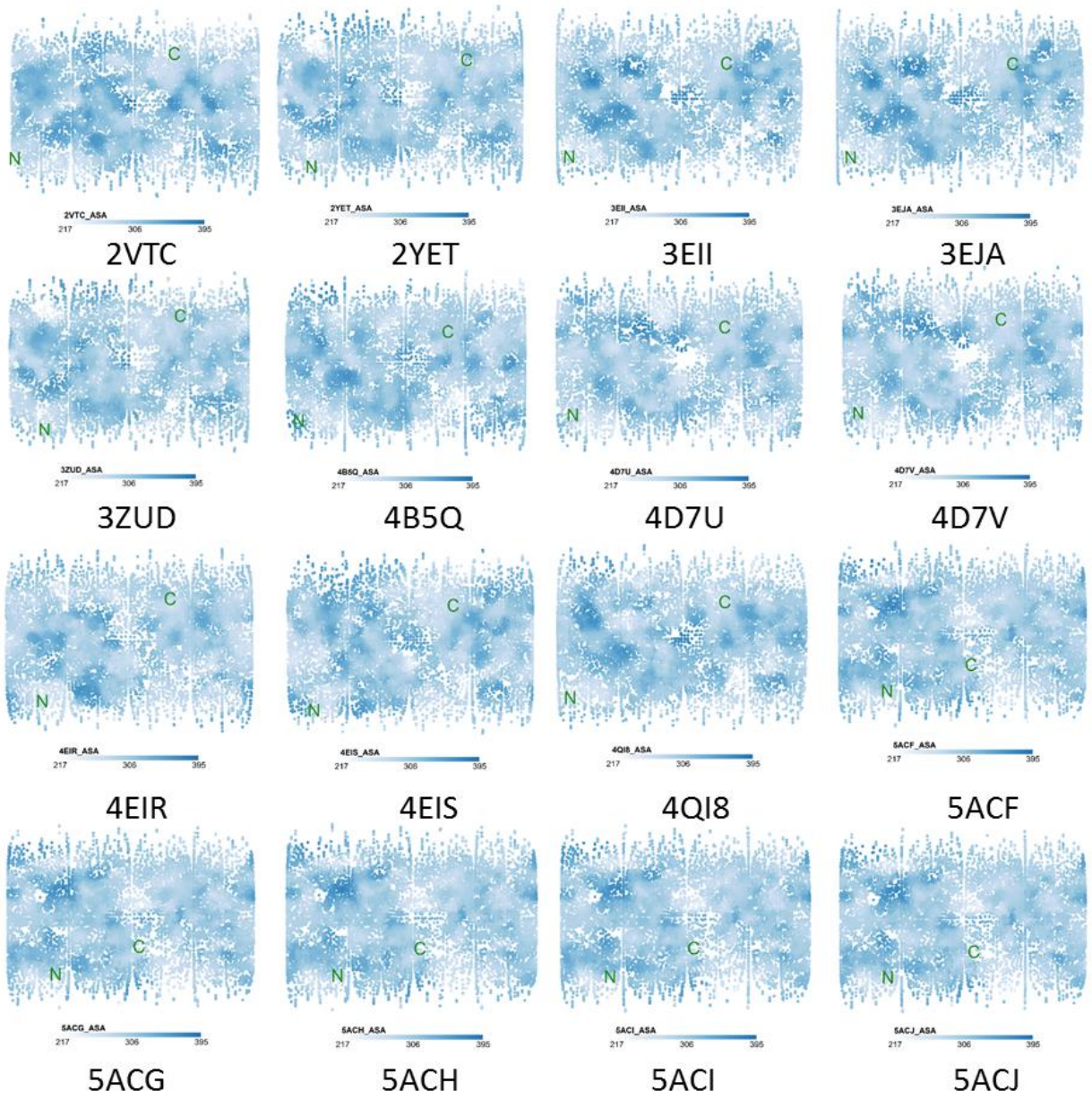


Figure 8: Plots showing Accessible Surface Area of 16 structures of AA9 calculated by Structuprint.

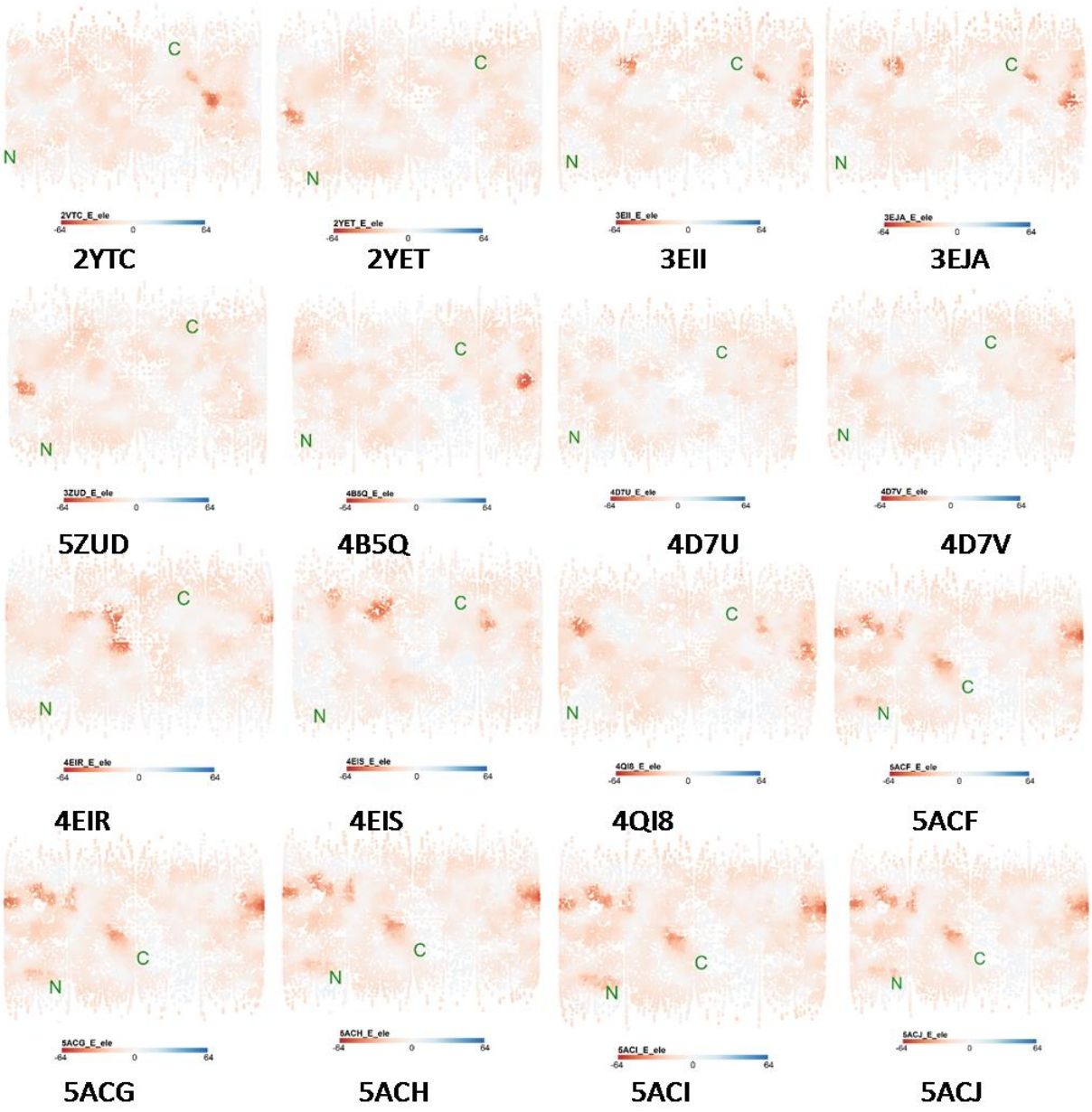


Figure 9: Plots showing electrostatic energy of 16 structures of AA9 calculated by Structuprint.

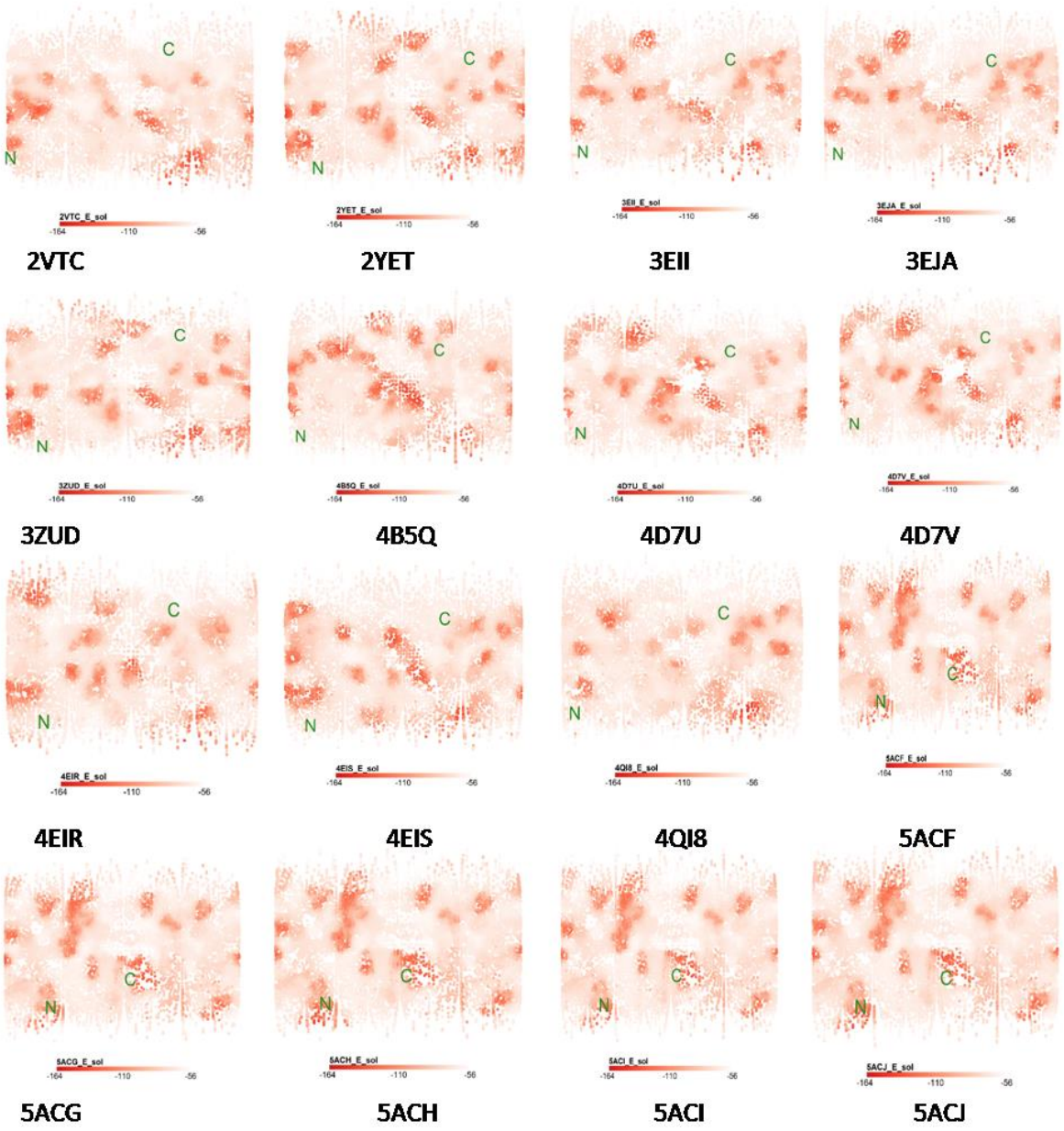


Figure 10: Plots showing solvation energy of 16 structures of AA9 calculated by Structuprint.

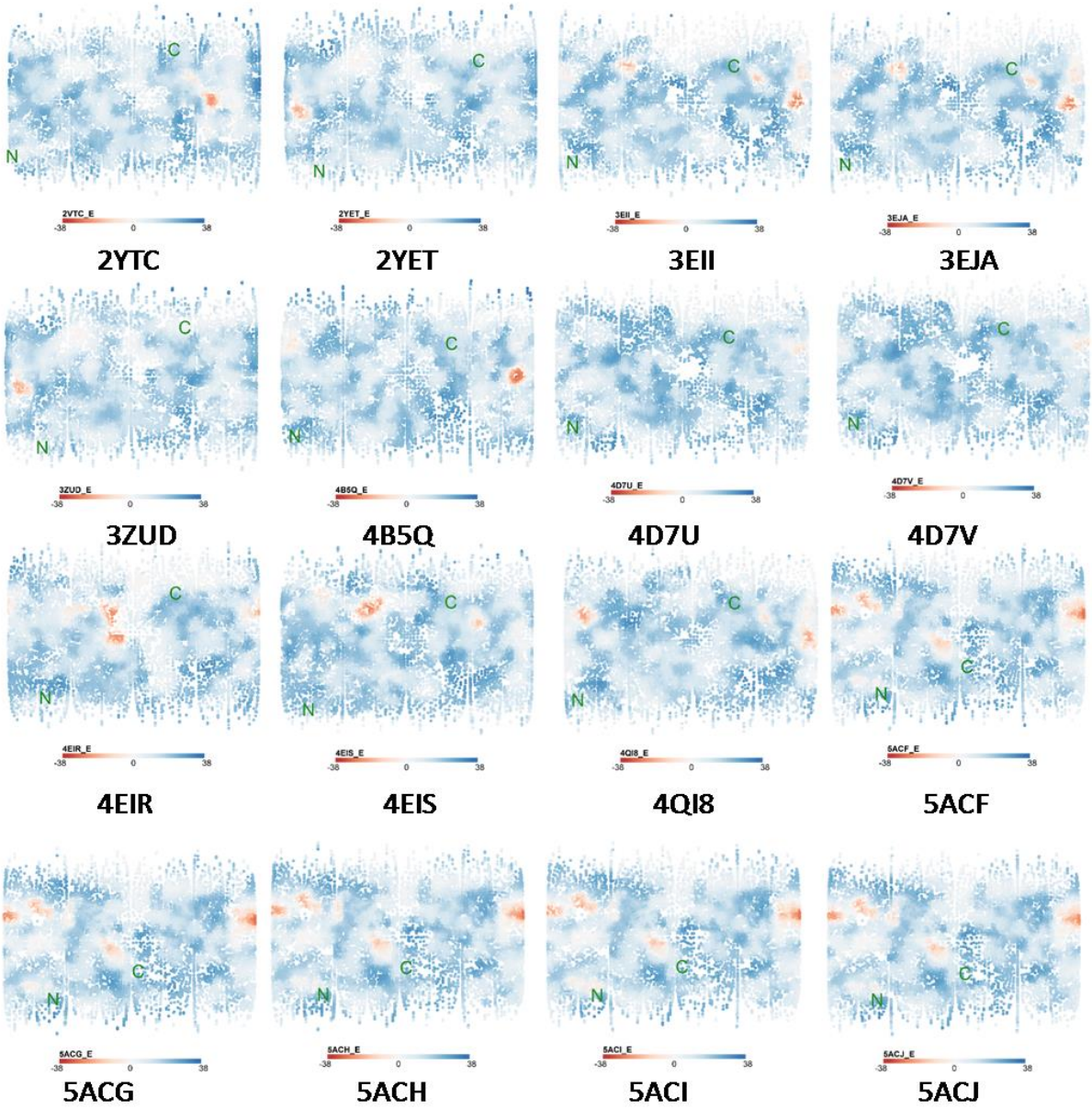


Figure 11: Plots showing potential energy of 16 structures of AA9 calculated by Structuprint.

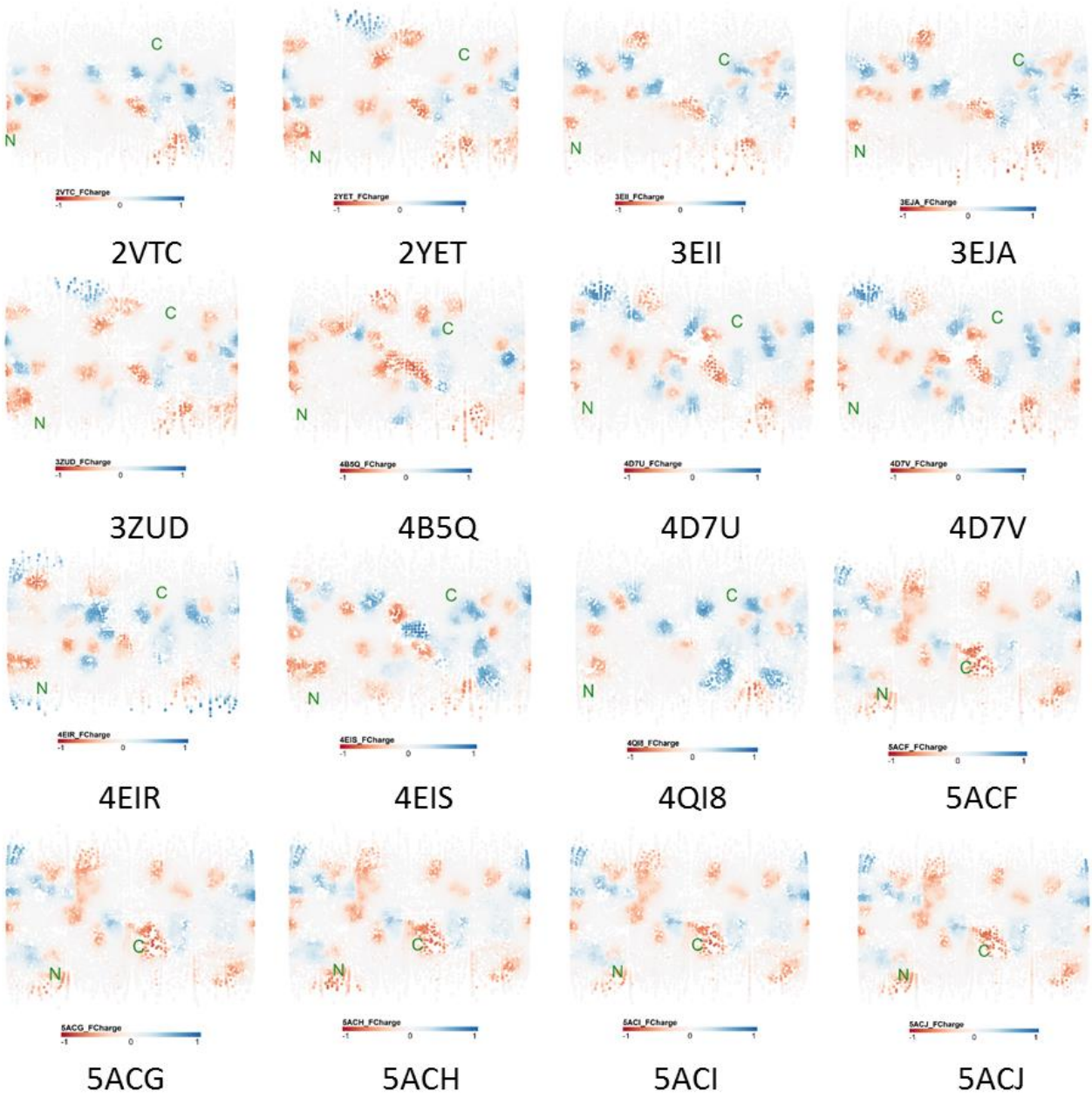


Figure 12: Plots showing formal charge of 16 structures of AA9 calculated by Structuprint.

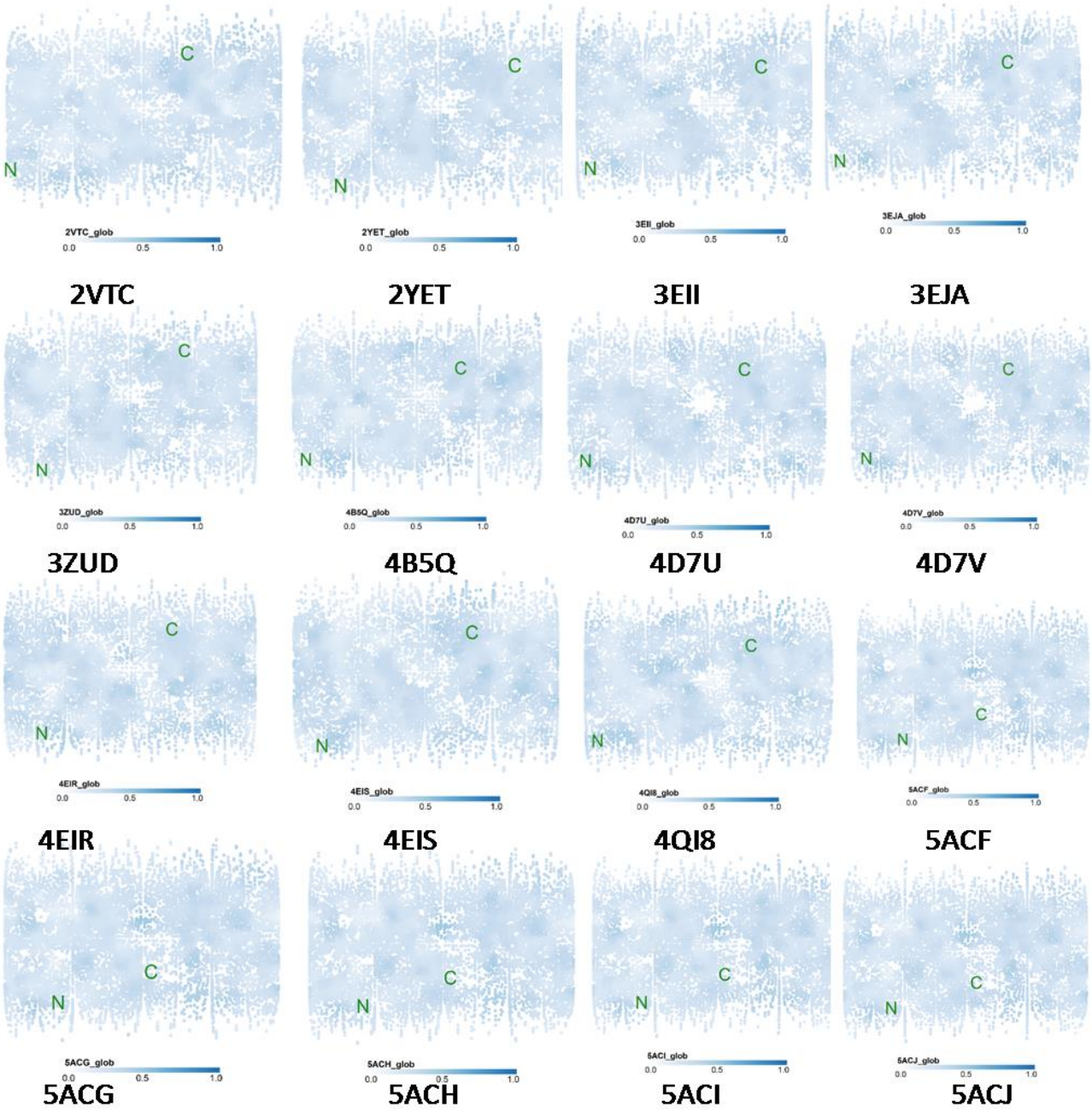


Figure 13: Plots showing globularity of 16 structures of AA9 calculated by Structuprint.

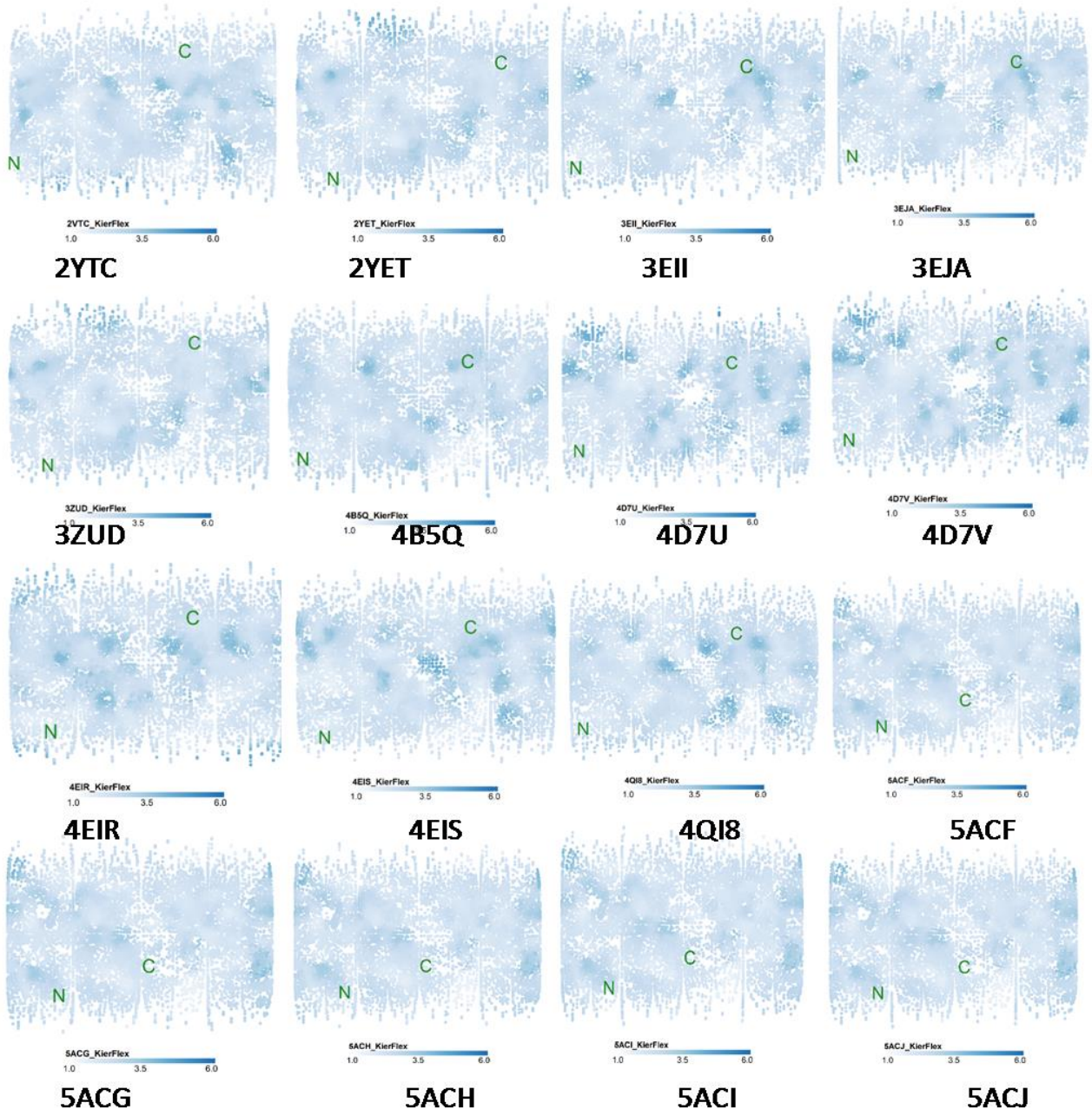


Figure 14: Plots showing Kier molecular flexibility of 16 structures of AA9 calculated by Structuprint.

Figure 15 shows **fractional van der waals** i.e. This is the sum of the vi such that |qi| is less than or equal to 0.2 divided by the total surface area. The vi are calculated using a connection table approximation. Values ranging between 91.5 and 147.0 were more than values ranging from 36.0 to 91.5.

Figure 16 shows **total polar surface area** which is area of a molecule defined as the surface sum over all polar atoms, primarily oxygen and nitrogen. It has scale ranging from 56 to 134. Most of the proteins have values ranging from 56 to 95 but there are few families with more values ranging from 95 to 134. Figure 17 shows the **van der waals surface area** i.e. A polyhedral representation is used for each atom in calculating the surface area. The results depict that most of the structures of AA9 were lying between the range 159.5-226.0 on a scale of 93.0-226.0.

Figure18 shows the **rugosity** i.e. is a measure of small-scale variations of amplitude in the height of a surface. Values ranging from 1.0-1.5 were more comparatively to the values under the range 1.5-2.0 on a scale of 1.0-2.0. Figure 19 shows the variations on the basis of water **accessible surface area** for AA10 structures. It is calculated using a radius of 1.4Å for water molecule. The results depict that most of the structures of AA10 were lying between the range 217-306 which infers that the residues have less accessibility to the water molecule whereas some of them showed higher values ranging up to 395 which infers that some of the structures were having more accessibility to the water molecule. Figure 20 shows the **electrostatic potential energy** as calculated by Structuprint for AA10 structures. Electrostatic potential energy used to describe the potential energy in systems with time variant electric fields. Its scale reads −38 to 38 with a red to blue scale representing the negative and positive electrostatic potential respectively. The graphs have varying abundance of negative electrostatic potential energy.

Figure 21 shows the plots for **solvation energy** i.e. the process of attraction and association of molecules of a solvent with molecules or ions of a solute or the amount of energy linked with dissolving a solute in a solvent. The results depict that some of the structures of AA10 have more solvation energy ranging from -110 to -56 and most of them had large negative values ranging from -164 to -110. Figure 22 shows the **potential energy** calculated for AA10 structures. The potential energy is measured in joules resulting from conservative Coulomb forces. Its scale

ranges from −64 to 64. All structures have a net negative potential energy. Figure 23 shows the **total charge of the molecule** (sum of formal charges). It is the charge assigned to an atom in a

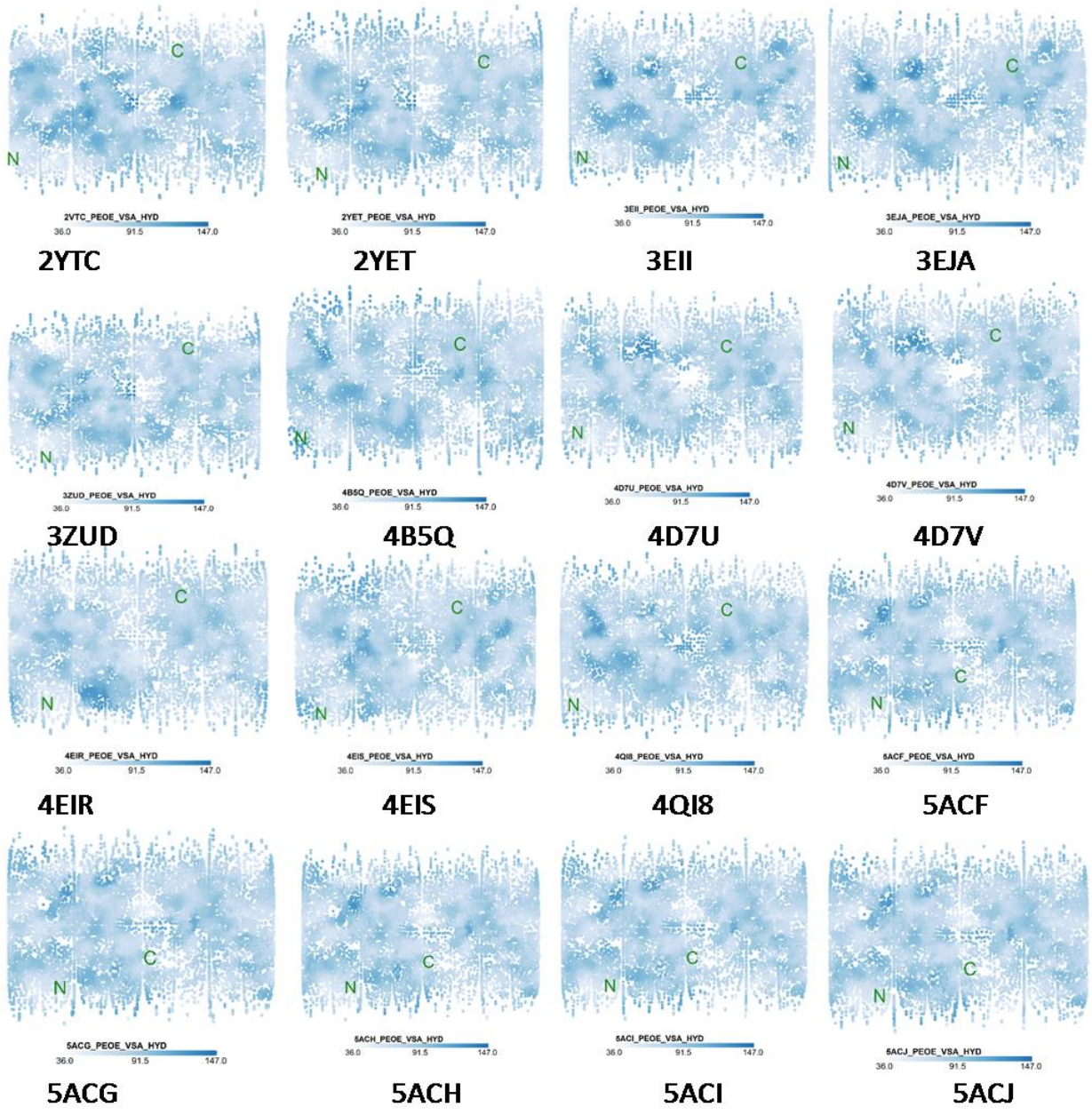


Figure 15: Plots showing fractional hydrophobic van der Waals surface area of 16 structures of AA9 calculated by Structuprint.

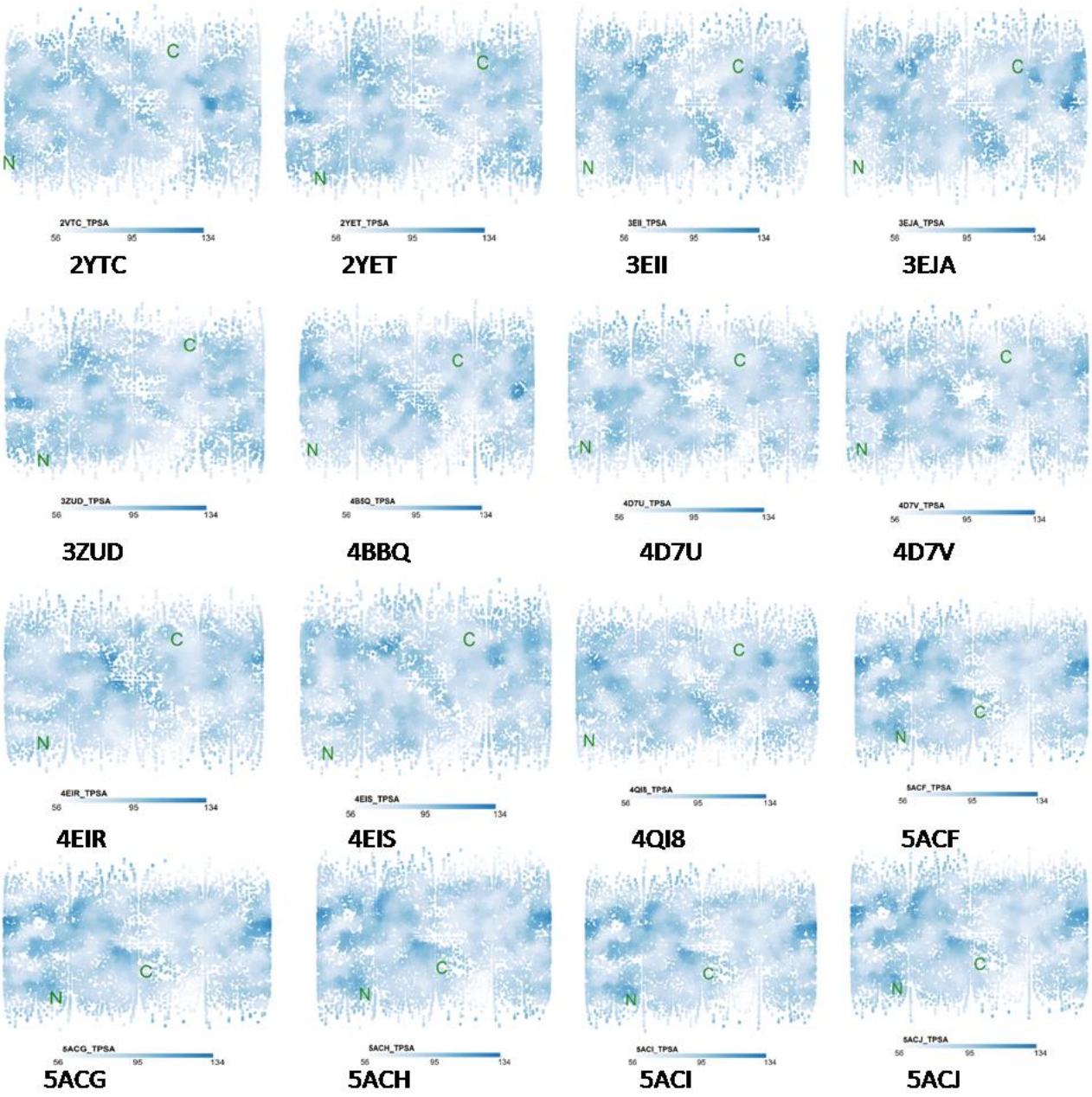


Figure 16: Plots showing total polar surface area of 16 structures of AA9 calculated by Structuprint.

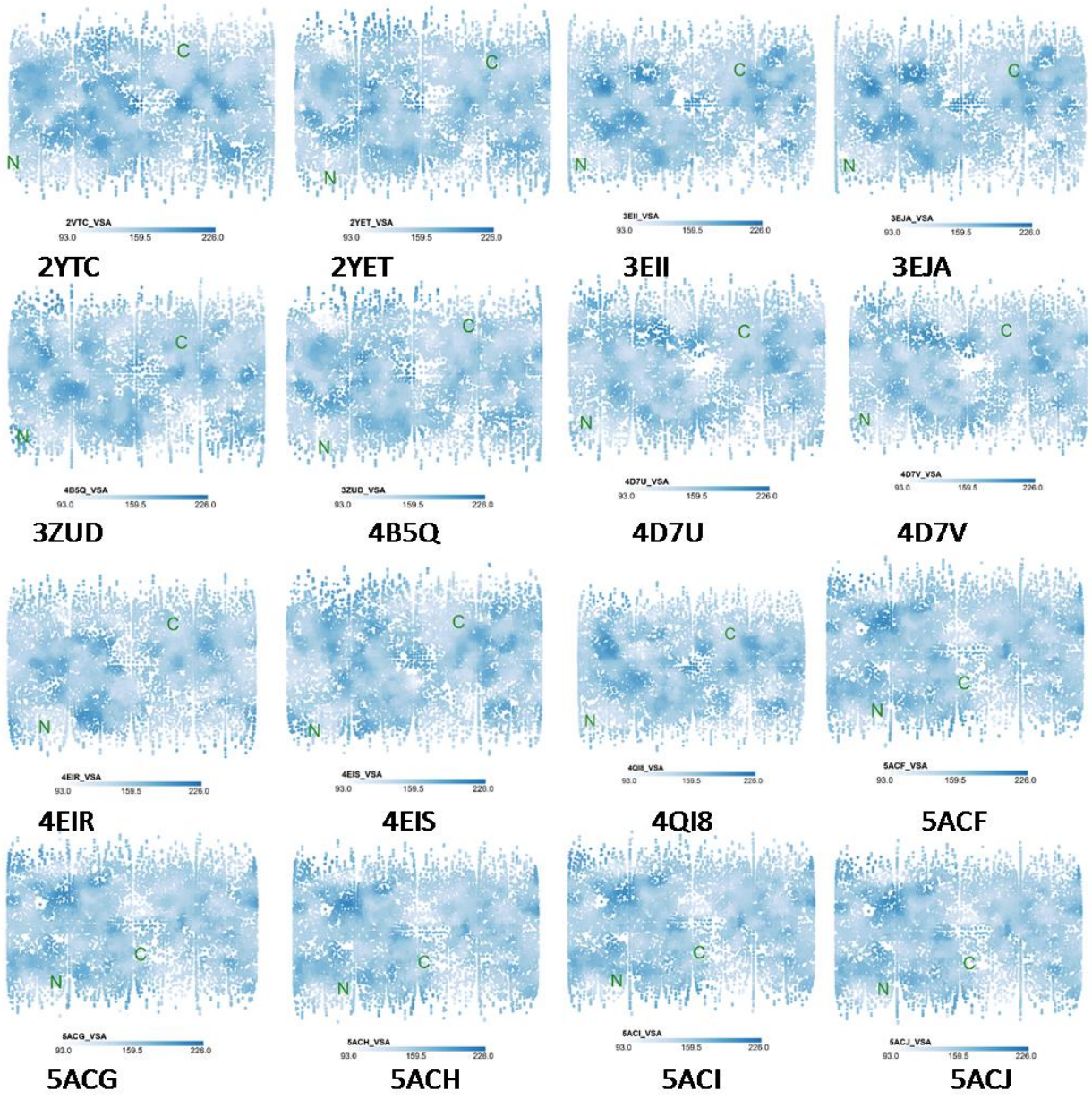


Figure 17: Plots showing van der waals surface area of 16 structures of AA9 calculated by Structuprint.

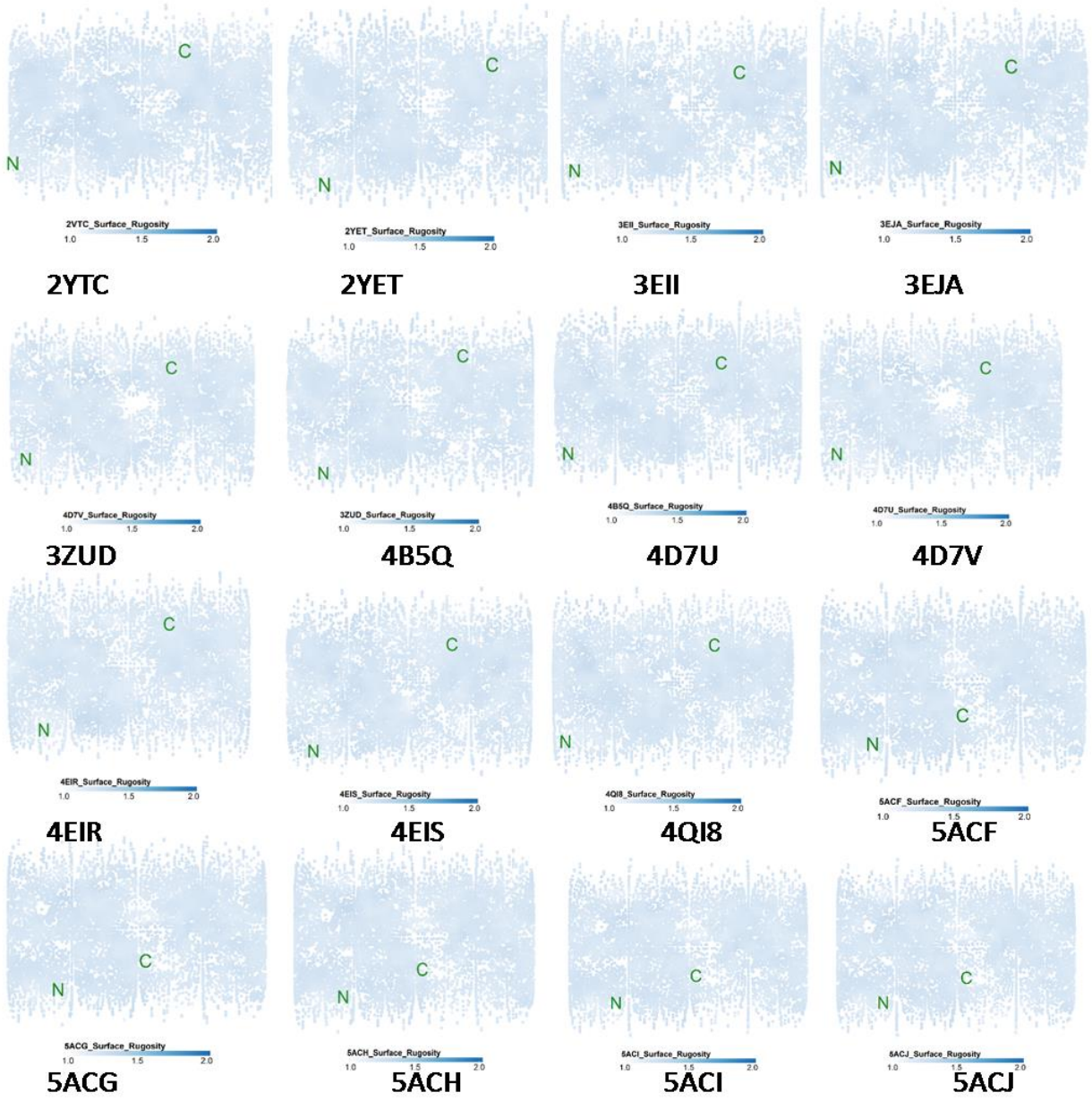


Figure 18: Plots showing surface rugosity of 16 structures of AA9 calculated by Structuprint.

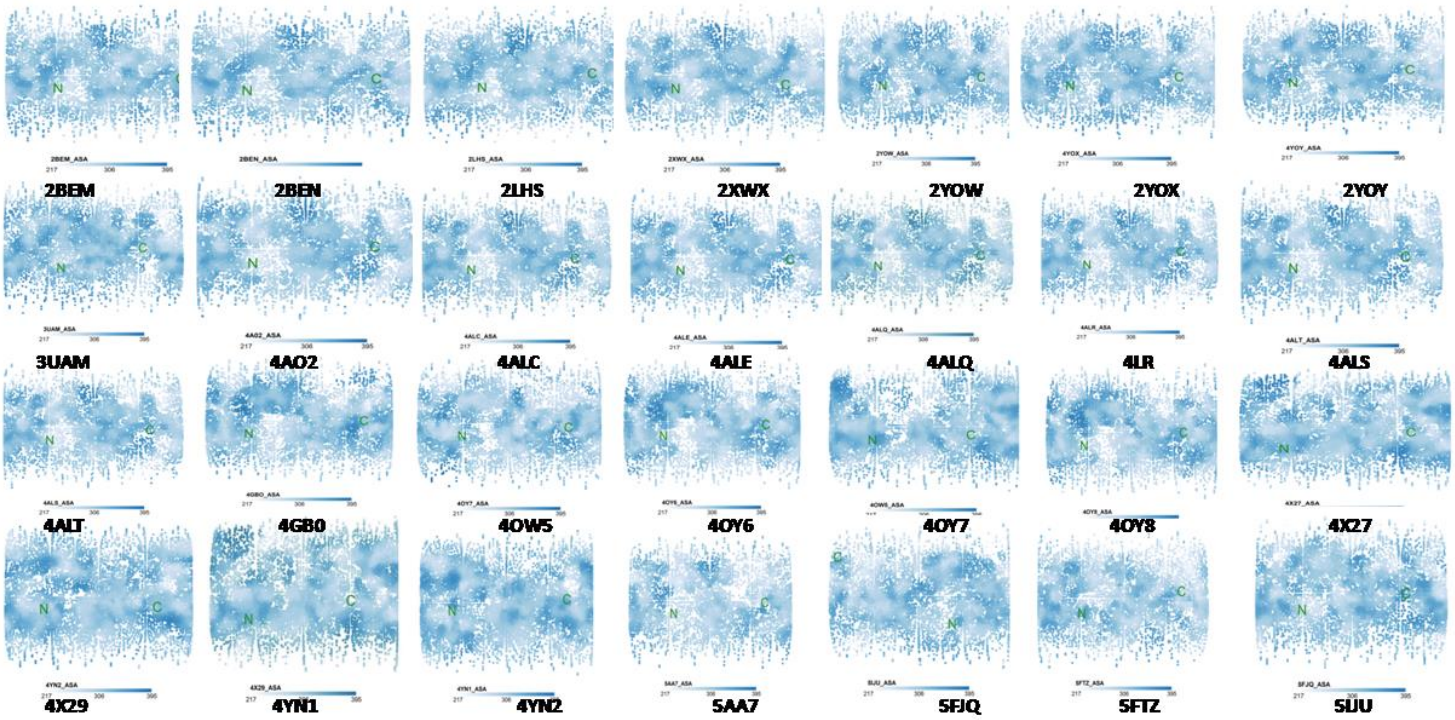


Figure 19: Plots showing Accessible Surface Area of 27 structures of AA10 calculated by Structuprint.

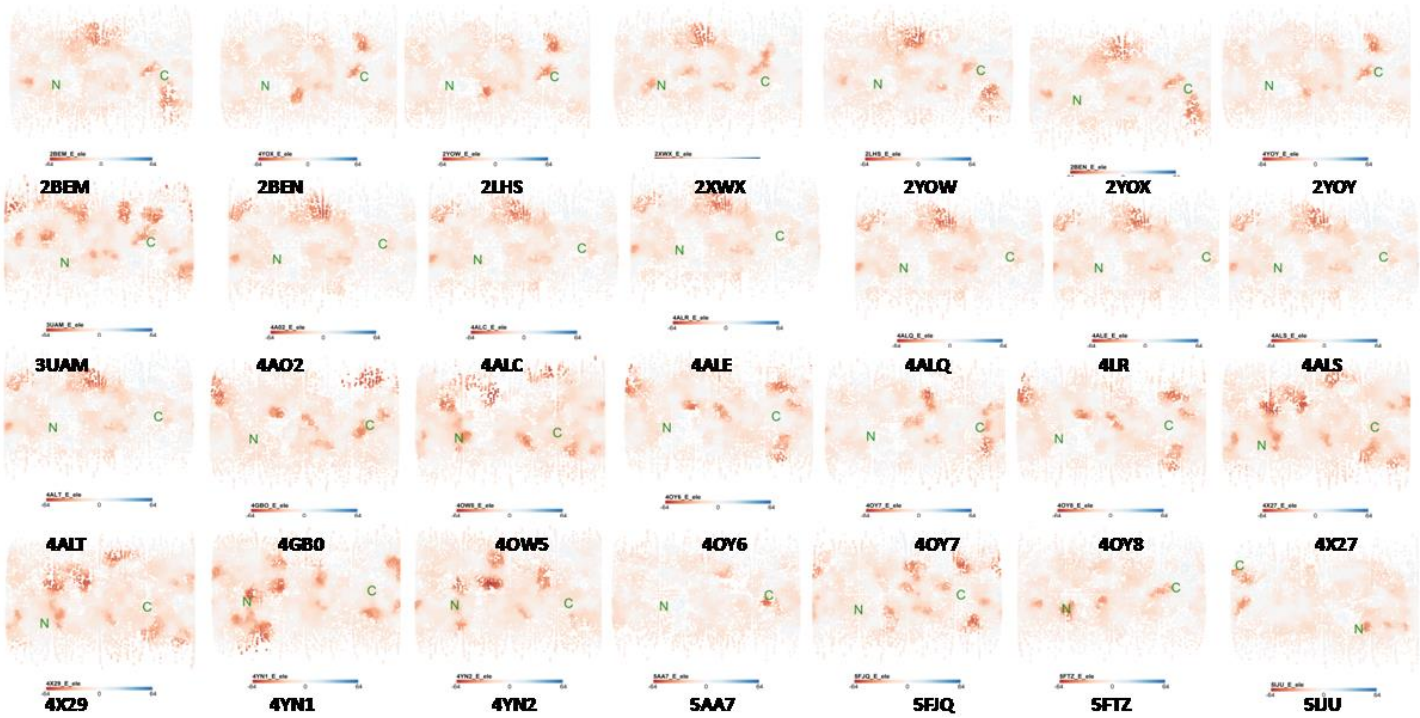


Figure 20: Plots showing electrostatic energy of 27 structures of AA10 calculated by Structuprint.

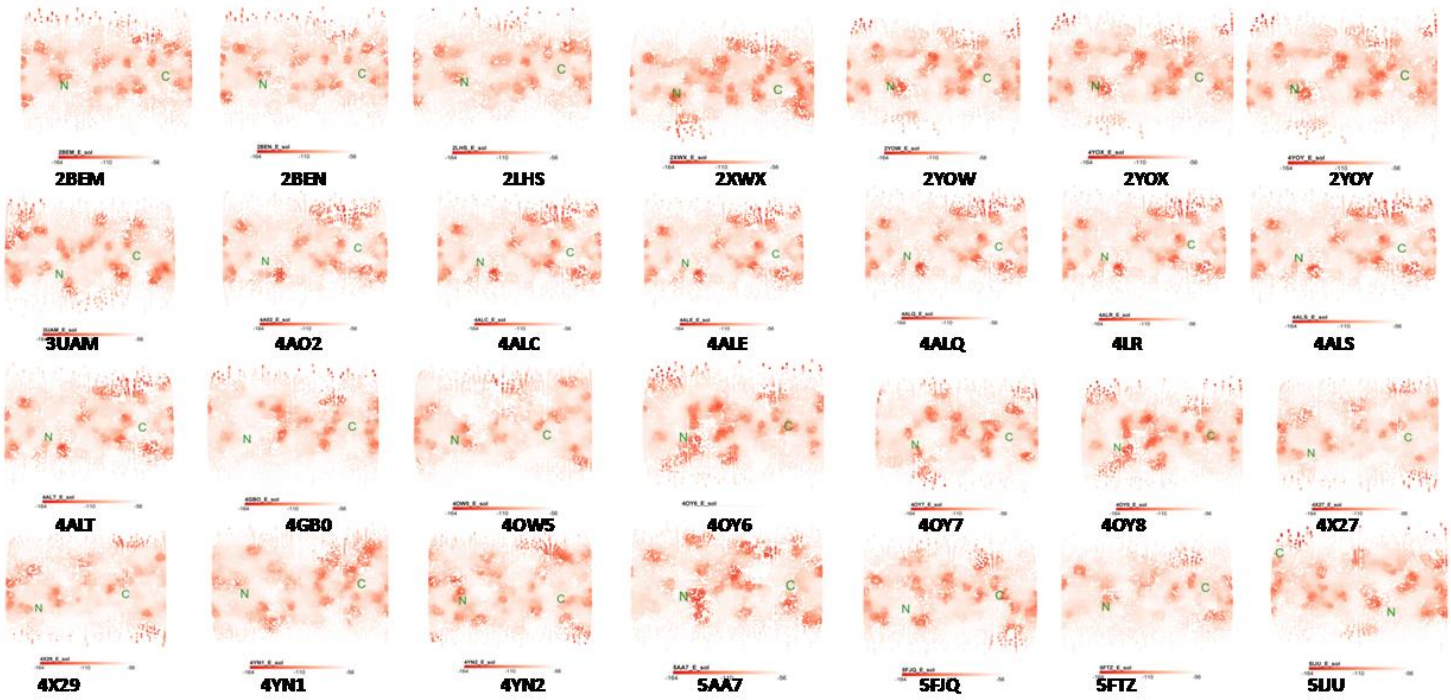


Figure 21: Plots showing solvation energy of 27 structures of AA10 calculated by Structuprint.

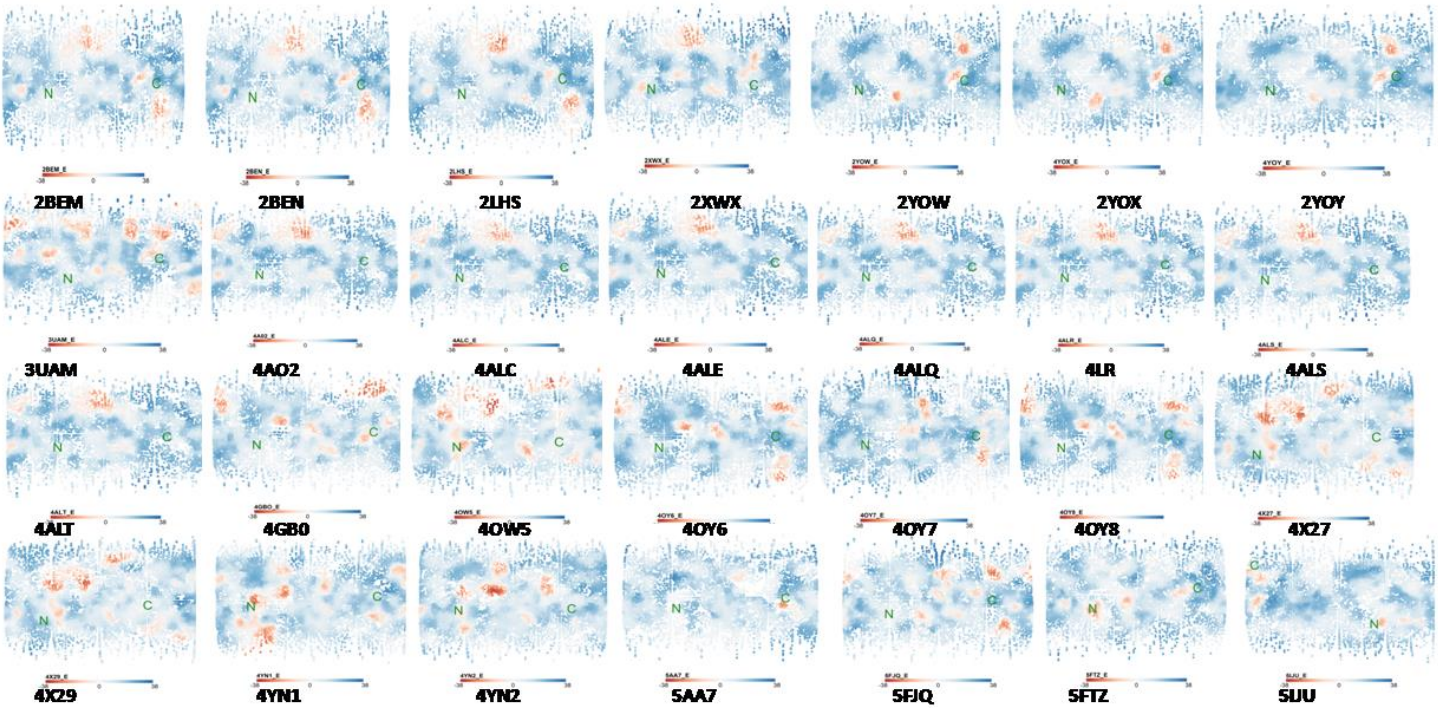


Figure 22: Plots showing potential energy of 27 structures of AA10 calculated by Structuprint.

molecule. The red dots in the graph illustrate the presence of the total negative charge and the blue dots illustrate the presence of the total positive charge. Its scale ranges from -1 to 1. Some of the protein structures have more total negative charge than that total positive charge and vice versa. Figure 24 the **globularity** i.e. Globularity or inverse condition number (smallest eigenvalue divided by the largest eigenvalue) of the covariance matrix of atomic coordinates. A value of 1 indicates a perfect sphere while a value of 0 indicates a two- or one-dimensional object. Most of the values were ranging between 0 and 0.5 on a scale of 0 to 0.5. Figure 25 shows the **Kier-Flex molecular flexibility index** which is derived from the Kier alpha modified shape descriptors. Most of the values were ranging between 1.0 and 3.5 only few of them were in the range of 3.5 and 6.0. Figure26 shows the **fractional van der waals** surface area this is the sum of the vi such that |qi| is less than or equal to 0.2 divided by the total surface area. The vi is calculated using a connection table approximation. Most of the values were ranging between 34.0 and 91.5 only few of them were in the range of 91.0 and 147.0 on the scale of 34.0 to 91.5. Figure 27 shows **total polar surface area** which is area of a molecule defined as the surface sum over all polar atoms, primarily oxygen and nitrogen. It has scale ranging from 56 to 134. Most of the proteins have values ranging from 56 to 95 but there are few families with more values ranging from 95 to 134. Figure 28 shows the **Van der Waals surface area** of AA9 structures**.** A polyhedral representation is used for each atom in calculating the surface area. The results depict that most of the structures of AA10 were lying between the range 93- 226. Figure 29 shows the **rugosity** i.e. the individual plots for AA11 structures are shown in Figure 30 (pdb id: 4mah) and Figure 31 (pdb id: 4mai), respectively. They show that these two structures have similar trend among the eleven properties analyzed using Structuprint.

The individual plots for AA11 structures are shown in Figure 30 (pdb id: 4mah) and Figure 31 (pdb id: 4mai), respectively. They show that these two structures have similar trend among the eleven properties analyzed using Structuprint.

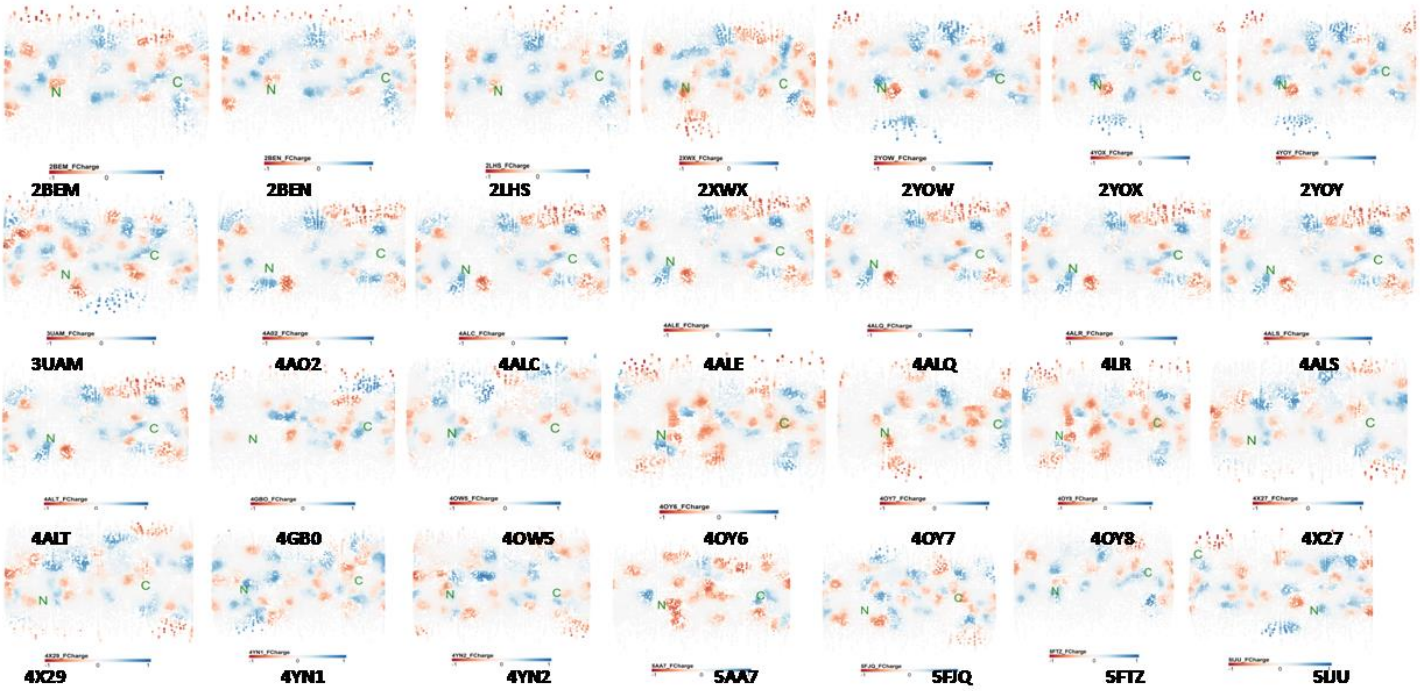


Figure 23: Plots showing formal charge of 27 structures of AA10 calculated by Structuprint.

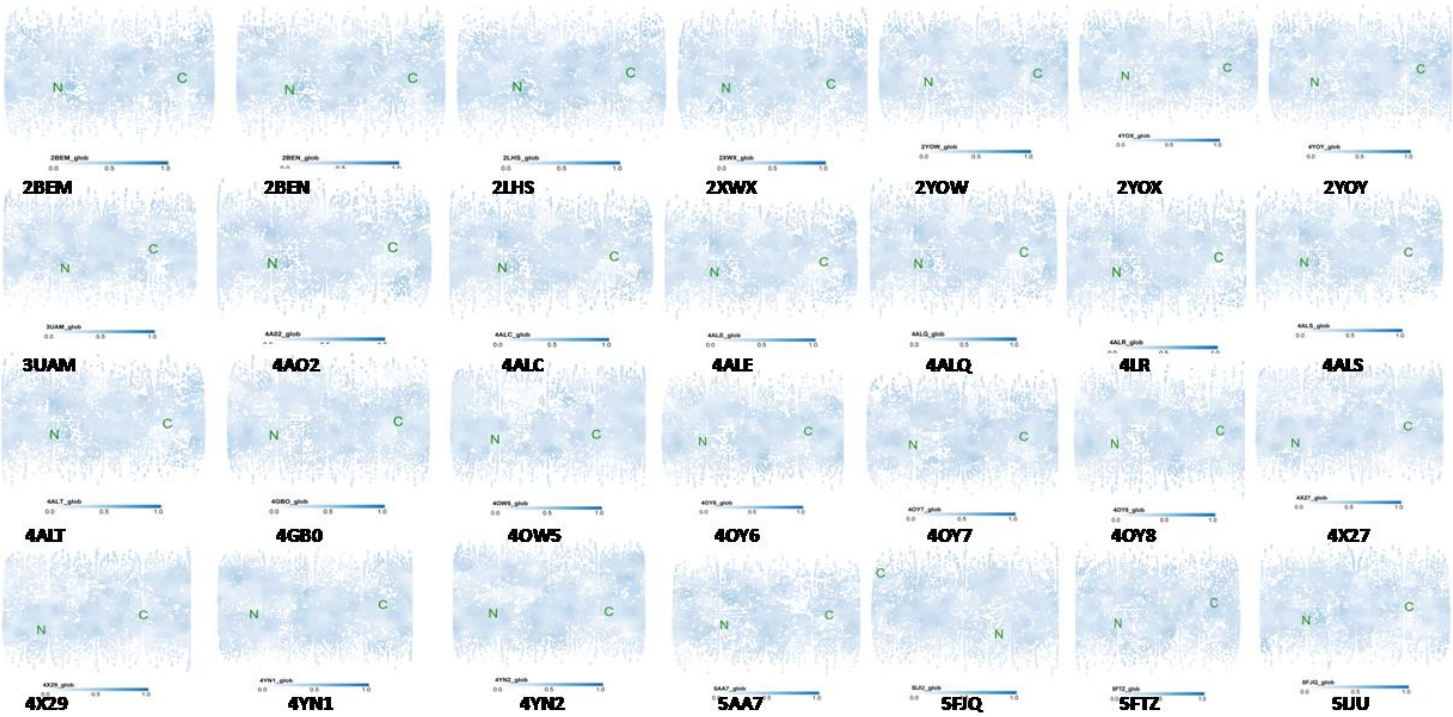


Figure 24: Plots showing globularity of 27 structures of AA10 calculated by Structuprint.

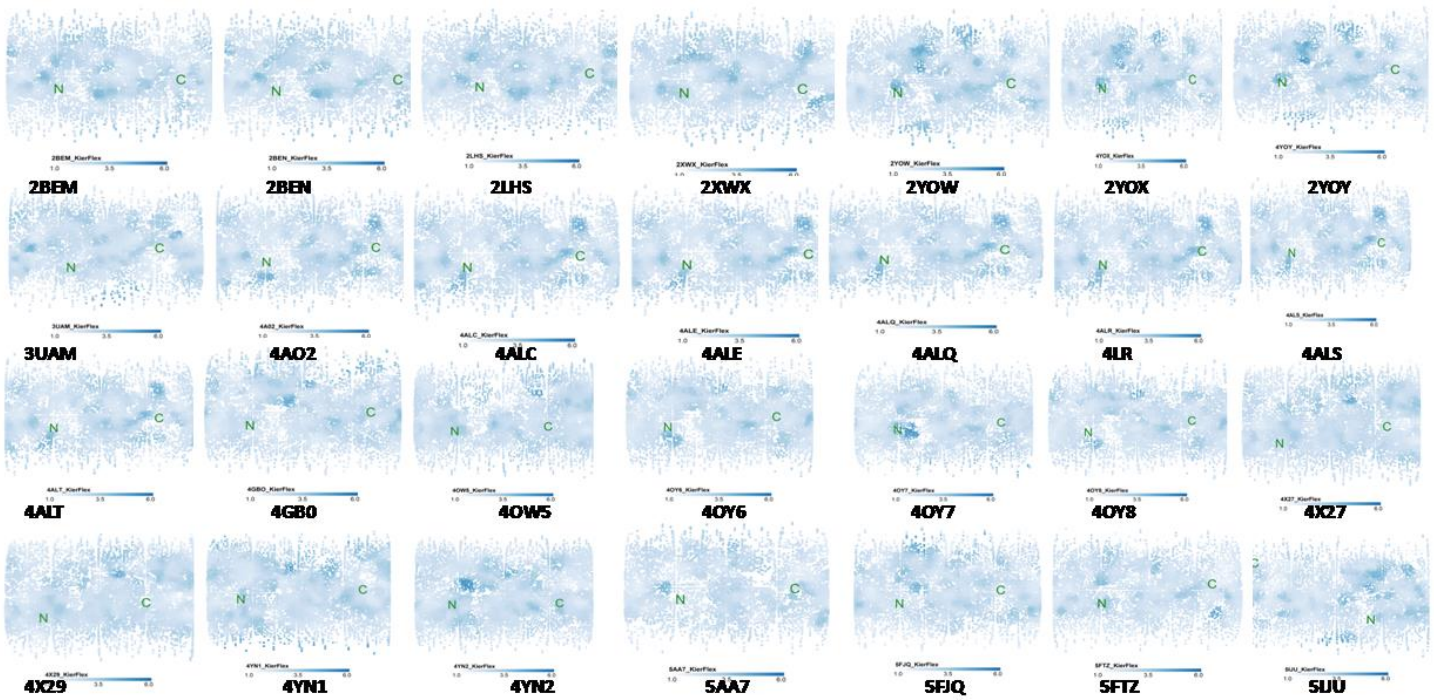


Figure 25: Plots showing Kier molecular flexibility of 27 structures of AA10 calculated by Structuprint.

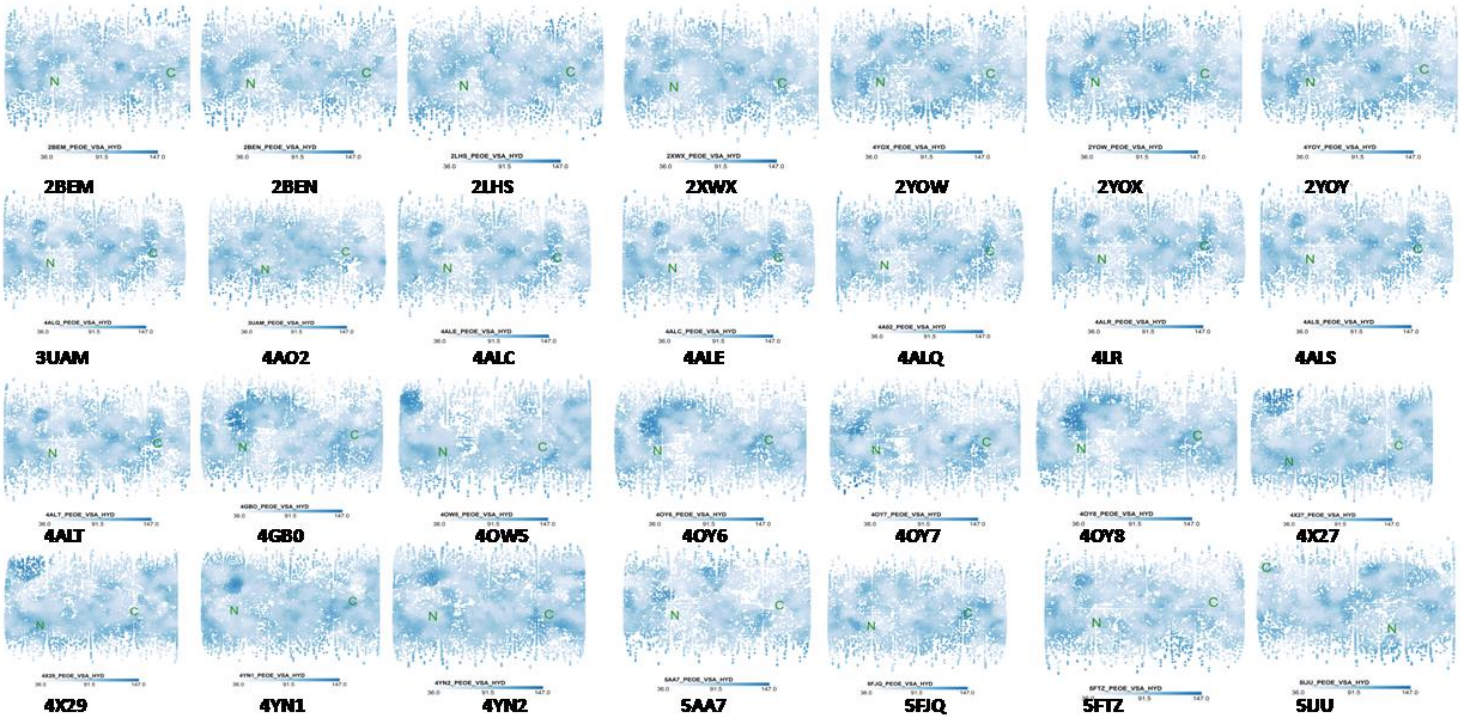


Figure 26: Plots showing Fractional hydrophobic van der Waals surface area of 27 structures of AA10 calculated by Structuprint.

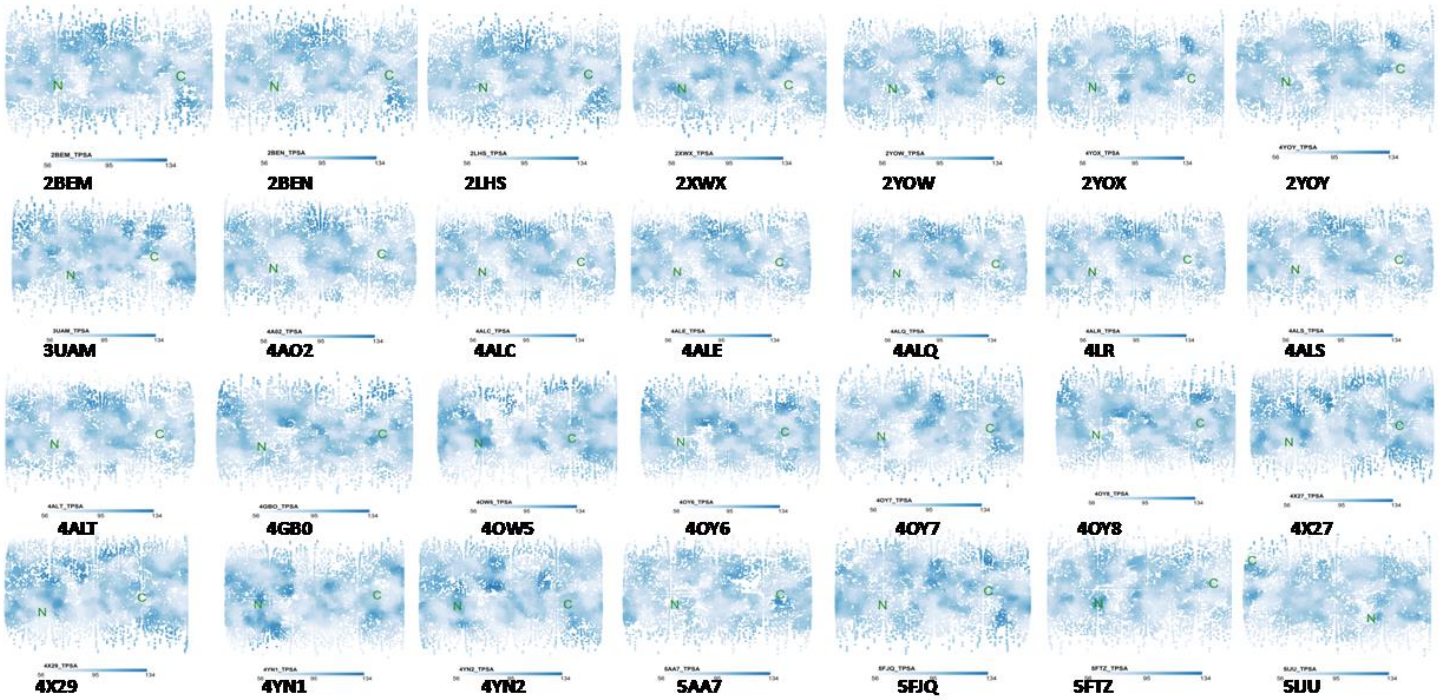


Figure 27: Plots showing total polar surface area of 27 structures of AA10 calculated by Structuprint.

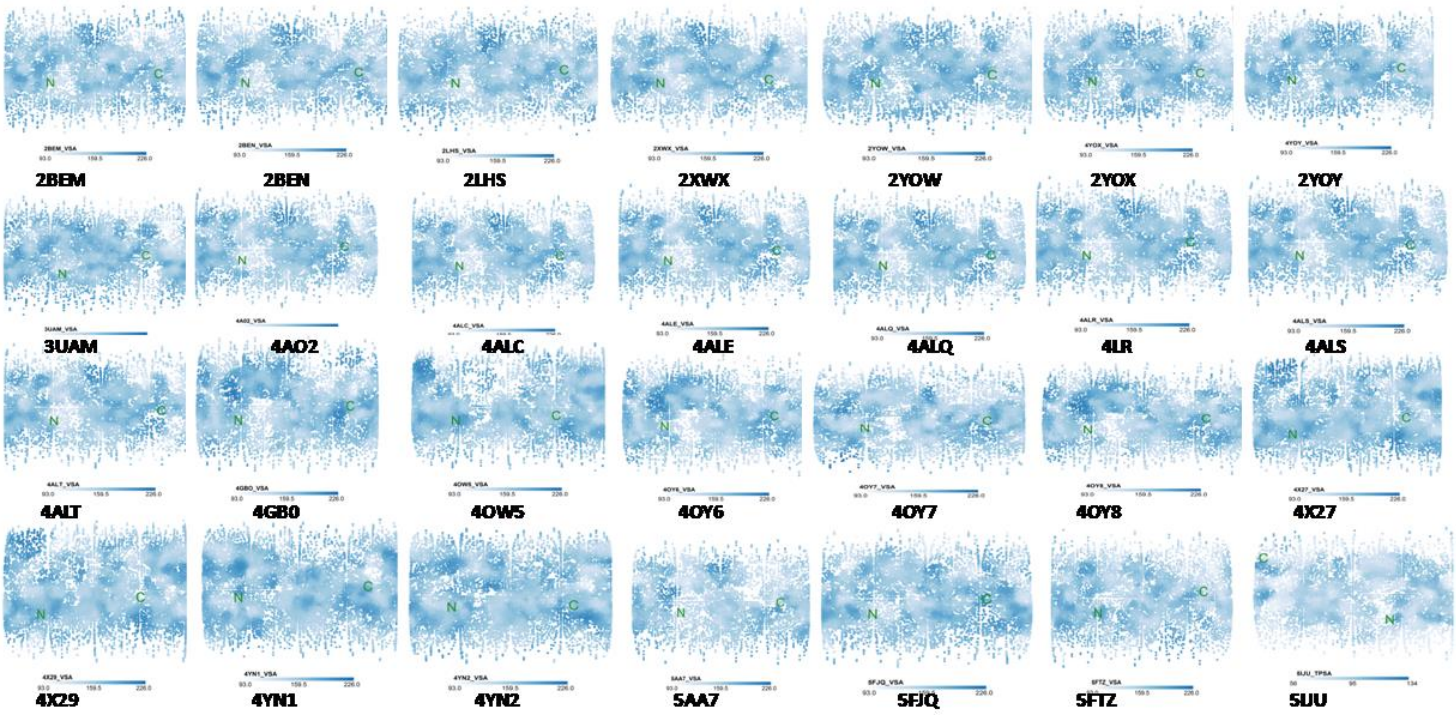


Figure 28: Plots showing van der waals surface area of 27 structures of AA10 calculated by Structuprint.

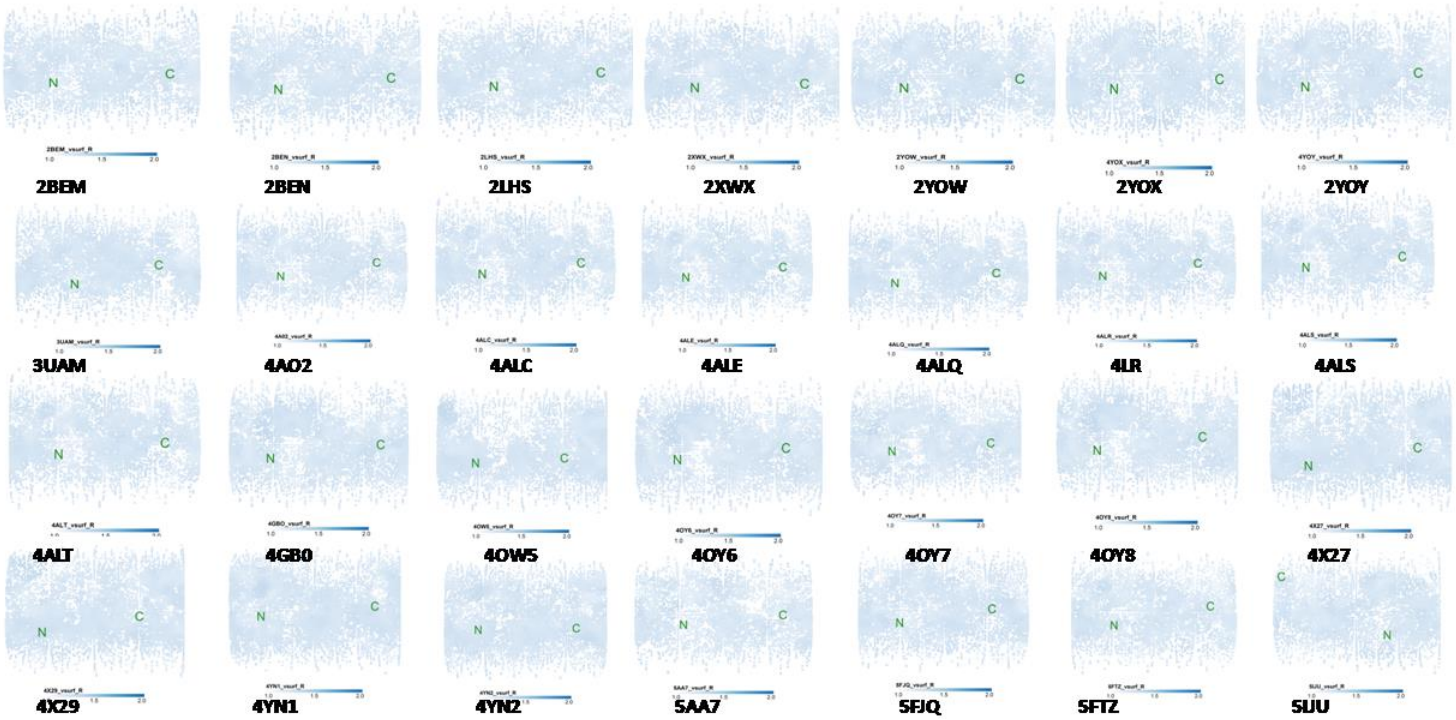


Figure 29: Plots showing surface rugosity of 27 structures of AA10 calculated by Structuprint.

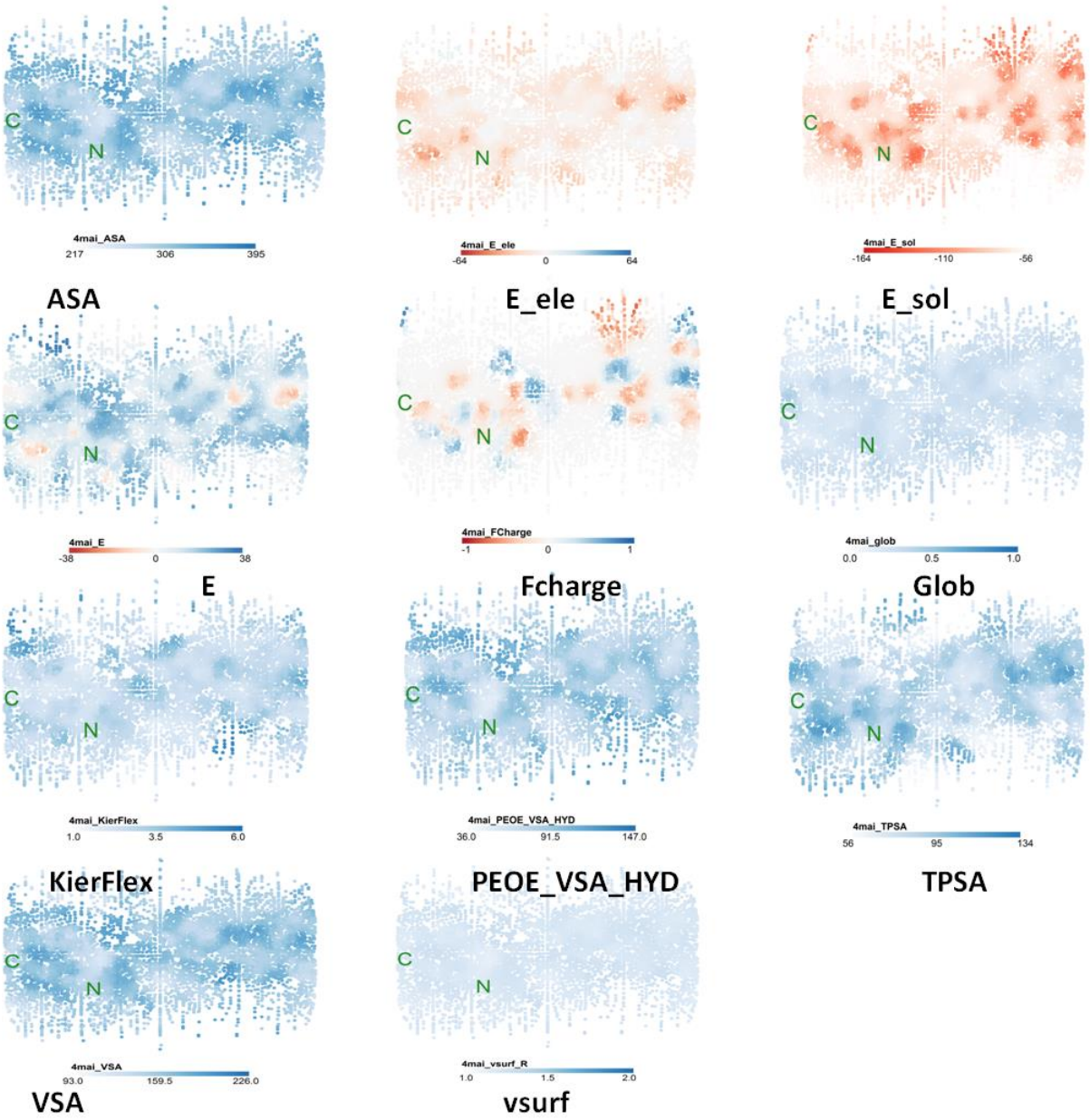


Figure 31: Plots of AA11 structure (pdb id: 4mai) for eleven properties calculated by Structuprint

Prediction of residues involved in Allostery

In Statistical coupling analysis, we observed the conserved residues. Figure 32 is showing the similarity matrix of different sequences of AA10 family and a histogram for the same. We could have derived the covariance matrix too but we are majorly focusing on similarity here. The histogram shows significantly narrow distribution with mean pairwise identity between sequence of about 22% and range of 10-40% suggesting most sequences are equally dissimilar. Figure 33 shows the positional conservation and as we had larger number of sequences as compared to the length of the sequence, the results are significant showing positional conservation at places other than histidines. Then we extracted eigenvectors and figure 34a) shows top three eigenvectors plotted in 3D space. Figure 34 b) shows them separately in 2 dimensional space and figure 34 c) shows the sequence correlation plotted against potential correlation of AA10 sequences in 3D space. In figure 34 d) different sectors, as already explained earlier, are defined according to two different cut-offs. Then after defining the sectors we mapped them with the 2BEM structure in PyMOL and Figure 35 is showing a) the side view of the structure and b) the top view of the same. It confirms that there are residues (glutamic acid, alanines, tryptophans etc.) other than histidines which might be contributing towards the allostery of the structure.

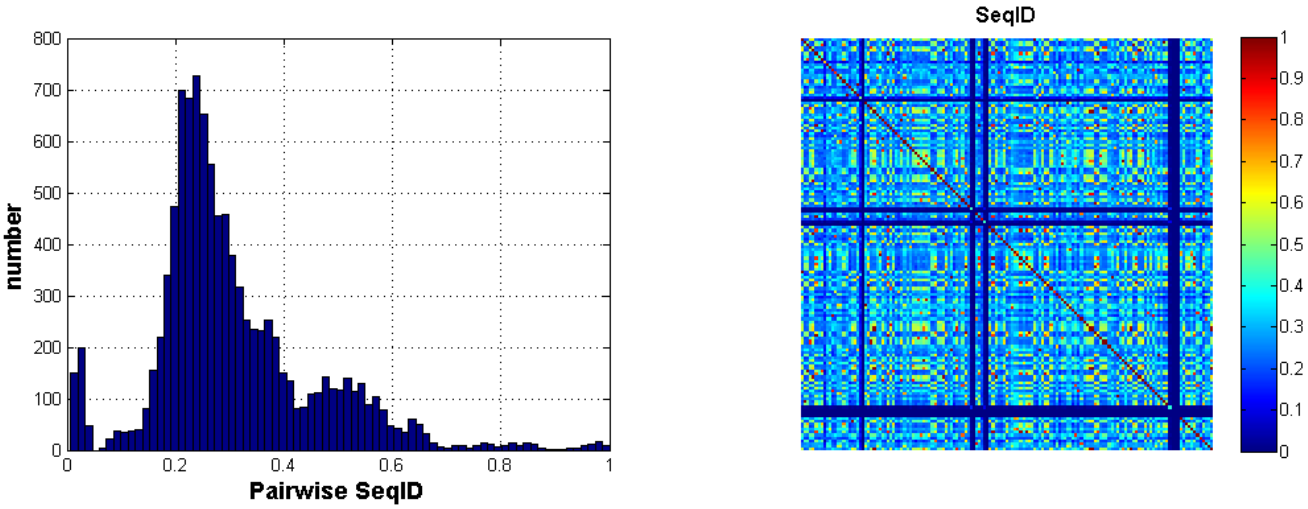


Figure 32: Sequence Correlations of AA10 A) histogram of the similarities between pairs of sequences, B) The similarly matrix

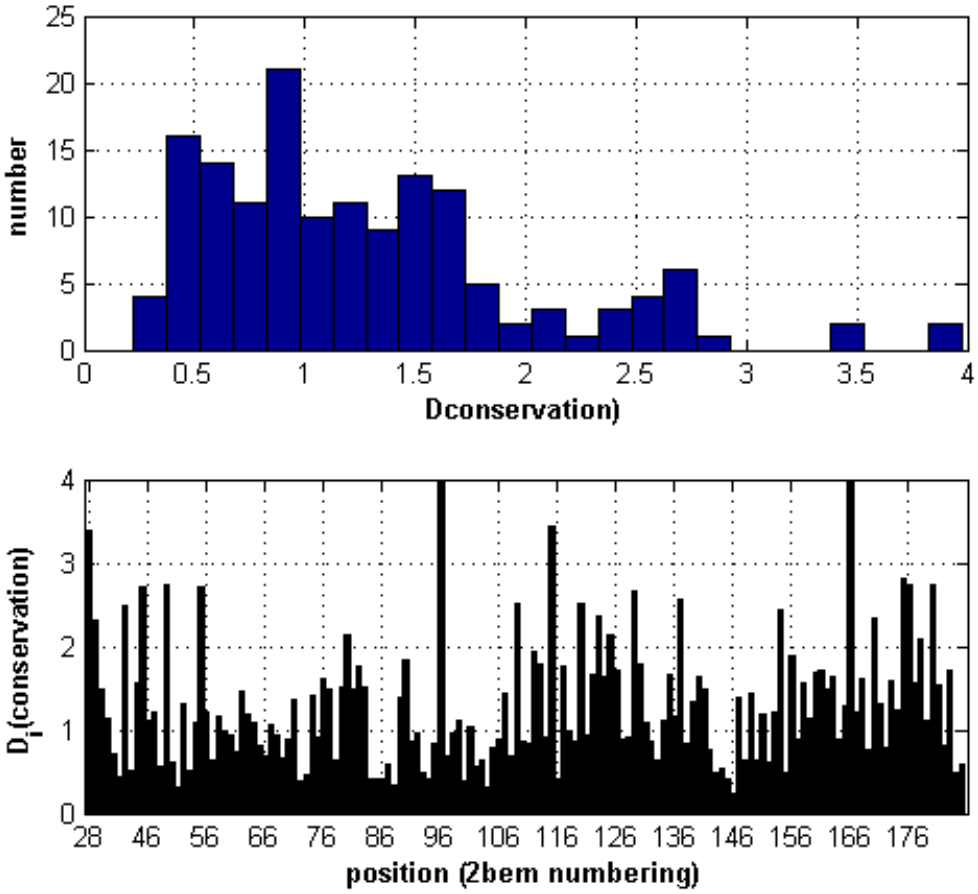


Figure 33: Positional Conservation AA10

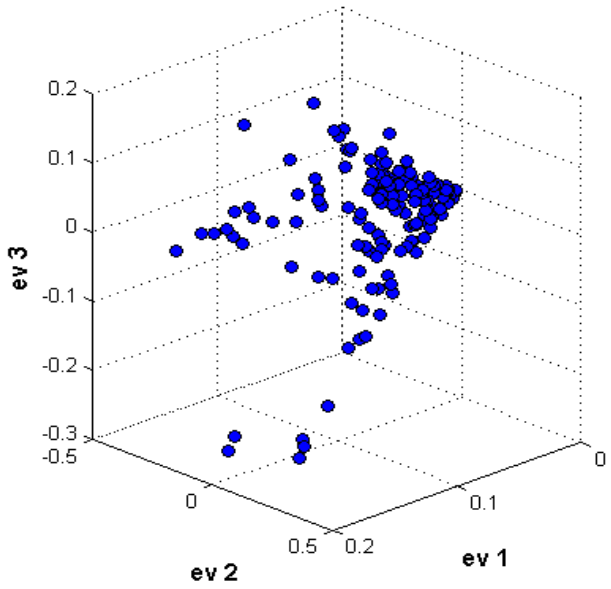


Figure 34(A): 3-D plots of the top three eigenvectors

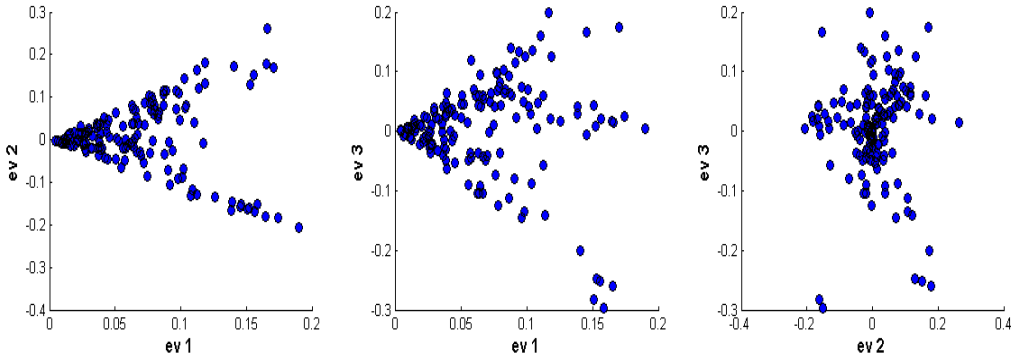


Figure 34(B): 2-D Plots of the top three eigenvectors

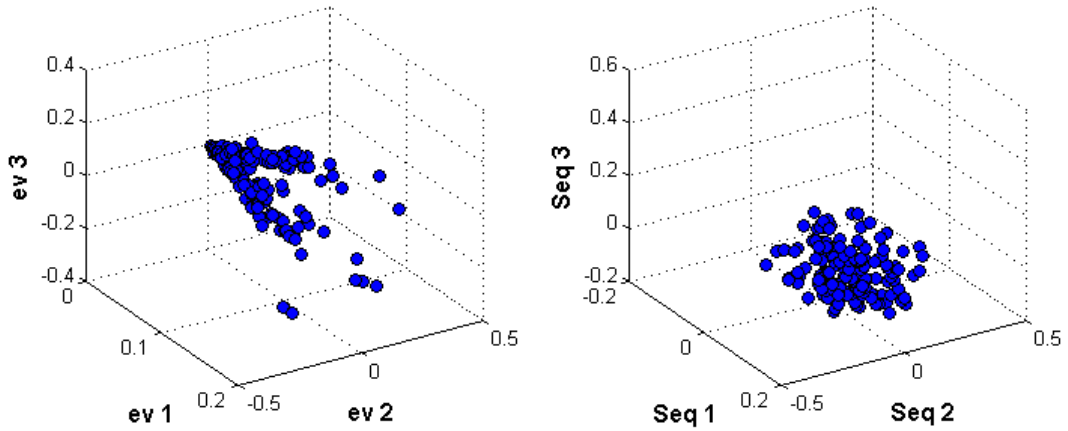


Figure 34(C): Mapping Sequence correlation by positional correlation of AA10

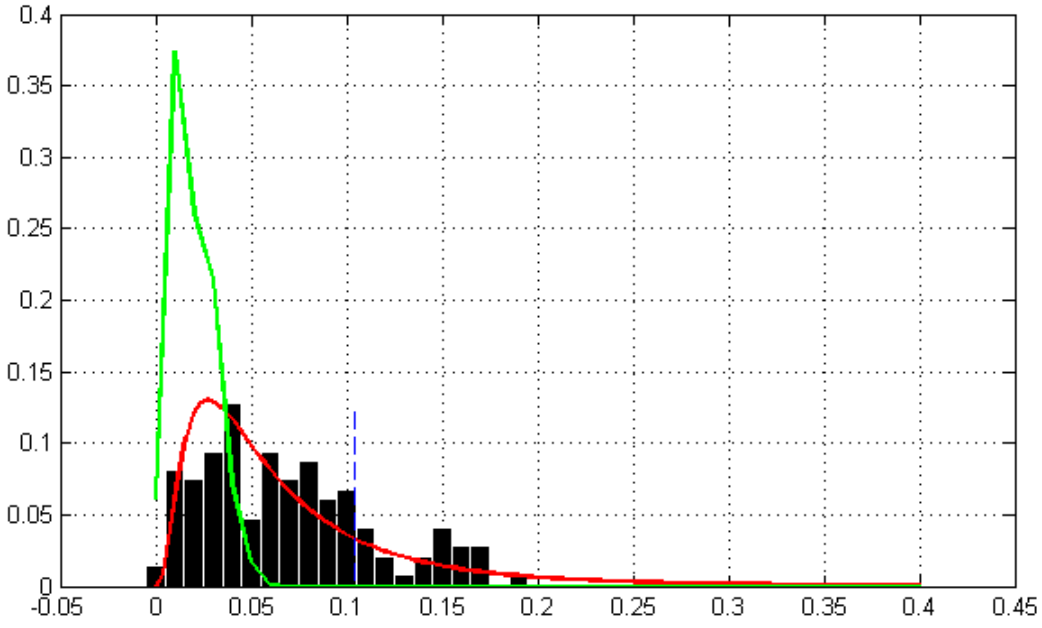


Figure 34(D): Defining Sectors according to the top Eigen mode vectors of AA10

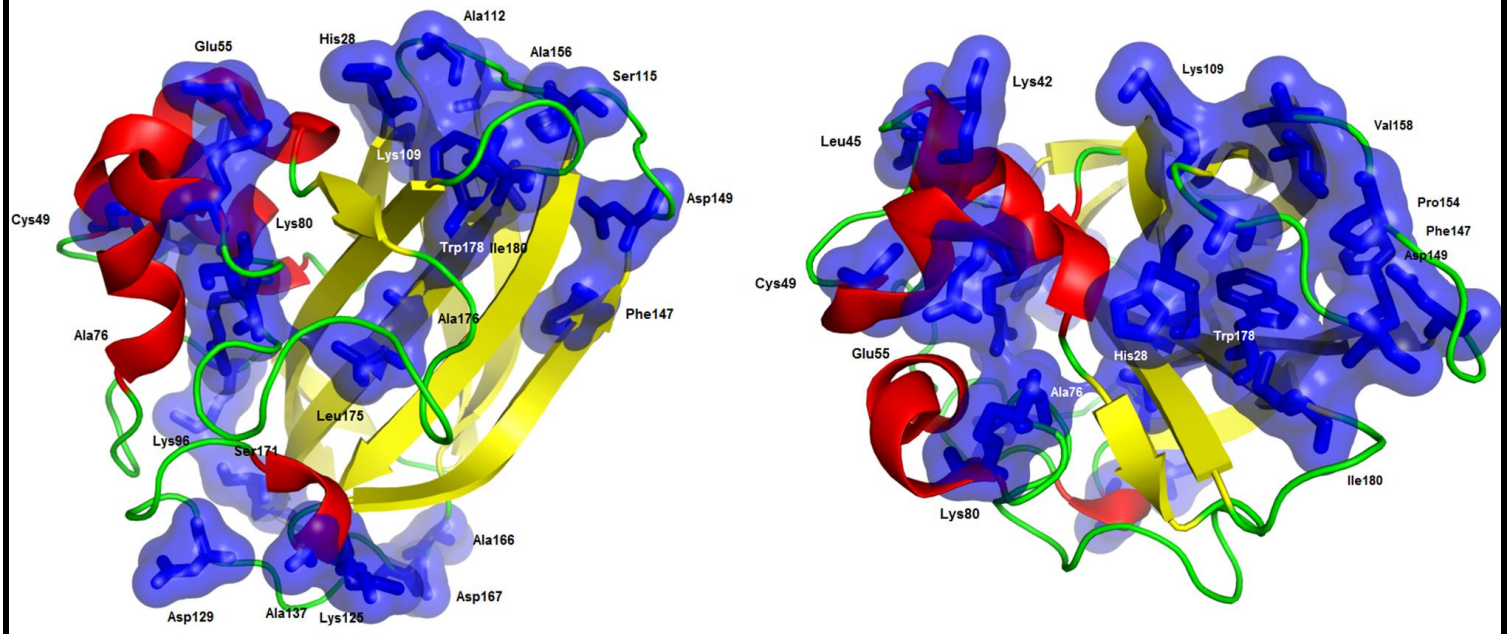


Figure 35: (A) Side view by Statistical Coupling Analysis (SCA) and (B) Top view of residues that are predicted.

CONCLUSIONS

- The Elastic network models indicate dynamics at the substrate binding side.
- The cross-correlation maps show complex interplay of other residues towards substrate binding
- There are differences within the families in terms of physiochemical properties, such as Formal Charge.
- Statistical Coupling Analysis show that there are residues on the nonsubstrate binding side that possibly impart allostery

REFERENCES

1. S. Horn, G. Vaaje-Kolstad, B. Westereng and V. Eijsink, "Novel enzymes for the degradation of cellulose", *Biotechnology for Biofuels*, vol. 5, no. 1, p. 45, 2012.
2. McNamara, J. Morgan and J. Zimmer, "A Molecular Description of Cellulose Biosynthesis", Annual Review of Biochemistry, vol. 84, no. 1, pp. 895-921, 2015
3. Rogers and Kara, *The Science of Nutrition*, 1st ed. New York: Britannica Educational Pub. in association with Rosen Educational Services, 2013.
4. 2016. [Online]. Available: http://Chem.gla.ac.uk. [Accessed: 22- Dec- 2016].
5. A. Bakan, L. Meireles and I. Bahar, "ProDy: Protein Dynamics Inferred from Theory and Experiments", *Bioinformatics*, vol. 27, no. 11, pp. 1575-1577, 2011.A. Bakan, A. Dutta, W. Mao, Y. Liu, C. Chennubhotla, T. Lezon and I. Bahar, "Evol and ProDy for bridging protein sequence evolution and structural dynamics", *Bioinformatics*, vol. 30, no. 18, pp. 2681-2683, 2014.
6. D. Kontopoulos, D. Vlachakis, G. Tsiliki and S. Kossida, "Structuprint: a scalable and extensible tool for two-dimensional representation of protein surfaces", *BMC Structural Biology*, vol. 16, no. 1, 2016.
7. M. Wu, G. Beckham, A. Larsson, T. Ishida, S. Kim, C. Payne, M. Himmel, M. Crowley, S. Horn, B. Westereng, K. Igarashi, M. Samejima, J. Stahlberg, V. Eijsink and M. Sandgren, "Crystal Structure and Computational Characterization of the Lytic Polysaccharide Monooxygenase GH61D from the Basidiomycota Fungus Phanerochaete chrysosporium", *Journal of Biological Chemistry*, vol. 288, no. 18, pp. 12828-12839, 2013.
8. K. Frandsen and L. Lo Leggio, "Lytic polysaccharide monooxygenases: a crystallographer's view on a new class of biomass-degrading enzymes", IUCrJ, vol. 3, no. 6, pp. 448-467, 2016.
9. 9. M. Gudmundsson, S. Kim, M. Wu, T. Ishida, M. Momeni, G. Vaaje-Kolstad, D. Lundberg, A. Royant, J. Stahlberg, V. Eijsink, G. Beckham and M. Sandgren, "Structural and Electronic Snapshots during the Transition from a Cu(II) to Cu(I) Metal Center of a Lytic Polysaccharide Monooxygenase by X-ray Photoreduction", Journal of Biological Chemistry, vol. 289, no. 27, pp. 18782-18792, 2014.

10. 10. G. Vaaje-Kolstad, D. Houston, A. Riemen, V. Eijsink and D. van Aalten, "Crystal Structure and Binding Properties of the Serratiamarcescens Chitin-binding Protein CBP21", Journal of Biological Chemistry, vol. 280, no. 12, pp. 11313-11319, 2004.
11. A. Book, R. Yennamalli, T. Takasuka, C. Currie, G. Phillips and B. Fox, "Evolution of substrate specificity in bacterial AA10 lytic polysaccharide monooxygenases", Biotechnology for Biofuels, vol. 7, no. 1, p. 109, 2014.].
12. W. Beeson, V. Vu, E. Span, C. Phillips and M. Marletta, "Cellulose Degradation by Polysaccharide Monooxygenases", *Annual Review of Biochemistry*, vol. 84, no. 1, pp. 923-946, 2015.
13. K. Frandsen and L. Lo Leggio, "Lytic polysaccharide monooxygenases: a crystallographer's view on a new class of biomass-degrading enzymes", *IUCrJ*, vol. 3, no. 6, pp. 448-467, 2016.].
14. Agger, T. Isaksen, A. Varnai, S. Vidal-Melgosa, W. Willats, R. Ludwig, S. Horn, V. Eijsink and B. Westereng, "Discovery of LPMO activity on hemicelluloses shows the importance of oxidative processes in plant cell wall degradation", *Proceedings of the National Academy of Sciences*, vol. 111, no. 17, pp. 6287-6292, 2014.
15. P. Harris and D. Welner, et al. "Stimulation of Lignocellulosic Biomass Hydrolysis by Proteins of Glycoside Hydrolase Family 61: Structure and Function of a Large, Enigmatic Family", *Biochemistry*, vol. 49, no. 15, pp. 3305-3316, 2010.
16. V. Lombard, H. Golaconda Ramulu, E. Drula, P. Coutinho and B. Henrissat, "The carbohydrate-active enzymes database (CAZy) in 2013", Nucleic Acids Research, vol. 42, no. 1, pp. D490-D495, 2013.
17. S. Karkehabadi, H. Hansson, S. Kim, K. Piens, C. Mitchinson and M. Sandgren, "The First Structure of a Glycoside Hydrolase Family 61 Member, Cel61B from Hypocrea jecorina, at 1.6 Å Resolution", *Journal of Molecular Biology*, vol. 383, no. 1, pp. 144-154, 2008.
18. R. Quinlan, M. Sweeney, L. Lo Leggio, H. Otten, J. Poulsen, K. Johansen, K. Krogh, C. Jorgensen, M. Tovborg, A. Anthonsen, T. Tryfona, C. Walter, P. Dupree, F. Xu, G. Davies and P. Walton, "Insights into the oxidative degradation of cellulose by a copper metalloenzyme that exploits biomass components", *Proceedings of the National Academy of Sciences*, vol. 108, no. 37, pp. 15079-15084, 2011.
19. M. Wu, G. Beckham, A. Larsson, T. Ishida, S. Kim, C. Payne, M. Himmel, M. Crowley, S. Horn, B. Westereng, K. Igarashi, M. Samejima, J. Stahlberg, V. Eijsink and M. Sandgren,

"Crystal Structure and Computational Characterization of the Lytic Polysaccharide Monooxygenase GH61D from the Basidiomycota Fungus Phanerochaete chrysosporium", *Journal of Biological Chemistry*, vol. 288, no. 18, pp. 12828-12839, 2013.

20. A. Borisova, T. Isaksen, M. Dimarogona, A. Kognole, G. Mathiesen, A. Várnai, Å. Røhr, C. Payne, M. Sørlie, M. Sandgren and V. Eijsink, "Structural and Functional Characterization of a Lytic Polysaccharide Monooxygenase with Broad Substrate Specificity", *Journal of Biological Chemistry*, vol. 290, no. 38, pp. 22955-22969, 2015.
21. X. Li, W. Beeson, C. Phillips, M. Marletta and J. Cate, "Structural Basis for Substrate Targeting and Catalysis by Fungal Polysaccharide Monooxygenases", *Structure*, vol. 20, no. 6, pp. 1051-1061, 2012.
22. K. Frandsen, T. Simmons, P. Dupree, J. Poulsen, G. Hemsworth, L. Ciano, E. Johnston, M. Tovborg, K. Johansen, P. von Freiesleben, L. Marmuse, S. Fort, S. Cottaz, H. Driguez, B. Henrissat, N. Lenfant, F. Tuna, A. Baldansuren, G. Davies, L. Lo Leggio and P. Walton, "The molecular basis of polysaccharide cleavage by lytic polysaccharide monooxygenases", *Nature Chemical Biology*, vol. 12, no. 4, pp. 298-303, 2016.
23. E. Chiu, M. Hijnen, R. Bunker, M. Boudes, C. Rajendran, K. Aizel, V. Oliéric, C. Schulze-Briese, W. Mitsuhashi, V. Young, V. Ward, M. Bergoin, P. Metcalf and F. Coulibaly, "Structural basis for the enhancement of virulence by viral spindles and their in vivo crystallization", *Proceedings of the National Academy of Sciences*, vol. 112, no. 13, pp. 3973-3978, 2015.
24. E. Wong, G. Vaaje-Kolstad, A. Ghosh, R. Hurtado-Guerrero, P. Konarev, A. Ibrahim, D. Svergun, V. Eijsink, N. Chatterjee and D. van Aalten, "The Vibrio cholerae Colonization Factor GbpA Possesses a Modular Structure that Governs Binding to Different Host Surfaces", *PLoS Pathogens*, vol. 8, no. 1, p. e1002373, 2012.
25. F. Moser, D. Irwin, S. Chen and D. Wilson, "Regulation and characterization ofThermobifida fusca carbohydrate-binding module proteins E7 and E8", *Biotechnology and Bioengineering*, vol. 100, no. 6, pp. 1066-1077, 2008.
26. A. Chaplin, M. Wilson, M. Hough, D. Svistunenko, G. Hemsworth, P. Walton, E. Vijgenboom and J. Worrall, "Heterogeneity in the Histidine-brace Copper Coordination Sphere in Auxiliary Activity Family 10 (AA10) Lytic Polysaccharide Monooxygenases", *Journal of Biological Chemistry*, vol. 291, no. 24, pp. 12838-12850, 2016.

27. Z. Forsberg, A. Mackenzie, M. Sorlie, A. Rohr, R. Helland, A. Arvai, G. Vaaje-Kolstad and V. Eijsink, "Structural and functional characterization of a conserved pair of bacterial cellulose-oxidizing lytic polysaccharide monooxygenases", *Proceedings of the National Academy of Sciences*, vol. 111, no. 23, pp. 8446-8451, 2014.
28. F. Aachmann, M. Sorlie, G. Skjak-Braek, V. Eijsink and G. Vaaje-Kolstad, "NMR structure of a lytic polysaccharide monooxygenase provides insight into copper binding, protein dynamics, and substrate interactions", *Proceedings of the National Academy of Sciences*, vol. 109, no. 46, pp. 18779-18784, 2012.
29. G. Vaaje-Kolstad, "Crystal Structure and Binding Properties of the Serratia marcescens Chitin-binding Protein CBP21", *Journal of Biological Chemistry*, vol. 280, no. 12, pp. 11313-11319, 2005.
30. S. Mekasha, Z. Forsberg, B. Dalhus, J. Bacik, S. Choudhary, C. Schmidt-Dannert, G. Vaaje-Kolstad and V. Eijsink, "Structural and functional characterization of a small chitin-active lytic polysaccharide monooxygenase domain of a multi-modular chitinase fromJonesia denitrificans", *FEBS Letters*, vol. 590, no. 1, pp. 34-42, 2015.
31. M. Gudmundsson, S. Kim, M. Wu, T. Ishida, M. Momeni, G. Vaaje-Kolstad, D. Lundberg, A. Royant, J. Stahlberg, V. Eijsink, G. Beckham and M. Sandgren, "Structural and Electronic Snapshots during the Transition from a Cu(II) to Cu(I) Metal Center of a Lytic Polysaccharide Monooxygenase by X-ray Photoreduction", *Journal of Biological Chemistry*, vol. 289, no. 27, pp. 18782-18792, 2014.
32. G. Vaaje-Kolstad, L. Bøhle, S. Gåseidnes, B. Dalhus, M. Bjørås, G. Mathiesen and V. Eijsink, "Characterization of the Chitinolytic Machinery of Enterococcus faecalis V583 and High-Resolution Structure of Its Oxidative CBM33 Enzyme", *Journal of Molecular Biology*, vol. 416, no. 2, pp. 239-254, 2012.
33. G. Hemsworth, E. Taylor, R. Kim, R. Gregory, S. Lewis, J. Turkenburg, A. Parkin, G. Davies and P. Walton, "The Copper Active Site of CBM33 Polysaccharide Oxygenases", *Journal of the American Chemical Society*, vol. 135, no. 16, pp. 6069-6077, 2013.
34. Helen M. Berman,John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov,Philip E. Bourne, “The Protein Data Bank” Nucleic Acids Res (2000) 28 (1): 235-242, 2000.

35. Q. Cui and I. Bahar, Normal mode analysis, 1st ed. Boca Raton: Chapman & Hall/CRC, 2006
36. K. Hinsen, "Analysis of domain motions by approximate normal mode calculations", Proteins: Structure, Function, and Genetics, vol. 33, no. 3, pp. 417-429, 1998.
37. M. Tirion, "Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis", Physical Review Letters, vol. 77, no. 9, pp. 1905-1908, 1996.
38. I. Bahar, A. Atilgan and B. Erman, "Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential", Folding and Design, vol. 2, no. 3, pp. 173-181, 1997.
39. T. Haliloglu, I. Bahar and B. Erman, "Gaussian Dynamics of Folded Proteins", Physical Review Letters, vol. 79, no. 16, pp. 3090-3093, 1997.
40. A. Atilgan, S. Durell, R. Jernigan, M. Demirel, O. Keskin and I. Bahar, "Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model", *Biophysical Journal*, vol. 80, no. 1, pp. 505-515, 2001.
41. S. Kundu, J. Melton, D. Sorensen and G. Phillips, "Dynamics of Proteins in Crystals: Comparison of Experiment with Simple Models", Biophysical Journal, vol. 83, no. 2, pp. 723-732, 2002.
42. Prediction of Protein Protein interactions from evolutionary information. From P. Bourne and J. Gu, Structural Bioinformatics, 1st ed. Hoboken: Wiley-Blackwell, 2009.
43. C. Tsai, A. del Sol and R. Nussinov, "Allostery: Absence of a Change in Shape Does Not Imply that Allostery Is Not at Play", *Journal of Molecular Biology*, vol. 378, no. 1, pp. 1-11, 2008.
44. M. Wu, G. Beckham, A. Larsson, T. Ishida, S. Kim, C. Payne, M. Himmel, M. Crowley, S. Horn, B. Westereng, K. Igarashi, M. Samejima, J. Stahlberg, V. Eijsink and M. Sandgren, "Crystal Structure and Computational Characterization of the Lytic Polysaccharide Monooxygenase GH61D from the Basidiomycota Fungus Phanerochaetechrysosporium", *Journal of Biological Chemistry*, vol. 288, no. 18, pp. 12828-12839, 2013.
45. Agger, T. Isaksen, A. Varnai, S. Vidal-Melgosa, W. Willats, R. Ludwig, S. Horn, V. Eijsink and B. Westereng, "Discovery of LPMO activity on hemicelluloses shows the importance of oxidative processes in plant cell wall degradation", *Proceedings of the National Academy of Sciences*, vol. 111, no. 17, pp. 6287-6292, 2014.

APPENDIX

- **First add the path.**

```
addpath C:\Users\133812\Desktop\SCA5_forDist\sca5
clear; close all
```

- **Alignment loading and conditioning.**

```
[labels_seq,algn_full]=get_seqs('AA10.free');
N_seq=size(algn_full,1);
cut_off=.2;
fraq_gaps=sum(isletter(algn_full)==0)/N_seq;
algn=algn_full(:,fraq_gaps<cut_off);
N_pos=size(algn,2);
pdb_id='2bem'; chain='A';
pdb=pdbread(['' pdb_id '.pdb']);
```

- **Sequence similarity matrix.**

```
[S]=sim_seq(algn);
listS=nonzeros(triu(S,1));
h_seqsim=figure; clf;
set(h_seqsim,'Units','normalized','Position',[0 0.3 0.9
0.5],'Name','Sequence Correlations: PDZ');
subplot(1,2,1);hist(listS,N_pos/2);
xlabel('Pairwise SeqID','FontSize',14,'FontWeight','bold');
ylabel('number','FontSize',14,'FontWeight','bold'); grid on
figure(h_seqsim);
subplot(1,2,2); imshow(S,[0 1],'InitialMagnification','fit');
colormap(jet); colorbar;
title('SeqID', 'FontSize',12,'FontWeight','bold');
```

- **Positional conservation.**

```
[D_glo]=cons(algn);
h_D=figure; set(h_D,'Units','normalized','Position',[0 0.6 0.4
0.4],'Name','Positional Conservation');clf
subplot(2,1,1);hist(D_glo,25); grid on;
xlabel('Dconservation)','FontSize',10,'FontWeight','bold');
ylabel('number','FontSize',10,'FontWeight','bold');
subplot(2,1,2);bar([1:numel(ats)],D_glo,'k'); grid on;
axis([0 numel(ats)+1 0 4]);
set(gca,'XTick',[1:10:numel(ats)]);
set(gca,'XTickLabel',ats([1:10:numel(ats)]));
xlabel('position (2bem
numbering)','FontSize',10,'FontWeight','bold');
ylabel('D_i(conservation)','FontSize',10,'FontWeight','bold');
```

- **SCA calculations.**

```
[AA10sca]=sca5(algn);
```

- **Spectral (or eigenvalue) decomposition.**

```
[spect]=spectral_decomp(AA10sca,100);
```

- **Structure of top eigenmodes.**

```
h_3Dtopmodes=figure;
set(h_3Dtopmodes,'Units','normalized','Position',[0 0.7 0.3
0.4],'Name','Top Eigenmodes - 3D'); clf;
scatter3(spect.evpos(:,1),spect.evpos(:,2),spect.evpos(:,3),'ko','S
izeData', 50, 'MarkerFaceColor','b');
az=136;el=20;view(az,el);
xlabel('ev 1','FontSize',12,'FontWeight','b');
ylabel('ev 2','FontSize',12,'FontWeight','b');
zlabel('ev 3','FontSize',12,'FontWeight','b');
```

- **2D plots of top three modes:**

```
h_2Dtopmodes=figure;
set(h_2Dtopmodes,'Units','normalized','Position',[0 0 1.0
0.4],'Name','Top Eigenmodes-2D'); clf;
subplot(1,3,1);
scatter(spect.evpos(:,1),spect.evpos(:,2),'ko','SizeData', 50,
'MarkerFaceColor','b');
xlabel('ev 1','FontSize',12,'FontWeight','b');ylabel('ev
2','FontSize',12,'FontWeight','b');
subplot(1,3,2);
scatter(spect.evpos(:,1),spect.evpos(:,3),'ko','SizeData', 50,
'MarkerFaceColor','b');
xlabel('ev 1','FontSize',12,'FontWeight','b');ylabel('ev
3','FontSize',12,'FontWeight','b');
subplot(1,3,3);
scatter(spect.evpos(:,2),spect.evpos(:,3),'ko','SizeData', 50,
'MarkerFaceColor','b');
xlabel('ev 2','FontSize',12,'FontWeight','b');ylabel('ev
3','FontSize',12,'FontWeight','b');
```

- **A mapping between positional and sequence correlations.**

```
[U,sv,V]=svd(AA10sca.pwX);
N_min=min(N_seq,N_pos);
Pi=U(:,1:N_min)*V(:,1:N_min)';
U_p=Pi*spect.evpos;
h_SectSeq=figure; set(h_SectSeq,'Units','normalized','Position',[0
0.1 0.6 0.4],'Name','Mapping Seq Correlations by Positional
Correlations'); clf;
h_SectSeq(1)=subplot(1,2,1)
```

```
scatter3(spect.evpos(:,1),spect.evpos(:,2),spect.evpos(:,3),'ko','S
izeData', 50, 'MarkerFaceColor','b');
az=58;el=30;view(az,el);
xlabel('ev 1','FontSize',12,'FontWeight','b');
ylabel('ev 2','FontSize',12,'FontWeight','b');
zlabel('ev 3','FontSize',12,'FontWeight','b');
h_SectSeq(2)=subplot(1,2,2)
scatter3(U_p(:,1),U_p(:,2),U_p(:,3),'ko','SizeData', 50,
'MarkerFaceColor','b');
az=58;el=30;view(az,el);
xlabel('Seq 1','FontSize',12,'FontWeight','b');
ylabel('Seq 2','FontSize',12,'FontWeight','b');
zlabel('Seq 3','FontSize',12,'FontWeight','b');
```

- Sector definition.

```
h_secdef=figure;
set(h_secdef,'Units','normalized','Position',[0 1 .5
0.3],'Name','Top Eigenmode'); clf;
p_cutoff=0.8;
secpos = [];
histogram of the data
xhist=[0:.01:.4]; % make bins for the histogram based on the full
range of the data
[yhist]=hist(spect.evpos(:,1),xhist);
bar(xhist,yhist./N_pos,'k');hold on;grid on
```

- distribution fitting

```
pd=fitdist(spect.evpos(:,1),'lognormal');
x_dist=[min(xhist):(max(xhist)-min(xhist))/100:max(xhist)];
area_hist=N_pos*(xhist(2)-xhist(1)); % for proper scaling of the
pdf
pdf_jnk=pdf(pd,x_dist);
scaled_pdf=area_hist.*pdf_jnk;
plot(x_dist,scaled_pdf./N_pos,'r-','LineWidth',1.5);
```

- here, we make the cdf, and define sectors:

```
cdf_jnk=cdf(pd,x_dist);
clear sec cutoff_ev
[jnk,x_dist_pos_right]=min(abs(cdf_jnk-(p_cutoff)));
cutoff_ev = x_dist(x_dist_pos_right)';
```

- we obtain the indices of sector positions given the cutoffs

```
[sec.def] = find(spect.evpos(:,1)>cutoff_ev);
sprintf('%g+',str2num(char(ats(sec.def))))
sec.cutoff=cutoff_ev;
figure(h_secdef); line([cutoff_ev cutoff_ev],[0
max(yhist)/N_pos],'LineWidth',1,'LineStyle','--
','Color','b');sec.col=2/3;
```