Morpheus: A Fragment-Based Algorithm To Predict Metamorphic Behaviour In Proteins Across Proteomes

A Thesis
Submitted towards the partial fullfilment of
BS-MS dual degree programme by

V Vijay Subramanian

Under the guidance of

Anand Srivastava

Associate Professor, IISc Bangalore

(May 2024 – March 2025)

Date: 16 March 2025

Indian Institute of Science Education and Research Pune

Certificate

This is to certify that this dissertation entitled "Morpheus: A Fragment-Based Algorithm To Predict Metamorphic Behaviour In Proteins Across Proteomes" submitted towards the partial fulfillment of the BS-MS degree at the Indian Institute of Science Education and Research, Pune represents original research carried out by "V Vijay Subramanian" at "Indian Institute of Science, Bangalore", under the supervision of "Anand Srivastava" during academic year May 2025 to March 2026.

SUPERVISOR:
ANAND SRIVASTAVA
Associate Professor
Indian Institute of Science,
Bangalore

STUDENT:
V VIJAY SUBRAMANIAN
20201035
BS-MS
IISER Pune

COMMITTEE MEMBERS:
Supervisor: Prof. Anand Srivastava
Co-Supervisor: Dr. Rajeswari Appadurai
Expert: Prof. M S Madhusudhan

DATE:
16/03/2025

Declaration

I, hereby declare that the matter embodied in the report titled "Morpheus: A Fragment-Based Algorithm To Predict Metamorphic Behaviour In Proteins with Proteome Level Predictions" is the results of the investigations carried out by me at the "Indian Institute of Science, Bangalore" from the period 01-05-2025 to 01-03-2026 under the supervision of Professor Anand Srivastava and the co-supervision of Dr. Rajeswari Appadurai and the same has not been submitted elsewhere for any other degree.

STUDENT:
V VIJAY SUBRAMANIAN
2020135
BS-MS
IISER Pune

DATE:
16/03/2025

This thesis is dedicated to my parents,
Hari Subramanian and Kishore sir.

Acknowledgements

My thesis work was greatly enriched by the support and guidance of many individuals. I am truly grateful for being in a position to seek and receive their help.

I would first like to express my sincere gratitude to Professor Anand for welcoming me into his lab and consistently giving this project his time and attention. I have grown to admire his attitude towards research, open-source development and the meticulous care he gives to each project. My time in this lab has been instrumental in my learning and growth. I would like to wholeheartedly thank Rajeswari for being my constant source of support during the crucial moments of our project. Her insights and way of thinking greatly shaped the outcomes of my thesis. The confidence that both Anand and Rajeswari have placed in this project—and in my capabilities—carried me a long way. I would also like to extend my gratitude to my thesis expert, Professor Madhusudhan, who initially got me interested in Bioinformatics.

Working (and trying to work) in the lab was never not fun. My time in the lab has been an immensely rewarding experience, both intellectually and personally. I have had the privilege of learning from and working alongside an incredible group of people in the lab: Dipam, Raghav, Jai, Rashmi, Prachi, Namrata, Vinodhini, Prithvi, Mouli, Santosh, Harini, Irawati, Satya, Krishnakanth, Neha, Srinivasan, Gaurav, Sridip and Sangeetha. From learning different ways of thinking about science to losing 20 consecutive rounds in poker (I'm still sour about that), it has been quite an experience. Through conferences and weekly group meetings, I realize the importance of interaction and exchange of ideas in fostering innovation. A lot of the inputs at group meetings really helped me in improving my way of thinking. It made me think (and criticize) about my own project, which I feel is essential for growth in science. As I am close to leaving the lab, I now realize, more than ever, how open and intellectually stimulating the lab environment has been, and I am grateful for the role it has played in shaping me as a researcher.

A heartfelt thanks to my friends—Lakshmy, Mridul, Megha, Joan, Ajma, and many more—who held together through times of shared joy and sorrow. I'm happy to have found friendship among such a group of genuine folks.

I want to express my gratitude to Department of Science and Technology (DST) for providing the Kishore Vaigyanik Protsahan Yojana (KVPY) fellowship which supported my scientific endeavors.

I would like to acknowledge my family for their immense support throughout my research. I am thankful to my parents for supporting my decisions towards research and academia. They try their best to ensure that I do not get overwhelmed by research and responsibilities. I would like to acknowledge the significant contribu-

tions my brother, V Harikrishnan, made throughout the project: from helping me set up Linux for the first time on my system, to building and deploying the web server for Morpheus. I frequently kept seeking his help, even during his work hours (like with the daunting PBS scheduler script). I am deeply grateful to my uncle, Hari Subramanian, for his invaluable guidance and the countless hours he spent discussing potential ideas and identifying improvements in my thesis work. His support has been instrumental from the very beginning—encouraging me to apply to IISER, engaging in lengthy discussions on scientific topics, and recommending books, blogs, and lectures that have significantly shaped my scientific perspective.

Contributions

Contributor name	Contributor role
Rajeswari Appadurai and Anand Srivastava	Conceptualization Ideas
Rajeswari Appadurai	Methodology
Vijay Subramanian	Software
Vijay Subramanian	Validation
Vijay Subramanian	Formal analysis
Vijay Subramanian	Investigation
Anand Srivastava	Resources
Rajeswari Appadurai	Data Curation
Harikrishnan Venkatesh	Website Development
Vijay Subramanian	Writing - original draft preparation
Anand Srivastava	Writing - review and editing
Anand Srivastava	Supervision
Anand Srivastava	Project administration
Anand Srivastava	Funding acquisition

This contributor syntax is based on the Journal of Cell Science CRediT Taxonomy[1]

[1] https://journals.biologists.com/jcs/pages/author-contributions

Abstract

While the majority of proteins adhere to the traditional one-sequence/one-structure model, an increasing number of proteins have been found to possess dual structures and functions. These "fold-switching" or "metamorphic" proteins can undergo significant structural changes under native conditions or in response to specific cellular stimuli. This ability to "switch folds" allows proteins to perform multiple functions, enabling cells to tightly regulate various biochemical processes. Despite their importance in health and disease, not much work has been performed that could assign a fold-switching protein from sequence information. Here, we present a fragment-based approach to build a classifier that can predict metamorphic behaviour in protein sequences with an average accuracy of 84.7% and a Matthew's correlation coefficient (MCC) of 0.698. To develop our classification algorithm, we use the structural data from the experimentally solved structures in the protein data bank (PDB) and AlphaFold Protein Structure Database, which is a repository of computationally solved structures. Our classifier works by analysing the diversity of structures the fragments within a query protein occupy. The algorithm takes in a query protein sequence as input, fragments the sequence using a sliding window, performs a fragment search across the database and gives a binary output prediction using a support vector machine (SVM) to perform the classification. We employed our algorithm on 57 different proteomes consisting of a total of 601,218 proteins (one sequence per gene). We identified about 10% of these proteins have the ability to fold switch, significantly expanding the known metamorphic reservoir of proteins. Potential candidates with high confidence scores are shortlisted for further experimental evaluations. Additionally, we have built a web server for Morpheus in the event that the user's protein of interest is absent from the data published (http://mbu.iisc.ac.in/~anand/morpheus). Our work is the first implementation of a proteome-level predictor of metamorphic proteins, and it helps in the selection of potential candidate protein sequences to evaluate their metamorphic behaviour through further experimentation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Protein Structure

Proteins are functional bio-molecules of the cell present in all organisms of life. Amino acids are the monomeric building blocks of proteins, and they can be characterized by their amino acid sequence. Proteins are created in all life forms by stringing together amino acids in a linear chain to form a final protein sequence. The chemical structure of an amino acid is shown in Figure 1.1. The α carbon of amino acids is bonded to four different chemical groups: an amino group, a carboxyl group, a hydrogen atom and a variable side chain. There are 20 different standard amino acids, each one differing only in the side chain 'R'.

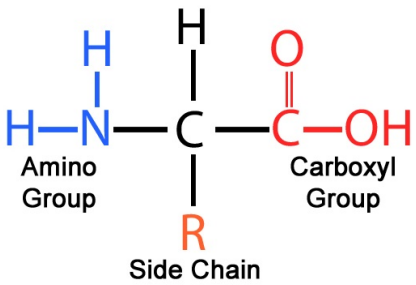

Figure 1.1: **Amino acid chemical structure:** The chemical representation of an amino acid, which contains an Amino group, a carboxyl group and a variable side chain 'R'

Depending on the side-chain 'R', amino acids have different physical properties such as polarity, hydrophobicity, etc. The amino acids in the protein sequence interact with each other (and the solvent) through electrostatic and Van der Waals forces. They attract/repel and fold into a final three-dimensional structure. This 3-D structure is closely correlated to the function that the protein undertakes. There are about 20,000 different proteins in the human body, and analyzing the structure of these proteins becomes crucial in understanding their functional interdependence. The structure of any protein can be understood from 4 different hierarchical levels. These hierarchies are explained in the following subsections.

1.1.1 Primary Structure of Proteins

This level corresponds to the amino acid sequence of the protein. This can be represented as a sequence of alphabets, where each alphabet corresponds to an amino acid. The code for the standard amino acids is shown in table 1.1.

Amino Acids	Single Letter Code
Glycine	G
Alanine	A
Valine	V
Leucine	L
IsoLeucine	I
Threonine	T
Serine	S
Methionine	M
Cysteine	C
Proline	P
Phenylalanine	F
Tyrosine	Y
Tryptophane	W
Histidine	H
Lysine	K
Argenine	R
Aspartate	D
Glutamate	E
Asparagine	N
Glutamine	Q

Table 1.1: Amino acid codes for the 20 standard amino acids.

Amino acids can be grouped into different categories based on charge, polarity and hydropathy. The grouping is shown in figure 1.1.1. Amino acids with a polar side group are more likely to be on the surface of proteins. By interacting with water, the solvent, they make proteins soluble in water. Alternatively, amino acids with non-polar side groups avoid water and form the insoluble core of proteins. The polarity of amino acids is one of the driving forces that shape the final 3D structure of proteins.

Cysteine (C), Glycine (G) and Proline (P) display unique properties that set them apart from the rest of the amino acids. Cysteine (C) has a reactive -SH group as the side-chain. One Cysteine can oxidize with another Cysteine to form a C-S-S-C bond called a disulfide bond. This kind of interaction can stabilize different parts of the protein sequence and bring them together. Glycine (G) is the smallest amino acid due to hydroegen being the side chain. This gives glycine the ability to fit into tight spaces and have relatively lesser steric hindrance. It is also the only amino acid which does not exhibit L/R stereo-isomerism due to the presence of two hydrogen groups. Unlike other amino acids, proline has a cyclic ring that is produced by the

formation of a covalent bond between its R group and the amino acid group on C_α. This makes proline rigid, and its presence creates a kink in the protein chain.

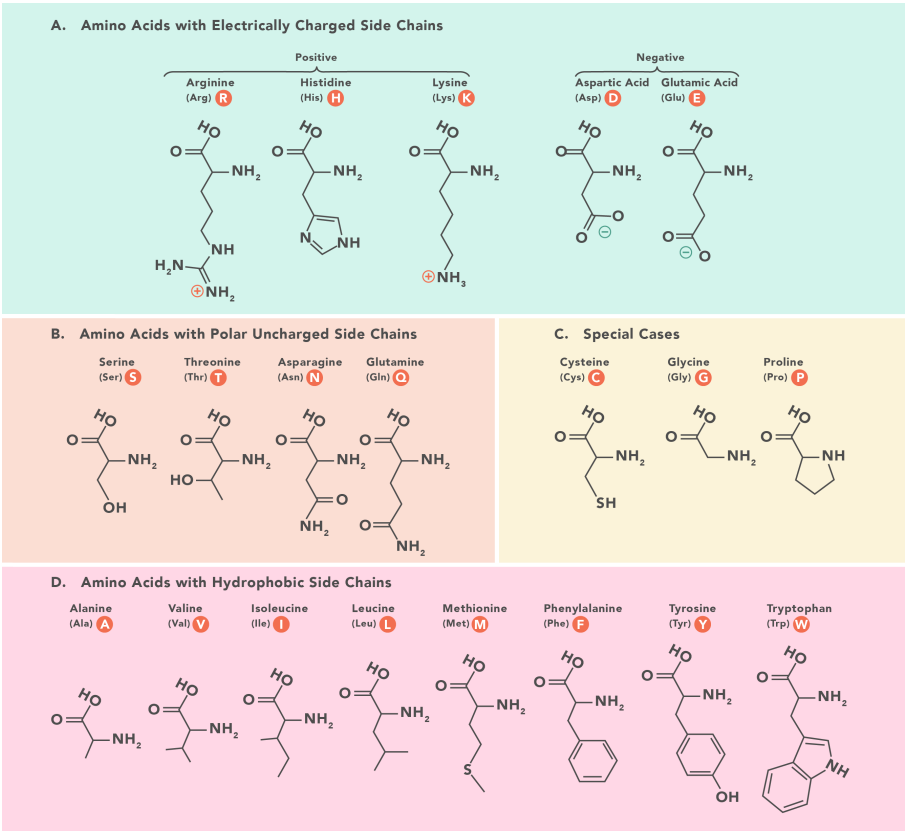

Figure 1.2: **Amino acid categorization:** 20 standard amino acids are grouped on the basis of charge, polarity and hydropathy. Image borrowed from the internet - technologyworks.com

.

1.1.2 Secondary Structure of Proteins

The secondary structure of proteins is the local spatial conformation of the polypeptide backbone, excluding the side chains. The two most common secondary structure elements are α-helices and β-sheets. The schematic and hydrogen bonding of both α helix and β sheet is shown in Figure 1.3.

α helix: When Polypeptide segments assume a regular spiral or helical conformation, such a secondary structure is called the α helix. Here, the carbonyl carbon of each amino acid is hydrogen-bonded with the amide hydrogen of the amino acid, four residues towards the C-terminus. Hence, the regular α helix has a pitch of 3.6 amino acids per turn.

β sheet: This secondary structure consists of laterally packed β strands. Each β strand is a short, nearly fully extended polypeptide chain. The hydrogen bonding

pattern in a β sheet takes place between carbonyl oxygen from one sheet and the amide hydrogen from a parallel sheet.

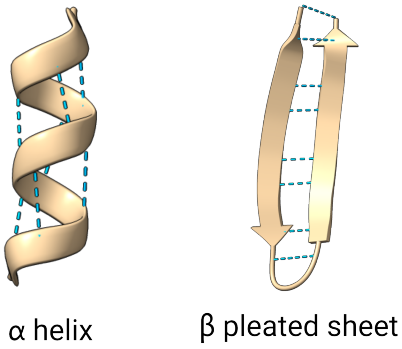

Figure 1.3: α **helix and** β **sheet:** Secondary structure depiction of α helix, β sheet and their hydrogen bonding patterns.

1.1.3 Tertiary Structure of Proteins

The tertiary structure of proteins is the arrangement of secondary structures in 3-dimensional space. Figure 1.4 shows the tertiary structure of human myoglobin, one of the first proteins that was experimentally solved. Myoglobin belongs to the globin superfamily of proteins, and human myoglobin contains 154 residues. These amino acids are arranged in 3D space in the form of 8 helices connected by loops.

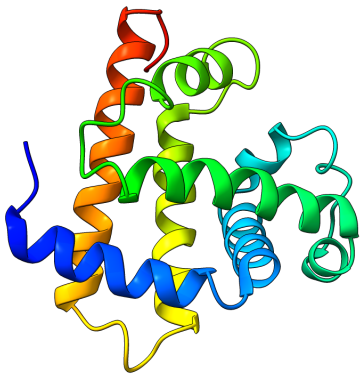

Figure 1.4: **Myoglobin:** The 3D structure of one of the first proteins that was solved: myoglobin. (PDB:1MBN). The residues are coloured from blue to red, starting from the N-terminal.

1.1.4 Quaternary Structure of Proteins

When two or more polypeptide chains are held together by non-covalent bonds, they form a multimeric protein. The relative arrangement of the polypeptide sub-units in space refers to the quaternary structure of proteins. These sub-units can consist of the same polypeptide chains (homomeric structure) or of different subunits (heteromeric structure).

1.2 Anfinsen's Paradigm

For the first time, in 1961, Anfinsen demonstrated that all the information for a protein to fold into its native state is contained within the amino acid sequence. Anfinsen *et al.* (1961); Anfinsen (1973). This is what was known as the 'Thermodynamic hypothesis'. In its normal physiological conditions (solvent, pH, ionic strength, temperature, and other), the native three-dimensional structure of a protein is the one in which the Gibbs free energy is the lowest. This would imply that the native conformation is determined by the totality of the inter-atomic interactions and, thereby, the amino acid sequence of the protein.

One of the series of experiments that supported the thermodynamic hypothesis is the homogeneous refolding into the native state of the so-called "scrambled" ribonuclease. Ribonuclease A is a 124 residue-long amino acid with 8 SH groups present on the 8 Cysteines. In its native tertiary structure, Ribonuclease A forms 4 disulfide bonds. When the fully reduced protein is allowed to re-oxidize under denaturing conditions, many disulfide-bonded isomeric forms are obtained. If the denaturing agent is removed and a small amount of sulfhydryl group is introduced to the "scrambled" protein, disulfide interchange takes place, and the mixture is converted into a homogeneous product of the original ribonuclease protein.

1.2.1 Challenges to Anfinsen's Paradigm

Many recent studies reveal a growing class of proteins that pose a challenge to Anfinsen's paradigm. These proteins defy the dogma by either performing its intended function but still not having a stable tertiary structure or by having multiple stable protein folds. These classes of proteins are listed below:

- Intrinsically Disordered Proteins
- Metamorphic Proteins
- Misfolding Proteins

Intrinsically disordered proteins are a class of proteins that a lack a stable three dimensional structure in their native state. This implies that the protein sequence, in the case of these proteins, encodes information for the protein to exist in an ensemble of disordered states. Metamorphic proteins are those proteins that have the ability to inter-convert between two stable native conformations. The protein sequence of metamorphic proteins encodes information for degenerate structures.

Furthermore there are proteins that undergo aggregation and misfolding. For example, Prions are stable conformations of proteins that differ from their native state.

1.3 Intrinsically Disordered Proteins

Anfinsen's paradigm has expanded from "one sequence, one fold" to" one sequence, an ensemble of conformations" with highly labile but functional intrinsically disordered proteins (IDPs) Dunker *et al.* (2001). IDPs are those proteins that lack a stable three dimensional structure under physiological conditions and exhibit high conformational heterogeneity. Different parts of the polypeptide chain might sample different secondary structures undergoing an order $\leftrightarrow$ disorder transition in a timescale of nanoseconds. This structural plasticity allows IDPs to function in signal transduction, molecular recognition and have the ability to interact with multiple binding partners. They perform their function through an ensemble of conformations.

When protein structure is determined through X-ray crystallography, some segments of the proteins do not have a discernable electron density. Although the protein crystallized, these regions showed considerable structural heterogeneity, and hence, the electron diffraction resonance in this region did not lead to discernable electron density. These 'unobserved' constitute the intrinsically disordered region of the protein. In some proteins, it was discovered that although these segments do not have a stable, folded structure, they are important for the function of the protein.

IDPs often contain a high proportion of polar and charged amino acids while having lesser amounts of hydrophobic residues, which prevent them from folding into stable structures. Many Intrinsically disordered proteins are also associated with causing neurodegenerative diseases in humans Uversky *et al.* (2008).

1.4 Metamorphic Proteins

Metamorphic proteins are a class of proteins that have the ability to exist in two conformations that are stable and significantly different in their secondary or tertiary structures. Murzin (2008); Jain *et al.* (2014) The two conformations are of comparable stability that interconvert in the absence of other ligands or cofactors. The two conformations correspond to highly dissimilar folds and possibly two different functions that the protein undertakes. The schematic energy landscape diagram of a metamorphic protein is depicted in Figure 1.5.

Often, there is a biological trigger, such as temperature, salt conditions, or the redox state associated with the conformational change that follows. A few of the well-studied metamorphic proteins include the following: (a) RfaH, a virulence

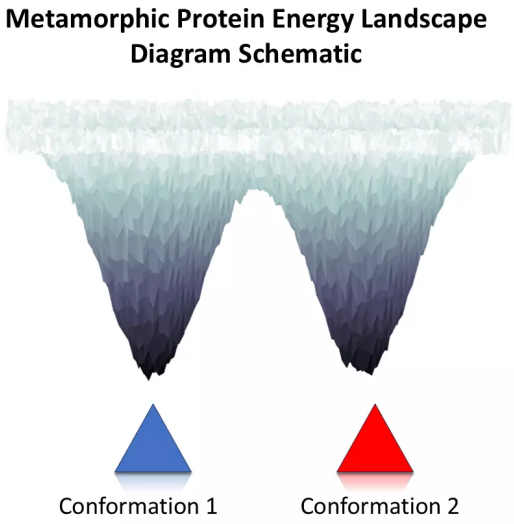

Figure 1.5: **Metamorphic protein energy diagram schematic:** Schematic of the energy landscape of a metamorphic protein with two structurally different conformations.

factor found in E. coli. RfaH is a multi-domain protein where the inter-domain interaction can modulate the conformation of the C terminal domain between an alpha-helical hairpin and a beta-barrel conformation. The all-alpha fold of RfaH functions to inhibit its activity, while the all-beta fold enables RfaH to activate transcription and facilitate translation Burmann *et al.* (2012); Artsimovitch and Ramírez-Sarmiento (2022) (b) KaiB, the cyanobacterial protein switches fold to regulate the phosphorylation/de-phosphorylation phases of the circadian clock KaiABC which allows it to maintain a periodicity of 24 hours. Chang *et al.* (2015) (c) XCL1, the human Lymphotactin is a metamorphic protein that undergoes a conformational rearrangement from a monomeric α+β chemokine fold to a novel dimeric all-β fold Tuinstra *et al.* (2008). The canonical chemokine fold performs as an agonist for XCR1 (a G-protein coupled receptor), while the novel dimeric fold binds to surface glycosaminoglycans Khatua *et al.* (2020). The trigger for the fold-switch in Lymphotactin is temperature, and under physiological conditions, both folds are populated. Known metamorphic proteins are pervasive across proteomes from many kingdoms of life and are estimated to constitute 0.5 - 4 % of the Protein Data Bank (PDB).Porter and Looger (2018)

Another class of proteins that are closely related to metamorphic proteins are called fold-switching proteins. They are those proteins that show the fold-switching nature where there is a change in conformation, but this change is irreversible. This makes the two structures non inter-convertible, hence creating the difference between the two classes: metamorphic and fold-switching proteins. Note the distinction between metamorphic proteins and fold-switching proteins that Porter and coworkers put forth:Kim and Porter (2021) the main difference between metamorphic and fold-switching proteins is that the secondary structure remodelling events of fold-

switchers can be either reversible or irreversible but that of metamorphic proteins must always be reversible. Hence, metamorphic proteins form a subset of fold-switching proteins.

In the following subsections 1.4.1 - 1.4.3, three examples of metamorphic proteins: RfaH, KaiB and Lymphotactin, will be discussed in more detail.

1.4.1 RfaH

RfaH is a bacterial transcription factor part of the NusG family of proteins. Regulators of this family of proteins are the only universally conserved family of transcription factors. Werner (2012). The C Terminal domain of a canonical protein of the NusG family is usually arranged in the form of a β barrel. While for RfaH, the C-terminal domain is arranged in the form of an α helical hairpin. This is the fold-switching region of the metamorphic protein RfaH. The trigger that controls this fold-switching behaviour is the inter-domain interaction between the N-terminal and C-terminal.

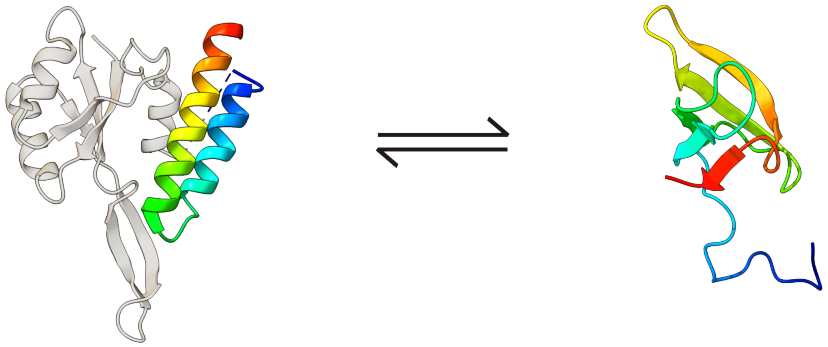

Figure 1.6: **Rfah:** The two different conformations of the C-Terminal domain of RfaH. PDB ID: 2OUG (left), 2LCL (right)

In free RfaH, the RNA-Polymerase binding site is blocked by the C-terminal domain helix. Only in the presence of a small DNA element, called ops (operon polarity suppressor), the inter-domain contact of RfaH is reduced, and the C-terminal domain undergoes a structural transformation to form a β barrel structure.

While the structural studies of wild-type RfaH show a α helical hairpin in the context of the whole protein sequence, NMR studies of RfaH, when only the C-terminal domain is studied, show experimental characteristics of β barrel secondary structure. Figure 1.6 shows the two conformations α helical hairpin and β barrel that the C-terminal domain of RfaH takes.

1.4.2 KaiB

The 105 residue cyanobacterial protein KaiB so far, is the only metamorphic protein discovered to be associated with the circadian rhythm of an organism. KaiB is one of three proteins in a system of KaiA, KaiB and KaiC, whose phosphorylation/de-phosphorylation cycle helps cyanobacteria entrain itself to the time of day.
The phosphorylation/de-phosphorylation cycle of KaiA-B-C works as follows. At the start of the day, Thr432 and Ser431 of KaiC are unphosphorylated. Towards Noon, Thr432 gets phosphorylated and KaiA binds to KaiC. Kim *et al.* (2008) This stimulates auto-phosphorylation. Once Ser431 is also phosphorylated, the domains of KaiC rearrange to accommodate a binding site for KaiB. KaiB further binds not only to KaiC but also to KaiA to make a ternary complex. This causes KaiA to unbind from KaiC, initiating the de-phosphorylation phase of the circadian rhythm oscillation.

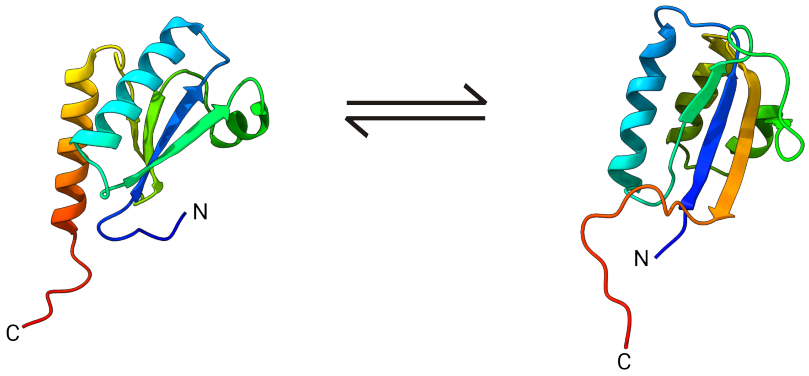

Figure 1.7: **KaiB:** The two different conformations of KaiB. The timescale involved in the fold-switch is of the order of hours. PDB ID: 5JYT (left), 2QKE (right)

In the free state, KaiB exists as a tetramer made of two asymmetric dimers. The tertiary structure of the individual monomer is shown in Figure 1.7 (right). Each of the monomers assumes a fold consisting of 4 β strands and 3 α helices. In the fold-switched state, Kaib assumes a thioredoxin fold, as shown in Figure 1.7 (left). The C terminal half of KaiB undergoes considerable rearrangement to switch from the novel tetrameric fold to the thioredoxin fold.

1.4.3 XCL1

XCL1, or the Human Lymphotactin, is a metamorphic protein present in *homo sapiens*. XCL1 is a chemokine, a small signalling protein, part of the immune response system which recruits white blood cells in humans. Lymphotactin is a 94 residue protein, which is one of the first proteins identified to be a metamorphic protein. Tuinstra *et al.* (2008)

The fold-switching in Lymphotactin takes place with temperature being the trigger involved. At 10° Celsius, it exists predominantly in the canonical chemokine fold, while at 40° Celsius, there is a dramatic rearrangement of secondary structure in order for the protein to dimerize and accommodate an additional beta-strand at the N-terminus. The two different conformations of Lymphotactin have distinct functions as a chemokine. The two functions that are originally carried out by the same fold in other chemokines are segregated in Lymphotactin. The two functions that are carried out by proteins of the chemokine family are:(i) Bind to G Protein-coupled receptor (GPCR) on the surface of leucocytes to stimulate migration and (ii) Bind to Glycosaminoglycans (GAG) in order to establish a concentration gradient for chemotaxis.

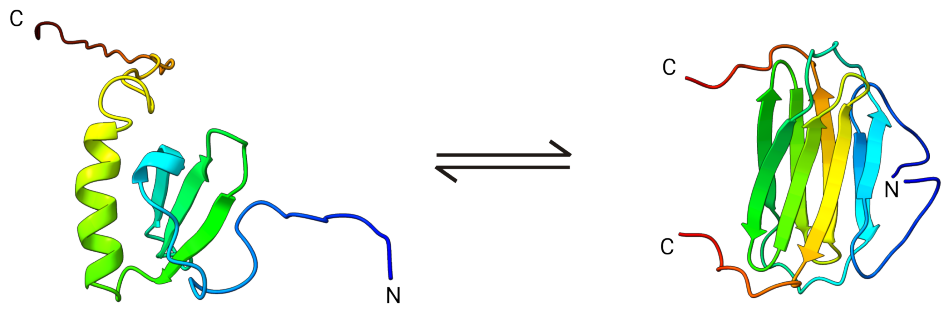

Figure 1.8: **Lymphotactin:** The two conformations of human Lymphotactin at different temperatures. Lymphotactin forms a homo dimer when the fold switch takes place. The timescale involved in the fold-switch is of the order of 1 second. PDB ID: 1J8I (left), 2JP1 (right)

In the case of Lymphotactin, the functions are segregated as follows:(i) Ltn10 is unable to tightly bind GAGs, it is an effective agonist for XCR1 (a type of GPCR) (ii) Ltn40 binds to GAGs with high affinity. Fox *et al.* (2015)

Figure 1.8 shows the two conformations involved in the metamorphic protein, human Lymphotactin.

1.5 Protein Structure Prediction

For the last 50 years, there has been a cumulative effort in the field of structural biology to computationally predict the 3D structure of a protein, given the amino acid sequence. Multiple groups have contributed to advances in protein structure prediction and the last decade has seen a revolutionary breakthrough in protein structure prediction. The subsections 1.5.1 and 1.5.2 describe two popular methods for ab initio prediction of protein structure.

1.5.1 Rosetta

Rosetta is a comprehensive software suite for modelling macromolecular structures developed by Prof. David Baker's group at the University of Washington. Rosetta began as a structure prediction tool and has consistently been a strong performer in the same since its inception. Rosetta is not a pure ab initio structure prediction method, as it uses protein fragment-based Monte Carlo search to predict the final structure of the protein.Abbass and Nebel (2020)

The structure prediction in Rosetta starts with acquiring information on the protein sequence, such as predicted secondary structure and a fragment library of similar fragments of sizes 3 and 9 residues, and uses this information to predict the native protein structure. Rosetta starts with an initial backbone conformation and then uses fragment insertion based on the fragments from the fragment library. These fragments are randomly sampled and inserted at each position. This initial structure is then used to perform a Monte Carlo sampling process to minimize their custom energy function. This type of fragment-sequence enhanced structure prediction has shown quite some success among its competitors.

However, when it comes to metamorphic proteins, Rosetta still is not able to predict the alternate conformations. It is not quite certain whether the ab initio based structure predicting tool of Rosetta can be altered/biased to predict the fold-switched state of metamorphic proteins. Further research needs to be performed in order to test this idea.

1.5.2 Alphafold

Alphafold is a deep learning based protein structure prediction tool that revolutionized the field of structural biology. AlphaFold works by leveraging deep learning models trained on vast datasets of known protein structures: The Protein Data Bank (PDB).Jumper *et al.* (2021) The PDB contains around 200,000 structures of proteins, including redundant protein structures and proteins bound with other molecules. Alphafold leverages both evolutionary data (Multiple sequence alignment) and structural data (PDB) in conjunction to predict protein structure. Information from these two different sources is allowed to interact with each other, making the predictions of Alphafold more robust.

Alphafold employs an advanced machine learning algorithm called a transformer network, which captures long-range dependencies between amino acids, enabling it to accurately model protein folding. The iterative process of using evolutionary and structural data along with the transformer architecture allows Alphafold to predict protein secondary structure up to atomistic accuracies.

Further, the Deepmind team has improved upon their model to create Alphafold3, which is now able to predict biomolecular interactions in protein complexes. Abramson *et al.* (2024) This includes not only protein-protein interaction but interactions

with nucleic acids, metal ions and ligands as well.

1.6 Alphafold Fails To Predict Fold-switching

Although Alphafold is able to predict structures of folded proteins up to a never-foreseen atomic level accuracy, studies have shown that it is unable to make predictions in the case of structurally degenerate proteins. In the case of metamorphic proteins, Alphafold always predicts the structure of one of the two alternate folds. When 98 pairs of fold-switching proteins were run through Alphafold, almost 94% of the time, only one of the experimentally solved structures were predicted. Chakravarty and Porter (2022)
Also among the conformation that alphafold predicts for metamorphic protein, it captures the ground state protein 70% of the time. Here ground state refers to either the preferred conformation of the protein suggested in literature (whenever available) or the conformation taken by it in isolation. Moreover, whenever Alphafold predicts the structure of a metamorphic protein, it gives a moderately high confidence score for the prediction. This is unlike the case of IDPs, where the disordered regions mostly have a low prediction confidence score.

This creates a need to either modify and tweak Alphafold to predict the alternate conformation or find new ways of predicting alternate conformations of metamorphic proteins. In the next chapter, we will discuss a few existing methods in literature to identify metamorphic proteins; including a method of biasing alphafold to predict the alternate conformations of metamorphic proteins.

Chapter 2

Existing Methods To Predict Metamorphic Proteins

2.1 Introduction

All the metamorphic proteins in the proteome can be called the "metamorphome". Estimates have been made, but it is unclear how much of this class of proteins is present in the proteome overall. Protein sequences must be computationally screened for possible metamorphic behaviour and then validated using experimental methods in order to uncover the metamorphome in its entirety. There have been previous attempts to predict metamorphic behaviour in proteins using protein sequence information. LiWang's group proposed one of the methods, which was built on the hypothesis that metamorphic proteins can "confuse" secondary structure prediction algorithms Chen *et al.* (2020). Leveraging this confusion, the authors devised the diversity index, a metric to quantify the uncertainty of assigned secondary structure by the structure prediction tool. Porter's group could also predict those fold-switching proteins that undergo alpha-helix to beta-sheet transition using sequence information alone to a high degree of accuracy Mishra *et al.* (2021). This particular type of fold-switching protein, where there is a major switch from alpha-helical conformation to a beta-sheet, constitutes a relatively small portion of the known metamorphic proteins overall. Porter's group put forth another approach that tackled a parallel problem where two different proteins with high levels of aligned similarity were assessed to see if they assumed the same fold or different ones Kim *et al.* (2021). In contrast to fold-switching proteins, the sequence-similar fold-switching proteins described before remodel their secondary structures in response to mutations.

Another set of approaches that are designed to address a broader problem are methods that employ Alphafold to study conformational changes. Osman (2023); Wayment-Steele *et al.* (2024) Although these methods have an advantage over bioinformatic approaches of being able to predict the two different 3D conformations, they tend to be computationally intensive approaches. Such methods would be highly useful when performing detailed studies of a few proteins but would not be the first pick when making high-throughput studies. When these methods are used

in conjunction with high-throughput bioinformatic approaches, good results can be achieved when compared to the results obtained individually. It is unclear whether the results obtained from Alphafold-based techniques are generative or based on memory; there is evidence which supports both sides of the argument. Chakravarty *et al.* (2023, 2024)

2.2 Diversity Index Based Classifier - Chen et al.

The main hypothesis that led to the creation of this classifier was that 'Metamorphic proteins can confuse secondary structure predicting tools'. Chen *et al.* (2020) Chen and co-workers noticed that when metamorphic protein sequences are used as input into four different secondary structure predicting tools - PSIPred, SPIDER2, SPIDER3, Porter5.0; the fold-switching region of those metamorphic proteins showed degenerate propensity in two or more secondary structure categories. This means that overlapping propensities are predicted for the same residues for helix (H), sheet (E), and loop(L). To test whether this degeneracy could be used to classify metamorphic proteins from monomorphic ones, they created a training dataset of metamorphic and monomorphic proteins and ran them on all four of these secondary structure predictors. Consequently, a significant number of the metamorphic proteins showed degeneracy/confusion in secondary structure prediction in the region in which they switch folds; the monomorphic proteins more or less showed a consistent prediction with a few exceptions here and there.

In order to quantify this degeneracy, the authors came up with a metric calling it the 'Diversity Index (DI)'. It is a polynomial form of Shannon entropy. Let h, e and l be the predicted propensity of a residue to be in helix, sheet and loop secondary structure, respectfully. The diversity index is then defined as shown in Eq. 2.1.

$$\text{DI} = \frac{1}{h^2 + e^2 + l^2} \tag{2.1}$$

The value of DI takes a minimum value of 1, in the case where one of h, e and l is equal to 1 and the remaining two equal to 0. DI takes a maximum value of 3 when $h = e = l = 1/3$. The diversity index is defined at each residue of the protein sequence. The final score for a protein sequence that is used for classifying the metamorphic behaviour is the maximum of the rolling average of the diversity index as shown in Eq. 2.2.

$$\text{DI score} = max \left\{ \frac{1}{WW} \sum_{k=1}^{L-WW+1} DI_k \right\} \tag{2.2}$$

Next, to implement a classification model, they use a threshold parameter which dictates whether the DI score is high enough to be classified as a metamorphic protein.

$$\text{Classification} = \begin{cases} \text{Metamorphic} & \text{DI score} > \text{DI}_{thres} \\ \text{Monomorphic} & \text{DI score} \leq \text{DI}_{thres} \end{cases} \tag{2.3}$$

There are two parameters involved in this classifier ,namely, WW (the window width of rolling average) and DI_{thres} (the threshold for DI score beyond which the algorithm classifies it as metamorphic). These two parameters were chosen based on which pair classified the training data the best. Each secondary structure predictor based classifier required a different set of parameters.

A 6-fold cross-validation scheme was implemented on the labelled training dataset of metamorphic and monomorphic proteins to test the accuracy of the model. Table 2.1 shows the accuracy measures that were achieved after the cross-validation.

Measure	Chen et. al
Sensitivity	0.41 (0.02)
Specificity	0.92 (0.02)
Precision	0.88
Accuracy	63.3% (1.6)
F1 Score	0.56
MCC	0.327 (0.12)

Table 2.1: The accuracy measures obtained upon 6-fold cross-validation for the Diversity Index (DI) based classifier developed by Chen and co-workers. The values in the brackets are the standard deviation across the six folds. For an explanation of these accuracy measures, check Sec. 3.6

2.3 JPred Based Classifier - Mishra et al.

The method developed by Mishra and co-workers is a sequence-based classifier for identifying those metamorphic proteins that undergo $\alpha - \text{helix} \leftrightarrow \beta - \text{strand}$ fold-switching transitions. Although the number of currently known metamorphic proteins contain a small fraction that undergoes $\alpha - helix$ to $\beta - strand$ transition, It is certainly a useful tool to have for a high-throughput screening of these types of metamorphic proteins.

This method hinges on two crucial observations: (i) Discrepancies in JPred (a protein secondary structure prediction server) prediction is a robust classifier of metamorphic proteins which undergo $\alpha - helix \leftrightarrow \beta$-strand transition, (ii) the fold-switching regions (FSRs) of some known metamorphic proteins have different secondary structure propensities when expressed by themselves than when they are predicted using the whole sequence. Using these two observations, Mishra and co-workers came up with a metamorphic protein classifier with accuracy measures shown in Table 2.2.

Measure	Mishra et. al
Sensitivity	0.64
MCC	0.71

Table 2.2: The accuracy measures attained by the classifier constructed by Mishra and co-workers. For an explanation of these accuracy measures, check Sec. 3.6

To quantify this discrepancy, the authors first ran both the parent sequence and the fold-switching region (FSR) through JPred and obtained the results. The resulting secondary structures were aligned corresponding to the sequence alignment obtained by aligning the parent sequence and FSR using a gap open/extension penalty of -1/-0.5. The secondary structure discrepancies between the predictions were summed residue-by-residue (1 for discrepancy, 0 for no discrepancy). This is summed up by equation 2.4

$$\text{Fraction}(\alpha - \beta\text{Discrepancy}) = \frac{\#\ \alpha - \beta \text{ substitutions}}{\text{len(FSR)}} \tag{2.4}$$

This high-throughput method can be further used as validation for Morpheus once we make predictions on data our model has not encountered before. Due to the high TPR value (true positivity rate) of the above algorithm, whatever sequences it does predict as metamorphic has a high chance of actually being a true positive hit.

2.4 Biasing Alphafold To Study Conformational Changes in Proteins

Since its inception, Alphafold has been extensively used and modified to utilise the full potential of the transformer-based deep learning model. One of the applications was to bias Alphafold to predict the alternate structure of proteins. Multiple groups have come up with methods to study conformational changes in proteins using Alphafold. Here we take a look at one of the methods where the input multiple sequence alignment (MSA) is modified in an attempt to predict the alternate conformation.Schafer and Porter (2023) While predicting the structure of proteins, one of the features that Alphafold leverages from the MSA are co-evolved amino-acid pairs.Göbel *et al.* (1994) Many early studies noted that co-varying amino-acid pairs tend to be in direct contact with each other. Hence the 3-dimensional arrangement of these amino acid pairs can be inferred from analysing these co-evolved pairs from homologous sequences in the multiple sequence alignment.

As Alphafold is unable to predict the alternate conformation of metamorphic proteins, the authors hypothesised that co-evolutionary contacts corresponding to the alternate fold are being masked by the homologous sequences of the other fold. In order to test this hypothesis, the input MSA to Alphafold was pruned to contain the subfamily-specific protein sequences that were the closest homologues to the query protein sequence. This way, when the MSA was filtered and provided to Al-

phafold, it was able to predict two conformations of a candidate sequence without experimentally determined structures.

Furthermore, the Authors constructed a blind predictive approach to predict metamorphic proteins and successfully identified fold switching in 13/56 known fold switchers (23%) with zero false positives. The low true positivity rate makes this method reliable when analyzing the positive predictions made, but its low true positivity rate also implies that the method misses a lot of the true positives while making high-throughput predictions.

While it is worthwhile trying to predict metamorphic proteins using sequence information alone, the approach falls short at making proteome-level predictions due to a high false negativity rate. Methods involving biasing Alphafold to predict the alternate conformations of metamorphic proteins are often unreliable, resource intensive and of high time complexity. Here, we present Morpheus, a fast fragment-based predictor of metamorphic proteins that can achieve an average cross-validation accuracy of 84.7% with Matthew's correlation coefficient of 0.70. To achieve this, we have implemented a Trie-based search to tremendously expedite the fragment-picking performance for large proteomes. In chapter 3, we introduce the methods used for the construction of our algorithm, Morpheus and detail the methods used in each step along the way to make predictions.

Chapter 3

Methods - Morpheus

Morpheus is able to classify protein sequences as metamorphic or monomorphic by leveraging the existing structural databases: the Protein Data Bank (PDB) and AlphaFold2. A database is curated, consisting of around 200,000 experimentally solved structures from the PDB and 1.2 million computationally solved structures from AlphaFold2. Given a query protein sequence, the algorithm is initiated by breaking the protein sequence into fragments of 7 residues using a sliding window. The curated database is searched for these identical fragments, and their structural information is extracted. The structural information of these fragments is then analysed and scored to make a binary prediction of the metamorphicity of the query protein (See Figure 3.1 for schematic). The computations required to create this structural classifier are quite involved and are detailed in the sections below.

3.1 Database Curation

Corresponding to each protein structure (whether experimentally or computationally solved), the following information is stored in a single file:

- Structure ID (PDB ID and Alphafold accession ID)
- Protein sequence
- Protein sequence cluster index
- Secondary structure (assigned by DSSP)
- pLDDT Score for each residue (applicable only for computationally solved structures)

Secondary structure: A secondary structure is assigned to each residue based on the coordinates of the residues in the protein structure. This assignment is performed by DSSP Kabsch and Sander (1983) which uses the internal coordinate dihedral angles (ϕ, ψ) for each residue and hydrogen bonding contacts to determine the secondary structure. The secondary structure assigned can take 1 of 8 possible outputs. They are α-helix (H), 3_{10} helix (G), π helix (I), poly-proline helix (P), extended strand (E), β bridge (B), turn (T) and bend (S). The letters in parentheses are the standard notation used by DSSP.

pLDDT score: The pLDDT score is a measure of the confidence of prediction (probabilistic Local Distance Difference Test) that Alphafold uses on its computational structure prediction. This pLDDT score (a value between 0 and 100) for each residue of the predicted protein, where the higher the value, the more confident the prediction by Alphafold. This score is extracted from all the computationally solved proteins from Alphafold2 and is saved in the database.

3.2 Fragment Picking

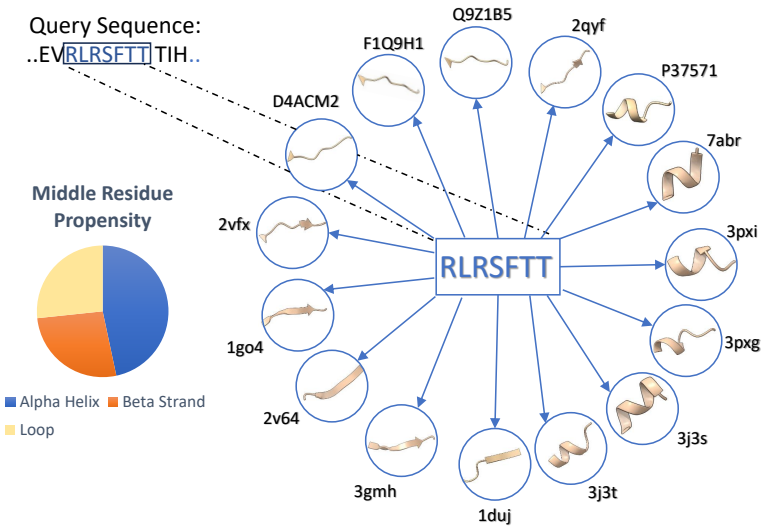

Figure 3.1: **Fragment picking schematic:** The query sequence fragment shown is a part of the metamorphic protein MAD2, and the cartoon representation of the 3D structures of proteins containing the fragments is shown in the circles. The PDB ID and the Alphafold Accession number are indicated in the representative hits.

The long query sequence is first broken into fragments using a sliding window of sequence size 7. Each fragment from the query sequence is searched for identical fragments across a database of protein sequences containing around 200,000 redundant proteins from Protein Data Bank Berman *et al.* (2000) and around 1.2 Million protein structure predictions from Alphafold2. In case a protein from PDB contains multiple chains, each chain's secondary structure is extracted and stored separately in the database. Given our fragment-picking database size, 7-mer fragments give the optimal number of hits while maintaining the secondary structure diversity of the hits. Once the hits are obtained, the secondary structure of the hits is extracted for the purpose of scoring. The hits are filtered to remove those fragments that contain at least one unobserved residue or a local pLDDT score of less than 70 in the case of experimentally solved structure and computationally predicted structure, respectively. The secondary structures of the proteins are assigned from the coordinates of the structures using DSSP Kabsch and Sander (1983). As we perform the fragment search on a redundant PDB database, many of the hits occurring from different protein structures may all be pertaining to the same protein sequence.

Hence, to eliminate the over-representation of proteins that have multiple redundant structures with the same sequence deposited in the PDB, we use a sequence clustering algorithm to cluster the database with a sequence identity threshold of 100%. The sequence clustering is performed using the fast and sensitive tool MMseqs2 Steinegger and Söding (2017). All sequence alignments with sequence identity scores of 100% and covering at least 80% (-c 0.80) of both sequences are grouped in one cluster. When performing the fragment picking for a query sequence, multiple fragments may be obtained from different protein structures within the same cluster; to account for that, when secondary structure probabilities are calculated, each hit from a cluster is weighted by the inverse of the total number of hits from the same cluster.

Figure 3.1 shows the fragment picking schematic for one 7-mer fragment `"RLRSFTT"` from MAD2 protein. The different structures containing the fragment sequence are depicted in the circles, and one structure from each cluster is shown. By analysing the secondary structure of the middle residue in each of the hits obtained, the middle residue propensity is calculated and depicted as a pie graph.

3.2.1 Trie-based Fragment Search

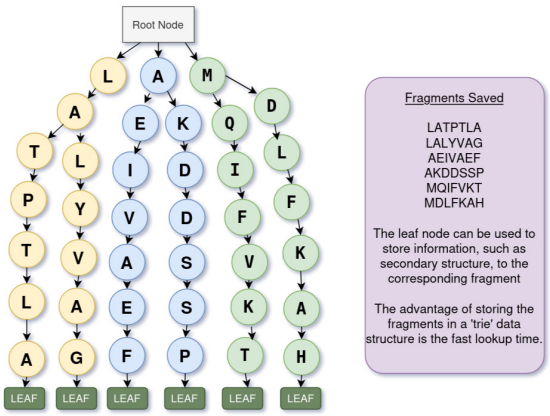

Figure 3.2: **Trie data-structure schematic:** Schematic representation of protein sequence fragment data stored in a trie with further information stored in the leaf node.

In order to perform fragment picking on a proteome, the fast lookup time of a trie data structure is leveraged. All the proteins in the proteome are broken into 7-mer fragments, and all the resulting 7-mers are stored in a trie. This trie now has all the possible fragments from the proteome that needs to be searched for. So now, all we have to do is iterate once through the database and whenever any fragment from the trie is encountered, append the structural information of that fragment in the leaf node of the trie corresponding to the same fragment. The schematic for the storage of protein sequence fragments in a trie data structure is shown in Figure 3.2.

To perform this, we have implemented the trie structure using the Python module 'pygtrie' developed by Google. Once an instance of the trie is created, the fragments from all the proteins in the proteome FASTA file are stored one by one in the trie. In our system, the trie-implemented search for a proteome with 20,000 proteins was completed within 2 hours.

3.2.2 Linear Search Implemented on JAVA

The fragment search is performed by iterating through the lines of the database with a fast linear search in Java, matching fragment sequences using Regular expression (REGEX). When run on our system (16-core AMD Ryzen 7 5700, 64 GB RAM), the search for one fragment across the database is completed at an average of 700 milliseconds. The schematic of linear search using a sliding window is illustrated in Figure 3.3.

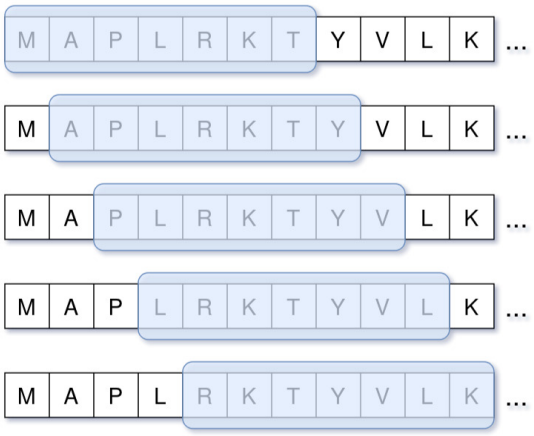

Figure 3.3: **Linear search schematic:** Schematic for a linear search using a sliding window over one sequence. Like this, all the sequences in the database are run over by a sliding window and searched for a particular sequence.

3.3 Training Data

The set of known fold-switching proteins and proteins that are highly likely to have a single fold is used as a training dataset. This dataset has been borrowed from existing literature Porter and Looger (2018); Chen *et al.* (2020). The dataset for fold-switching protein is initially curated by looking for existing protein structures, where the set is further filtered to remove structures flagged obsolete and structures that were later found not to be fold-switching due to erratum in structure. Additionally, the training dataset is expanded to include the designed metamorphic protein Solomon *et al.* (2023). This way, 189 fold-switching sequences and 196 monomorphic protein sequences are considered for the training dataset.

We have chosen to include fold-switching proteins for the construction of a metamorphic protein classifier for a couple of reasons: (a) There are only a handful of truly metamorphic proteins, and it is unlikely to train a model on such limited data meaningfully. (b) Our model predicts those metamorphic proteins where there is a large change in the secondary structure of the two conformations with high confidence. Hence, when predictions are made, analysing the highly confident hits could help in searching for potential metamorphic proteins.

The complete training data is presented in the appendix section A. The PDB IDs of the structures used for training and their chain IDs are provided.

3.4 Scoring Metrics

A particular fragment is scored based on the different secondary structures of the hits obtained from the database. We hypothesized that local interactions within the fragment of 7 residues play a major role in determining the secondary structure of the middle residue. Fragment-based ab initio structure prediction tools like RosettaGront *et al.* (2011) also choose to work with the secondary structure of the middle residue from the fragments that have been picked from their curated database for structure prediction. Guided by our hypothesis, we opted to analyse the secondary structure of the middle residue of the picked fragments. Therefore, the secondary structure of the middle residue is retrieved from each fragment hit obtained from our database. Since metamorphic proteins have the ability to interconvert between two distinct conformations, often having a change in secondary structure, the metric we choose should be able to quantify the ability of a fragment sequence to adopt distinct secondary structures. Multiple consecutive fragments displaying this ability would, in turn, aid the whole protein sequence to switch folds. Following this thought, we consolidated four different metrics: **(a)** Diversity Index, **(b)** Substitution Score, **(c)** Information Entropy, and **(d)** Uncertainty Score.

The diversity index has already been used for the classification of metamorphic proteins by Chen et al. Chen *et al.* (2020), we decided to implement the same metric in our model. The substitution-matrix-based scoring method is inspired by Porter and Looger's work Porter and Looger (2018), where it was initially used to elicit the difference in the secondary structure of highly similar protein sequences. Information entropy and uncertainty score are metrics that we have introduced for the purpose of classification of metamorphic proteins.
Diversity index and information entropy are defined as follows:

$$\mathrm{DI} = (h^2 + e^2 + l^2)^{-1} \tag{3.1}$$

$$\mathrm{Entropy} = -\,h\log(h) - e\log(e) - llog(l) \tag{3.2}$$

Where h, e and l are the calculated propensities of the middle residue to be in helix, sheet and loop respectively. The secondary structure propensities for each 7-mer are calculated from the hits obtained from the database by assessing what fraction of those hits have middle residues existing in helix, sheet and loop secondary structure.

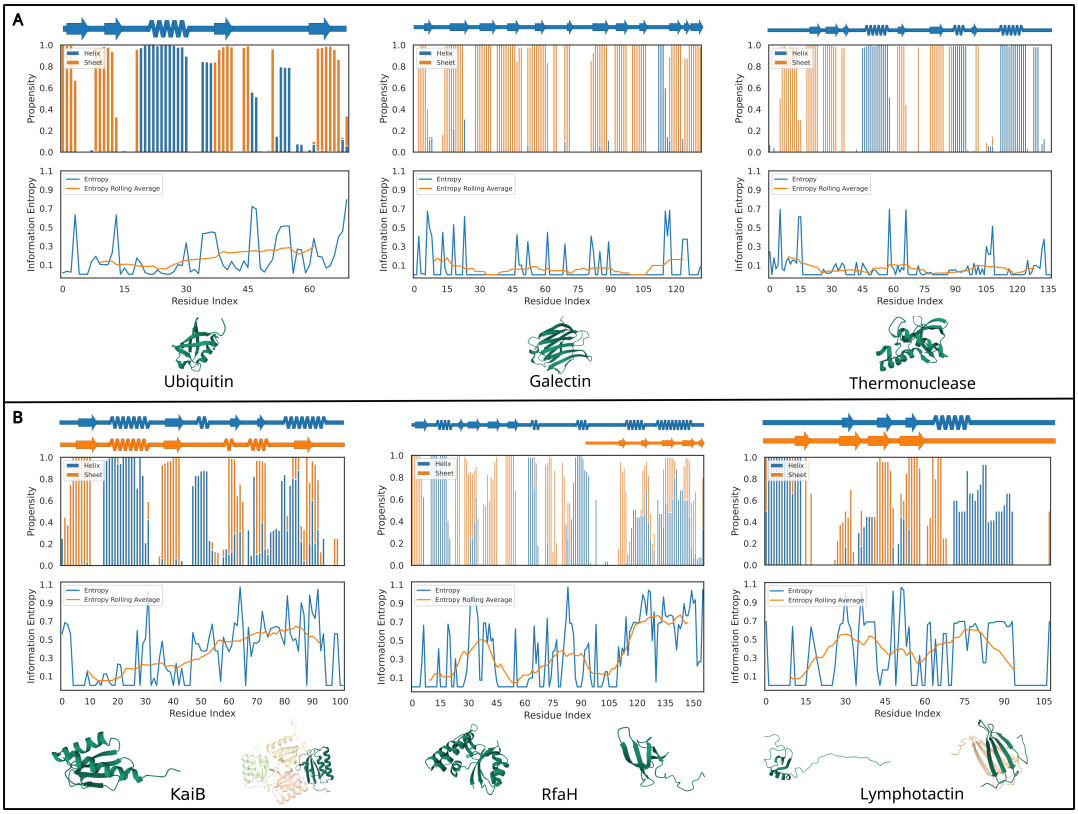

Figure 3.4: **Diversity metrics able to distinguish metamorphic proteins and capture regions of fold-switching:** The Secondary structure diagram annotated through an experimentally solved structure (top), secondary structure propensity plot (middle) calculated via fragment picking by Morpheus and entropy scores (bottom) are plotted for **(A)** Monomorphic proteins Ubiquitin, Galectin and Thermonuclease (PDB ID: 1UBQ, 2NN8 and 5KEE) **(B)** Experimentally known metamorphic proteins Kaib, RfaH and Lymphotactin (PDB ID: 5JYT and 2QKE, 2OUG and 2LCL, 1J8I and 2JP1). The experimentally solved structures are also depicted below each plot. The entropy score is plotted across residue number, the blue line plot is the entropy score for each fragment plotted corresponding to the residue number of the first residue, and the orange line plot is the rolling average of the entropy score with a window width of 18.

Hence, they are normalized and add up to 1. The diversity index and information entropy scores are calculated for all the fragments from their corresponding middle residues' propensities. The values that diversity and entropy scores attain across all possible input propensities is shown in Figure 3.5.

Figure 3.4 displays the diversity metrics for a few metamorphic and monomorphic proteins. Figure 3.4 shows that the region where the entropy score remains consistently high correlates well with the region of the protein that actually switches fold in the case of metamorphic proteins: KaiB, RfaH and Lymphotactin. As for the monomorphic proteins Ubiquitin, Galectin and Thermonuclease, the entropy scores remain low overall.

We noticed that when working with residues other than the middle residue in the

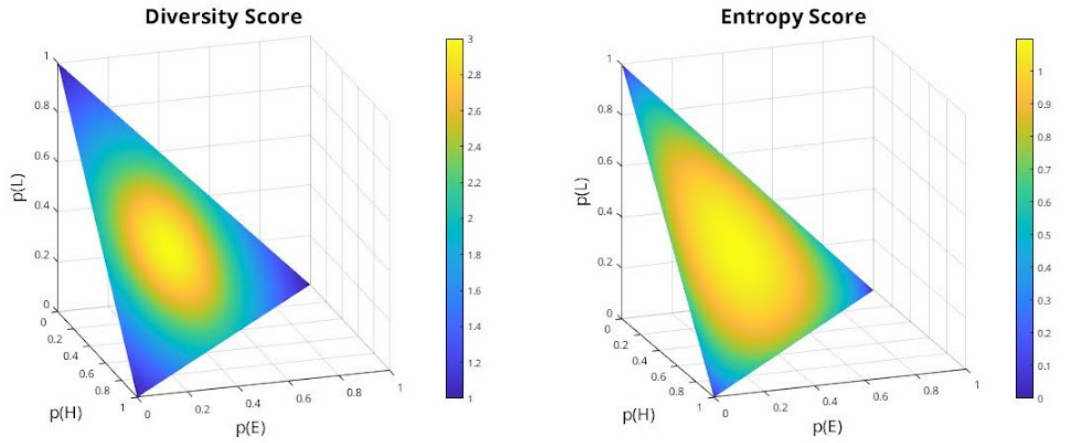

Figure 3.5: **Diversity vs Entropy score:** Diversity and Entropy score plotted as a function of the propensities of secondary structure p(H),p(E) and p(L). The triangle represents the region where the propensities are normalized, meaning they follow $p(H) + p(E) + p(L) = 1$.

7-mer, similar levels of accuracy were achieved. The accuracies when using all the different residues (except the terminal residues) were within one standard deviation of each other.(Table 3.2) Among these, the model working with the third residue gave the highest accuracy, higher than that of the middle residue. Moving forward, we decided to work with the third residue for the classification model.

In bioinformatics, substitution matrix-based scoring techniques are widely used. They are particularly helpful in quantifying the change or substitution of one element by another. We implemented a substitution matrix based scoring method in order to evaluate the ability of the fragments to adopt different secondary structure conformation. First, the secondary structure of the fragments in the original output from DSSP is taken. The original output from DSSP contains one of these 8 secondary structures: α-helix (H), 3_{10} helix (G), π helix (I), poly-proline helix (P), extended strand (E), β bridge (B), turn (T), bend (S). A reference hit is set as the hit that has the most commonly occurring 7-mer secondary structure pattern among all the hits. If there is no majority, a random hit is assigned as the reference hit. Then, every other hit is taken, pairwise aligned with the reference hit and scored according to the following substitution matrix:

$$\frac{1}{7}\begin{bmatrix} \text{ref/replace} & \text{H/I/G} & \text{E/B} & \text{T} \\ \text{H/I/G} & 0 & 0.7 & 0.3 \\ \text{E/B} & 0.7 & 0 & 1 \\ \text{T} & 0.3 & 1 & 0 \end{bmatrix}$$

This way, as we align each fragment sequence hit with the reference hit and apply the substitution matrix on the alignment, we get a score. For a sliding window fragment, all pairwise alignment with the reference hit is taken and among them, if there are multiple hits from the same sequence cluster, these scores are weighted with the inverse of the total number of hits that we obtain from the same sequence cluster. We then take the weighted average of these scores from all pairwise align-

ments and assign it to that sliding window fragment. In this manner, we obtain a score for each fragment corresponding to all sliding windows.

We also sought to quantify the confidence in the calculated diversity, entropy and substitution scores from fragment-picking for a particular sliding window based on the number of hits we obtained for the corresponding fragment sequence. Ideally, there exists a true value of the propensity for the middle residue to be in a helical or sheet-like structure; the propensity that we arrive at is an estimate that gets better and better based on the number of fragment hits from different sequence clusters that we have available to us to make an estimation. From Hoeffding's inequality Hoeffding (1994), we know that the difference in the true value and the estimation varies with the sample size in an exponential manner. Hence, the uncertainty in the calculated values of diversity metrics is estimated in the form of an exponential function. Given a fragment with a large number of hits (say N), we can analyse how the propensities calculated from 'n' randomly selected fragments vary from the true propensity (n $\rightarrow$ N). This analysis helps in modelling the functional form of uncertainty. The final form for uncertainty is estimated by the function:

$$\text{Uncertainty} = \exp\left(-0.5 \times \#\text{ of hits from unique clusters}\right) \tag{3.3}$$

Rolling average over a range for functions that are irregular and discretely defined gives an output that is a smoothened representation of the initial function. They accumulate and average the variation of the discrete function across a window. We theorized that a protein sequence having a contiguous region of high entropy scores would imply that the region as a whole would have a higher chance of switching folds. Hence we chose to make a rolling average of the scores to elicit these contiguous regions. So, the score that is taken as input for classification is the maximum of the rolling average of the diversity, entropy and substitution scores. The rolling window width is a parameter for which we are free to choose a value. An optimal value for the window width is chosen using the training data. It is found that a window width of 8 and 18 gives the best separation of metamorphic and monomorphic scores in the training data (SI: table S3). Hence, a step function, as shown below, is used for the rolling average window:

$$\text{Window width (WW)} = \begin{cases} 8 & \text{len(protein sequence)} \leq 25 \\ 18 & \text{len(protein sequence)} > 25 \end{cases} \tag{3.4}$$

The final score that is used as the input vector for the classification model is all of the aforementioned scores together. After the rolling average window width is decided based on the length of the protein sequence, the maximum value of the rolling average of diversity, entropy and substitution scores is calculated. The mean value of uncertainty score across all residues is also found and the four scores together is used as the feature space for the classification model as shown below:

$$\text{Final Score} = \begin{pmatrix} max\left(\frac{1}{WW}\sum_{j=0}^{j<WW}\text{Diversity}_{i+j}\right) \\ max\left(\frac{1}{WW}\sum_{j=0}^{j<WW}\text{Entropy}_{i+j}\right) \\ max\left(\frac{1}{WW}\sum_{j=0}^{j<WW}\text{Substitution}_{i+j}\right) \\ mean\,(\text{uncertainty}) \end{pmatrix} \tag{3.5}$$

In the formula for final score, i ranges from 3 to L-7+3 (the first two are skipped when the 3rd residue of the first sliding window is used) where L is the length of the protein sequence. In the case that for a particular fragment, no hits are found, we set the value of the diversity metric scores to minimum. Therefore, diversity is set to 1, entropy and substitution are set to 0. Uncertainty in that case achieves the maximum value of 1.5. This is done to reduce the false-positivity rate of the model in the case where we have no data to fix the propensity of the 3rd residue of the fragment and therefore the diversity of the fragment.

3.4.1 Optimizing The Scoring Metrics

3.4.1.1 Rolling window width

The Rolling average window width is optimized by considering a range of values from 5 to 25 and picking the one that gives the best Mattew's Correlation coefficient (MCC) when applied to the training dataset. A six-fold cross-validation scheme is applied when the MCC value is calculated with each of the window widths. The validation accuracy and MCC score are calculated, and the window width that gives the best MCC score is selected. Table 3.1 shows the different accuracy values attained when the window width is varied from 5 to 25.

3.4.1.2 Substitution Matrix

Similarly to optimizing the window width, the substitution matrix is optimized using a range of parameters the form shown below:

$$\begin{bmatrix} \text{ref/replace} & \text{H/I/G} & \text{E/B} & \text{L/C/S} & \text{T} \\ \text{H/I/G} & 0 & a & b & c \\ \text{E/B} & a & 0 & b & d \\ \text{L/C/S} & b & b & 0 & 0 \\ \text{T} & c & d & 0 & 0 \end{bmatrix}$$

The values for a,b,c and d were sampled first in a coarse manner, where the values of all the 4 parameters range from 0 to 1 with a step of 0.2 for each one. Then once a substitution matrix that gave the maximum MCC value for the substitution score is found, the sampling is made much finer with a step value of 0.05 for each of the

Window width	Accuracy	MCC
5	77.2	0.52
6	77.8	0.556
7	79.1	0.584
8	81.4	0.63
9	78.5	0.572
10	79.6	0.593
11	78.8	0.582
12	80.8	0.622
13	82.1	0.65
14	82.6	0.657
15	83.9	0.682
16	83.6	0.677
17	83.6	0.675
18	84.4	0.692
19	83.6	0.676
20	83.4	0.671
21	82.9	0.663
22	81.8	0.643
23	82.8	0.661
24	81.5	0.636
25	82.3	0.652
26	82.3	0.648
27	83.1	0.665

Table 3.1: The Accuracy and MCC measures are provided for each of the value of averaging window width ranging from 5 to 27. These measures are calculated by classifying the data using a quadratic SVM and performing a 6-fold cross-validation on the training dataset.

parameter. That way we arrived at a final substitution matrix as shown (d=0):

$$\frac{1}{7}\begin{bmatrix} \text{ref/replace} & \text{H/I/G} & \text{E/B} & \text{T} \\ \text{H/I/G} & 0 & 0.7 & 0.3 \\ \text{E/B} & 0.7 & 0 & 1 \\ \text{T} & 0.3 & 1 & 0 \end{bmatrix}$$

3.4.1.3 Fragment Size

The fragment size is optimized via trial and error. For the training dataset. Fragment picking was performed on the training dataset with the fragment sizes: 6, 7 and 8. When fragment picking is performed with a fragment size of 8, the number of hits, i.e. proteins containing the identical fragment, was quite low with the database currently in use. While this does not give us enough data to work with, using fragment size 6 incurred the problem of the fragment secondary structure diversity, being all over the place. The 6-mers did not seem to truly represent the diversity taken by the protein in question. Therefore the fragment size is set to

7 and the high accuracy in prediction seems to validate the choice of picking 7-mers.

3.4.1.4 Middle Residue

After fragment picking, the residue that was taken into consideration for the secondary structure analysis is the middle residue (4th residue of the 7-mer fragments). But upon considering other residues, there was a small change in accuracy measures. Although the accuracy measures when whichever residue is taken into consideration did not vary beyond a significant number standard deviation, we still chose to work with the one that gave the highest MCC score. Table 3.2 shows the variation of accuracy when the fragment in consideration is changed from the 2nd residue to the 6th residue of the 7-mer. Accordingly, we chose the final model to work with the 3rd residue.

Measure	2nd RSD	3rd RSD	4th RSD	5th RSD	6th RSD
TPR	0.82 (0.08)	0.788 (0.08)	0.777 (0.05)	0.874 (0.04)	0.787 (0.02)
SPC	0.852 (0.06)	0.904 (0.06)	0.863 (0.07)	0.737 (0.12)	0.886 (0.05)
PRE	0.844 (0.05)	0.887 (0.08)	0.845 (0.07)	0.771 (0.08)	0.872 (0.05)
ACC	83.42 (4.3%)	84.75 (6.0%)	81.87 (4.4%)	80.55 (6.3%)	83.93 (2.7%)
F1	0.829 (0.04)	0.834 (0.08)	0.808 (0.04)	0.818 (0.05)	0.827 (0.02)
MCC	0.6727 (0.08)	0.6983 (0.12)	0.6389 (0.09)	0.6185 (0.12)	0.678 (0.05)

Table 3.2: The Accuracy measures for the residue in consideration starting from the 2nd residue to the 6th residue of the 7-mer fragment.

3.4.1.5 Support Machine Vector Model

An SVM model with a quadratic kernel is chosen for making predictions on data the training model has not encountered before. Support vector machine is chosen as they achieve good performance on many classification tasks. Kernels make SVMs more flexible and able to handle nonlinear problems, and here as we are using a quadratic kernel, we do not have to worry about the model complexity. As we have a relatively small dataset to train with, this restricts the choices of many machine learning models we could use. A quadratic SVM model gives us good accuracy while still maintaining a low model complexity.
The hyperparameters box constraint and Kernel scale were optimized using a Bayesian optimization technique in MATLAB.

3.5 Prediction Model

A final prediction model is constructed using the aforementioned features. The training data is randomly split according to a 6-fold cross-validation scheme. Upon performing a 6-fold cross validation, an SVM model with a polynomial kernel of degree 2 is found to be the model that has highest average validation accuracy

while also maintaining a lower model complexity. As a result, a quadratic SVM is selected to provide predictions on data that the model has never encountered before. Figure 3.6 shows the decision boundary that is finally obtained after optimizing the support vector machine using a Gaussian optimization scheme for the SVM hyper-parameters.

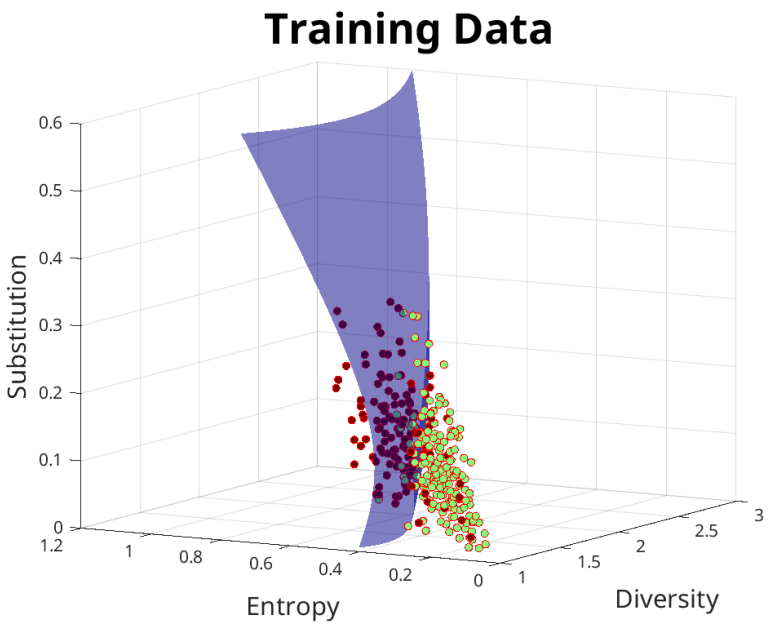

Figure 3.6: **The training dataset with 3/4 features:** The training dataset of known fold-switching proteins and highly likely monomorphic proteins are plotted in the feature space of the variables diversity score, entropy score and substitution score. The points in red depict fold-switching proteins, and the points in green depict monomorphic proteins. The decision boundary shows the optimized SVM model output when a quadratic kernel is used.

Although Figure 3.6 shows a 3D hyper-surface as the decision boundary, the actual decision boundary is a 4D hyper-surface corresponding to the feature space of the 4 scores used. The figure here is plotted for visualization purposes and to show the segregation of metamorphic and monomorphic proteins in the 3 feature space dimension. The 4 features are plotted in a parallel coordinates plot to see the mutual variation in all 4 coordinates with reference to each other (see Figure A.7). The final prediction model, uses all 4 of the features to make new predictions. Table 3.3 shows the diversity metrics for the known and experimentally verified metamorphic proteins and a few monomorphic proteins.

Figures A.1 - A.7 show how the decision boundary varies as the model is re-trained across each of the 6 partitions. The corresponding validation confusion matrix is also shown. The standard deviations shown in table 4.1 also gives an idea of how much variation is present across the different folds used for training. Figure S3.g shows the ROC (Receiver-operating characteristic) curve for the final optimized

Protein	Diversity	Entropy	Substitution	Uncertainty
Metamorphic				
RfaH	2.08	0.78	0.09	0.26
XCL1	1.78	0.56	0.02	0.46
KaiB	1.73	0.63	0.25	0.09
MAD2	1.66	0.55	0.17	0.05
Designed protein	1.58	0.44	0.16	0.32
HIV-RT1	1.53	0.43	0.13	0.02
IscU	1.47	0.36	0.13	0.74
MinE	1.47	0.43	0.09	0.20
Selecase	1.45	0.40	0.15	0.26
CLIC1	1.37	0.27	0.18	0.04
Monomorphic				
Luffaculin	1.53	0.42	0.08	0.41
Histone H3	1.50	0.51	0.048	0.00
beta-lactamase	1.44	0.35	0.17	0.01
GFP	1.39	0.34	0.15	0.02
Chymotrypsinogen	1.38	0.33	0.21	0.14
Ribonuclease	1.29	0.22	0.092	0.55
Trypsin	1.27	0.26	0.08	0.01
Ubiquitin	1.26	0.28	0.09	0.00
Caspase-3	1.10	0.11	0.05	0.01

Table 3.3: The diversity metrics of known metamorphic proteins and proteins that are highly likely to be monomorphic are ordered according to decreasing diversity scores.

SVM model.

3.6 Evaluation of Performance

The model performance is evaluated using different measures of accuracy. These measures include Accuracy (ACC), Sensitivity (Sn), Specificity (Sp), Precision, Matthew's Correlation coefficient (MCC) and F1-score which are commonly used measures for machine learning algorithms. They are defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP}$$

$$\text{Sn} = \frac{TP}{TP + FN}$$

$$\text{Sp} = \frac{TN}{TN + FP}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{F1} = \frac{2TP}{2TP + FP + FN}$$

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

where TP, FP, TN, and FN denote the number of true positives, false positives, true negatives, and false negatives, respectively. Different classification models such as

Support Vector Machines (SVM), k-nearest neighbour (KNN), Linear Discriminant Analysis (LDA), Decision Tree, etc. are evaluated on the training data. Finally one is picked to make predictions on data the model has not encountered while training. This decision is based on validation accuracy measures and lower model complexity. For example, a fine Gaussian SVM although gives high validation accuracy, the model complexity is quite high due to the fine nature of the kernel. Hence, to reduce chances of overfitting, a model with lower complexity is preferred. An SVM model with a polynomial kernel of the second degree is selected as the prediction model of the classification models that are evaluated.

3.7 Work Flow of the Algorithm

The workflow of the algorithm is depicted in the schematic figure 3.7.

The flow of information starts from the input query sequence, which is then used to perform fragment sequence search on the curated database. The database is initially clustered using a sequence clustering algorithm. This is to reduce over weighting redundant sequences in the database. The results of fragment picking is then scored using 4 different scoring methods: Diversity score, Entropy score, Substitution score and uncertainty. These four scores are fed into the SVM model. A binary output is produced indicating whether the prediction for the query protein is metamorphic or monomorphic.

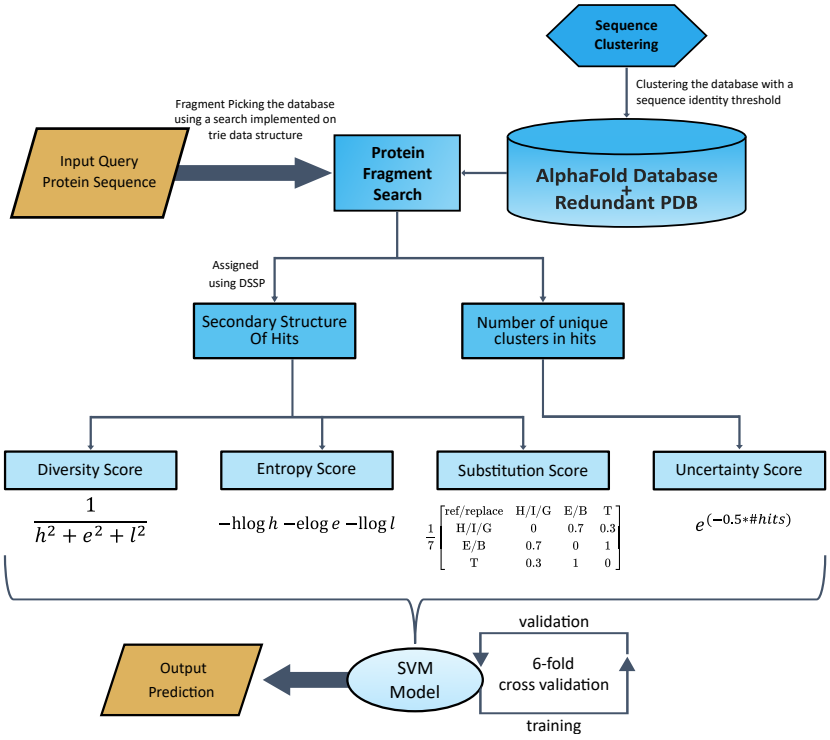

Figure 3.7: **workflow diagram:** Workflow is depicted starting with the fragment sequence search across the database. The hits are then scored for diversity of secondary structures shown across different structures through 4 different scoring techniques. An SVM model with a quadratic kernel is employed with the scores as the features. The SVM model is trained and a 6-fold cross-validation scheme is applied. Finally a binary output prediction is made using the trained SVM model.

Chapter 4

Results

Manuscript Submitted Online: The research work done as part of my MS thesis "Morpheus:A fragment-based algorithm to predict metamorphic behaviour in proteins across proteomes" is submitted online to the preprint server bioRxiv.org. The majority of the work described here is publically available with the Digital Object Identifier (DOI): 10.1101/2025.02.13.637956.

Having made an algorithm to predict the metamorphic behaviour in proteins using only protein sequence information, we used Morpheus to predict metamorphic proteins across proteins from 57 different proteomes. We also choose a few candidate metamorphic proteins from our predictions that are highly likely to switch folds based on a few criteria that are mentioned further on in the following subsections. In section 5.1 we select one of these candidate proteins, which was predicted to be metamorphic by Morpheus, and analyse it further to find any evidence of metamorphic behaviour reported. We also performed a TSNE based clustering of the names of the proteins predicted to be metamorphic in order to visualize them.

4.1 Accurately Predicting Metamorphic Proteins

Our Model can predict metamorphic proteins with an average cross-validation accuracy of 84.75% and Matthew's correlation coefficient of 0.698. Our approach surpasses existing sequence-based metamorphic protein predictors while achieving a low false-positive rate of 8.63%. These measures of accuracy present an exciting opportunity of predicting metamorphic proteins across different proteomes and comparing them. Table 4.1 depicts the varied measures of accuracy for the SVM model with the quadratic kernel used in Morpheus, and existing sequence-based predictor of metamorphic proteins.Chen *et al.* (2020)

Additionally, Plotting the diversity metrics and the residue-wise secondary structure propensities allows the user to estimate the approximate region that undergoes the fold switch. For example, in figure 3.4, residues around the C-terminal domain of the known metamorphic protein RfaH show a contiguous region with a high entropy score. These residues overlap well with the experimentally shown region of

Measure	Morpheus	Chen et. al
Sensitivity	0.78 (0.08)	0.41 (0.02)
Specificity	0.91 (0.03)	0.92 (0.02)
Precision	0.89 (0.03)	0.88
Accuracy	85.0% (3.8)	63.3% (1.6)
F1 Score	0.83 (0.05)	0.56
MCC	0.699 (0.07)	0.327 (0.12)

Table 4.1: The validation accuracy measures for Morpheus and the method by Chen and co-workers averaged across the cross-validation partitions are mentioned above. The numbers in the parentheses represent the standard deviation for these values calculated across the cross-validation partitions. The values for Chen et. al. directly taken from the values mentioned in the publication; precision and F1-score were inferred from the other values, hence we could not provide a standard deviation value.

fold-switching, which is the entire C-terminal domain of the protein. When performing further experiments on candidate proteins, this information might prove to be useful. Additionally, the nature of such fold-switch that may happen in the putative hits (such as $\alpha-$helix to $\beta-$sheet transition) can also be inferred by looking at the secondary structure propensity plot of the candidate protein sequence.

One way to test the goodness of a classification model is to plot the 'Receiver-operating characteristic' (ROC) curve. The ROC curve is drawn by calculating the true positive rate (TPR) and false positive rate (FPR) at every possible threshold (in practice, at selected intervals), then graphing TPR over FPR. Figure 4.1 shows the ROC curve for Morpheus. The 'Area under the curve' 'AUC' value for morpheus is 0.908. AUC values for any classification model always fall between 0.5 and 1. A model with an AUC in between 0.8 and 0.9 is indicative of a good discriminatory power between the classes.
The χ^2 feature importance score is also shown in Figure 4.1 for the four features in Morpheus.

4.2 Proteome-level Predictions reveal wide-spread varying levels of Metamorphicity

Now that we have an algorithm that is fast and able to accurately predict metamorphic proteins, we sought to make predictions on proteins from different organisms. Specifically, we use UNIPROT database of all (one sequence per gene) proteins from a particular organism to run on Morpheus. Currently, We have run our algorithm on 57 different proteomes which includes a consolidated database of viral proteins. The 57 organisms are chosen from all kingdoms of life, from multicellular eukaryotes to unicellular prokaryotes. Analyzing different proteomes offers distinct advantages. For instance, studying an archaeal species would be beneficial, as its smaller proteome size reduces complexity and eliminates concerns related to post-translational modifications. Further, a few archaen species live in extreme environments, such

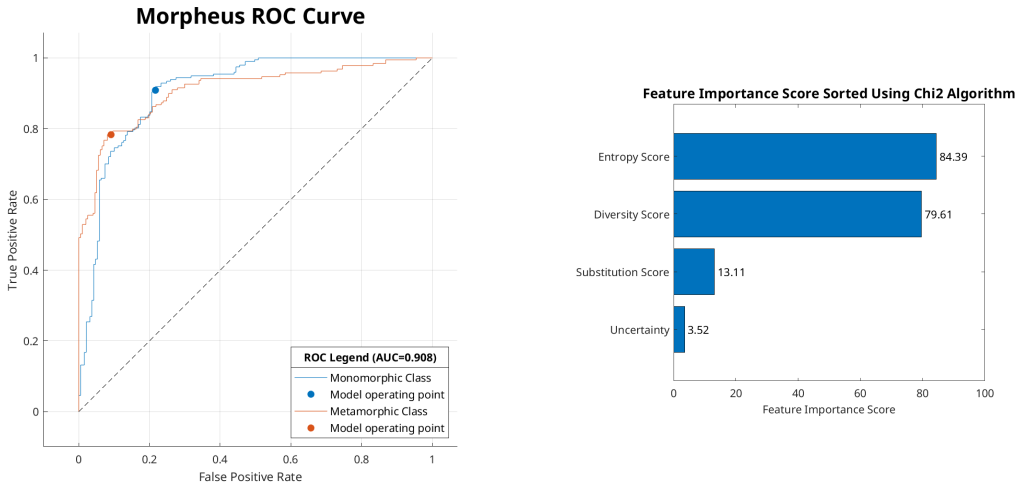

Figure 4.1: **ROC Curve and χ^2 scores:** (left) ROC curve (Receiver-operating characteristic curve) for the final classifier model of Morpheus for both the classes and (b) The χ^2 scores for the four features diversity, entropy, substitution score and uncertainty is shown with the values mentioned across each bar. The ROC curve is plotted considering the metamorphic protein class as the positive class. The area under the curve (AUC) for ROC curve is 0.908.

as very high temperatures or salinity, so the proteins involved in those organisms would have evolved to withstand such harsh conditions. We further noticed in the training dataset, viral proteins are more frequently populated among metamorphic proteins. Porter and Looger (2018) Hence we decided to include predictions on a consolidated database of viral proteins.All the data from running our algorithms on the 57 proteomes are presented here. Furthermore, a web-server has been set up to execute the algorithm for proteins that are not included in our local runs of the proteins from 57 proteomes. (http://mbu.iisc.ac.in/~anand/morpheus)

Besides fulfilling a researcher's curiosity, analysing the human proteome for metamorphic proteins can stand a chance at helping global healthcare and improving therapeutics by furthering the understanding of protein fold switching and the different physiological triggers involved within. Research has already shown fold-switching proteins associated with a number of diseases such as Malaria Jain *et al.* (2014), Gastric cancer Li *et al.* (2018), SARS-Cov 2 Costello *et al.* (2022) and more Lei and Takahama (2012). Expanding the reservoir of predicted metamorphic proteins help in paving ways for new research and may help in improving theraputics.

Figure 4.2 shows the diversity metrics for all the proteins from the human proteome. The decision boundary helps us make binary predictions of the metamorphic behaviour once we calculate the diversity metrics of these proteins. Further, based on the distance from the decision boudnary a confidence score can be assigned to each protein, indicating how confident we can be about the prediction. We also notice that of the proteins that are predicted metamorphic by our algorithm, a strong proportion of them are intrinsically disordered proteins (IDP). This can be rationalized

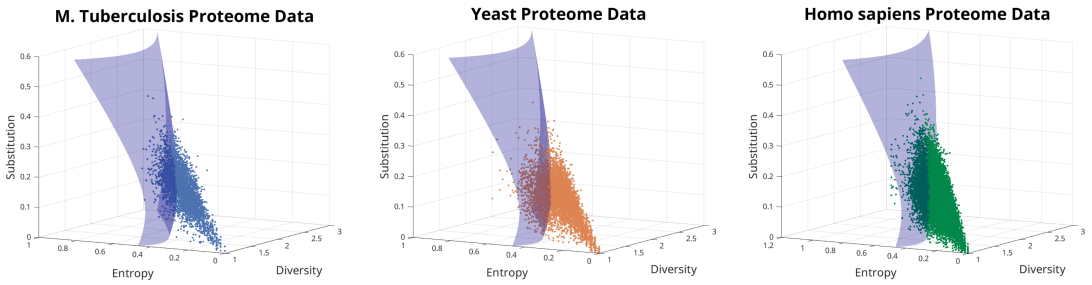

Figure 4.2: **Distribution of proteins from 3 proteomes on the feature space overlayed with the decision boundary enabling new predictions of metamorphic proteins.** A scatter plot visualizing 3,995, 6,060, and 20,579 proteins from the tuberculosis, yeast, and human proteomes, respectively, within the feature space defined by diversity score, entropy score, and substitution score. The SVM decision boundary is also depicted. Our algorithm predicts 1,050 (26.2%) tuberculosis proteins, 1,346 (22.2%) yeast proteins, and 4,696 (22.8%) human proteins as metamorphic of the entire proteome.

as our algorithm quantifies the diversity in secondary structures adopted by a particular fragment in the query protein with the help of a structural database.IDPs that exhibit highly varying secondary structures across different deposited structures, both experimentally and computationally sovled, are often reported as containing a high diversity of secondary structure. It is a natural consequence of the fragment-picking algorithm. When analyzing the final prediction set, each protein can be annotated with the percentage of disordered residues present in the sequence using existing protein disorder prediction tools such as IUPRED and PONDR. Given the high accuracy of many of these tools, employing a consensus approach from multiple predictions can further refine the dataset. This will ensuring a more reliable selection of metamorphic proteins among those that are predicted by our algorithm.

Figure 4.3 shows what fraction of the proteomes are predicted metamorphic by our algorithm. Intrinsically disorderd proteins are also included in the calculation of the histogram plot. To get an idead of the distribution of percentage disorder among those predicted metamorphic, the bars are also plotted with a colour intensity gradient. This gradient depicts the distribution of disorder among the proteins that are predicted to be metamorphic, with full intensity showing the completely ordered proteins and zero intensity showing completely disordered proteins. As the field of metamorphic proteins is still in its infancy and there is significant amount to understand regarding their fold-switching mechanisms; the features that exactly define a metamorphic protein and distinguish them from other conformational changes are still shaping up with ongoing research. Therefore, we decided to retain those proteins that are predicted to be intrinsically disordered in our final prediction set. We believe the conformational plasticity in proteins is a spectrum ranging from entirely monomorphic proteins on one end, which show the lowest conformational heterogeneity; to intrinsically disordered proteins, which show the highest conformational heterogeneity and lack a stable secondary structure most of the time. In the middle

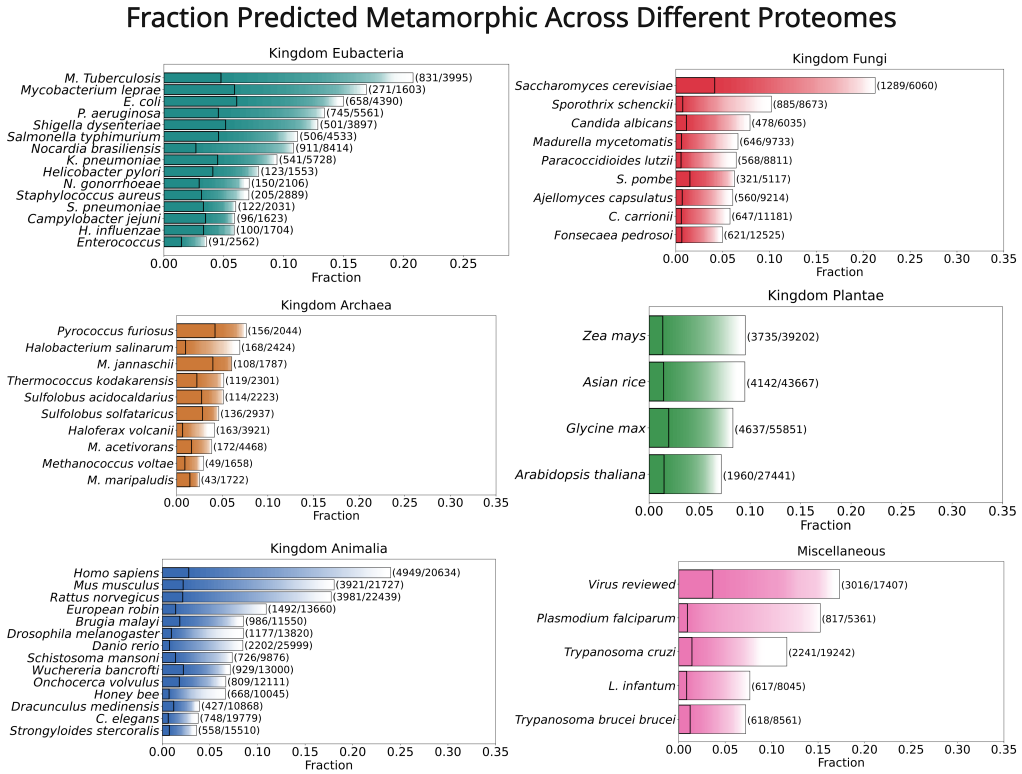

Figure 4.3: **The distribution of metamorphic proteins predicted by Morpheus across different proteomes:** The fraction of proteins predicted to be metamorphic by Morpheus across proteomes of various species. The distribution of disordered proteins among these predicted metamorphic proteins is depicted using a color intensity gradient, where $\alpha = 1$ (full intensity) represents completely ordered proteins and $\alpha = 0$ (white) represents those proteins that are fully disordered. Disorder prediction is performed for each identified metamorphic protein, and the fraction of these residues which are disordered is calculated. A vertical black line indicates the fraction of proteins with $\leq 1\%$ of residues predicted to be disordered.The figure is divided into subplots corresponding to different kingdoms of life. An additional subplot, labeled "Miscellaneous," includes viral species (from a curated database of manually reviewed viral proteomes) along with species from other kingdoms such as Excavata and Protista.

lies a set of proteins, including metamorphic, fold-switching, molten globules and multi-domain proteins with disordered regions. The distinction between metamorphic, fold-switching and proteins with disordered regions is not a clear distinction, yet. Hence, discarding proteins containing regions of disorder from the prediction set would be introducing bias and disregarding the continuum of conformational heterogeneity.

After having made the prediction across different proteomes, we can analyse the results. Refer section 4.4 for a visual representation of the fraction of proteins that are predicted to be metamorphic in the human proteome. The set of protein sequences which are predicted to be metamorphic with a high confidence is a promising source for finding potential candidate protein sequences on which experimental validation

can be performed. In the following section, we provide a filtered set of proteins and a few criteria to obtain these candidate proteins.

4.3 Selected Few Candidate Predictions Demanding Further Research

In the search for strong candidate metamorphic proteins, we had the following criteria in mind to filter the prediction dataset: **(a)** The protein sequence should be predicted metamorphic with a good confidence according to Morpheus **(b)** the protein should have a range of residues where the helix and sheet propensity values are both non-zero for a contiguous region indicating the potential ability to fold-switch **(c)** less than 10% of the residues should be predicted to be disordered and not in the region of fold-switch predicted by Morpheus. The threshold 10% is placed to account for the disorder region at the C or N terminus, if present. IUPRED2A is used to perform the disorder prediction Mészáros *et al.* (2018). Based on these criteria in mind, we present a few of these proteins as candidates for experimental evaluation.

Uniprot ID	Organism	Protein Name	Length	Confidence	Uncertainty
O14625	Homo Sapiens	CXCL11	94	6.48	0.52
Q9Y2Y1	Homo Sapiens	DNA-directed RNA polymerase III subunit RPC10	108	4.83	0.18
Q4CMH7*	Trypanosoma cruzi	Guanine nucleotide-binding protein subunit beta-like protein	102	4.80	0.44
P26995	P. aeruginosa	Transcriptional anti-antiactivator ExsC	145	2.82	0.43
Q9UI95	Homo Sapiens	Mitotic spindle assembly checkpoint protein MAD2B	211	2.47	0.01
A0A1D6P714*	Maize	Extensin-like protein	129	2.27	0.61
A0A804PFY2*	Maize	Photolyase/cryptochrome alpha/beta domain-containing protein	162	1.86	0.45
A0A3P7DY23*	Wuchereria bancrofti	S1 motif domain-containing protein	258	1.23	0.39
Q4DL46*	Trypanosoma cruzi	Copper-transporting ATPase-like protein, putative	263	1.03	0.25
B8A185*	Maize	Secreted protein	112	0.72	0.34

Table 4.2: List of potential candidates for metamorphic proteins. 6 of the 10 protein sequences marked with * do not have any experimentally solved structures deposited in the PDB as of when this manuscript was written.

Table 4.2 lists a few proteins that are predicted to be metamorphic and are good candidates to be considered for further experimentation. Two protein sequences contained in the list, CXCL11 and MAD2B, are related to already known metamorphic proteins that are lymphotactin and MAD2A. The sequences of CXCL11 and MAD2B are not similar to their metamorphic counterparts. CXCL11 already has two solved structures deposited in the PDB solved via NMR and Electron microscopy. Although the two structures exhibit different secondary structures, fold-switching has not yet been investigated. The NMR structure provides evidence suggesting that CXCL11 may be a metamorphic protein, with temperature acting as the trigger.Booth *et al.* (2004)
So far, only one protein that is part of the circadian rhythm, KaiB, has been experimentally characterized to be metamorphic in nature Chang *et al.* (2015). Since circadian rhythm is periodic and helps the organism in synchronizing with the time of the day, it would be quite a curious find, if a new metamorphic protein joins

the circadian rhythm. In our dataset of predictions, one of the promising candidated is part of the Maize proteome. The protein called 'Photolyase/cryptochrome alpha/beta domain-containing protein'is involved in circadian rhythm and is predicted to be metamorphic by Morpheus. In plants, cryptochrome acts as a blue light receptor to entrain circadian rhythms. They also mediate a variety of light responses, such as the regulation of flowering and seedling growth. Mei and Dvornyk (2015). However, there are no solved structures for this particular protein in maize. Hence, this protein is a great avenue to conduct further research and experimentation.

4.4 Vizualizing The Protein Prediction From the Human Proteome

To vizualize the different proteins that were predicted metamorphic in the human proteome, the names of the proteins predicted metamorphic needed to be represented in a plot where the more distant the two proteins were, the more dissimilar were the names of the proteins. To accomplish this, we used an embedding to vectorize the names of the proteins that are predicted metamorphic. There are a lot of word vector embeddings that are used in machine learning, one of the embeddings is chosen and implemented on python. At this stage, the name of each protein sequence is extracted from UNIPROT website and the word embedding is used to convert the name, to a vector. Further, a T-SNE distance based clustering is performed to visualize the proteins in a plot. In Figure 4.4, each blue dot represents a single protein sequence and the coordinates are based on T-SNE algorithm and have no direct inference. We then notice that many small clusters of proteins/datapoints are visible, so to get an idea of the proteins that are clustered, we manually annotated the names that were common within the smaller clusters. Figure 4.4 shows the final annotated T-SNE clustering plot.

Here we can see that proteins from a lot of different functional aspects of the cell are predicted to be metamorphic. There are cell-membrane proteins, cystosolic proteins, voltage-gated ion channels, transcription factors, ribosome-related proteins, etc. A few of the common recurring themes are elaborated below.

1. **Motor Proteins:** Of the proteins that are predicted to be metamorphic in the human proteome, one large cluster that can be seen in the T-SNE plot are Kinase, dynamin and dynein related proteins. These proteins undergo a lot of changes when their functions are performed. Hence it is only rational that when the structures of these proteins are solved, they are found in different conformations. Fragment search in Morpheus is able to pick these changes and predicts them as metamorphic.

2. **Voltage gated ion channels:** These proteins are channel proteins that regulate the movement of ions through a membrane. When sufficient conditions

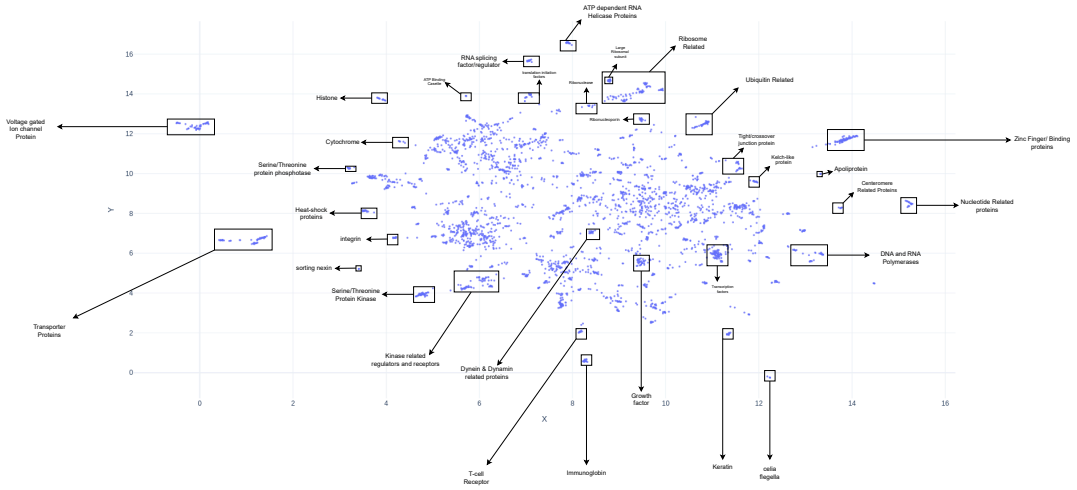

Figure 4.4: **Vizualizing prediction from Human proteome:** The names of the proteins that are predicted to be metamorphic in the human proteome are vectorized using a text embedding and then they are clustered using a T-SNE algorithm to visualize the closeness or similarity among the proteins

are met, the protein undergoes a conformational change that allows the passage of ions through the membrane. The resulting changes are also detected by Morpheus.

3. **disordered proteins** A large fraction of the proteins that are predicted to be metamorphic by Morpheus are partially or completely disorderd in nature. Disordered proteins appear either as missing density in crystallography or are captured by NMR structure solving methods. Sometimes, intrinsically disordered regions conformation are also predicted with higher confidence score by Alphafold. These might be the reasons why Morpheus detects those IDPs to have higher diversity and hence classify them as metamorphic proteins.

As we have only trained our model on proteins that are either fold-switching or having a highly stable single fold (monomorphic), Morpheus is not able to distinguish between other classes of proteins where the protein undergoes an order $\leftrightarrow$ disorder transition or undergoes a hinge like motion where there is no significant changes in secondary structure of the protein. These kinds of proteins need to be filtered out before our dataset can be used for further research. Intrinsically disordered proteins can be filtered out through the use of accurate disorder predicting tools such as PONDR or IUPRED.

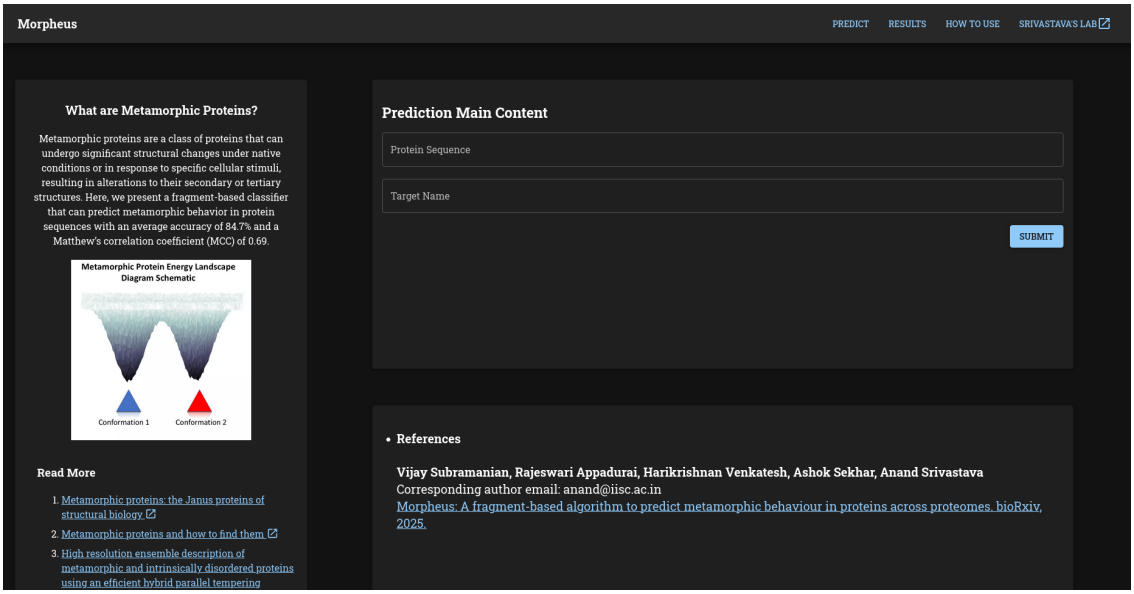

Figure 4.5: **Website:** Screenshot of the prediction page for Morpheus webserver. A query protein sequence can be given as an input and the prediction along with diversity metrics would be given as the output by the server.

4.5 Web-server Implemented For Morpheus

To allow easy access to the Morpheus Metamorphic Protein Classifier, we have made it available as a web application (https://mbu.iisc.ac.in/ anand/morpheus/). The application was made with the Next.js framework for the front end and the Spring boot framework for the back end. The core of the application consists of a search application that takes in a FASTA protein sequence and performs a REGEX-based search through a curated database of protein sequence and structure information. The application searches for identical protein sequence fragments of 7 residues within the database and returns the secondary structure information and pLDDT scores if applicable. We have adapted code from RCSB Advanced search sequence motif search. The frontend application allows the user to submit FASTA sequences, and it returns a yes/no classification along with a diagram that shows details like diversity score and entropy score for each residue.
Implementing a web application for Morpheus increases its reach and makes it easier for users to find out if their query protein has any metamorphic characteristics. Each job takes 2-3 minutes to run and scales linearly with the protein sequence length. The server handles multiple requests at the same time by executing jobs in parallel on the backend of the server. If, in the future, increased loads are received on the server, a public queue for job processing can be constructed.

Chapter 5

Conclusion

5.1 Examining CXCL11 - A Potential Candidate

CXCL11 (also known as ITAC) is one of three chemokines that binds to the CXCR3 receptor, with the highest affinity and strongest agonist activity amongst them. The other two chemokines being IP-10 and MiG, while CXCR3 is a G-Protein coupled receptor. ITAC's solution structure, determined using NMR spectroscopy, reveals a canonical chemokine fold but greater conformational flexibility compared to related chemokines. Booth *et al.* (2004) This conformational heterogeneity seems to be related to the temperature at which the NMR HSQC spectra is determined. This fact further gives us reason to believe that CXCL11 could be a metamorphic protein. Unlike the human Lymphotactin, ITAC does not dimerize at millimolar concentrations due to a β-bulge in β-strand 1, which disrupts the typical dimerization interface. The human Lymphotactin undergoes a conformational switch through dimerization with temperature being the trigger. So it appears that the conformational switch in CXCL11, if present, would be quite different from its known metamorphic relative Lymphotactin.

The NMR studies conducted by the authors on ITAC initially faced challenges due to spectral overlap and conformational heterogeneity. Refer to Figure 5.1 (left, taken directly from the source cited) for HSQC spectra at different temperatures (30°, 40° and 45°). The conformational heterogeneity seems to be higher at 30° than at 40° further hinting at the conformational trigger being temperature. A second sample, labeled with ^{15}N isotopes at leucine and valine residues, revealed multiple conformations under physiological conditions, with peak doubling and tripling in ^{15}N-^{1}H-HSQC spectra. This further provides evidence for the existence of an alternate fold for CXCL11. To answer whether there is a switch in secondary structure, further studies need to be performed. Temperature and pH adjustments showed that at 45°C and pH 4.5, ITAC predominantly adopts a single conformation. This implies that the population of ITAC tips more in the direction of a single conformation at higher temperatures, as is the case for Lymphotactin. However, even under these conditions, some heterogeneity persisted, particularly at residues 43, 51, and 52. This heterogeneity seems to be well captured by Morpheus as the diversity plots shows a peak at these regions in Figure 5.1 (right).

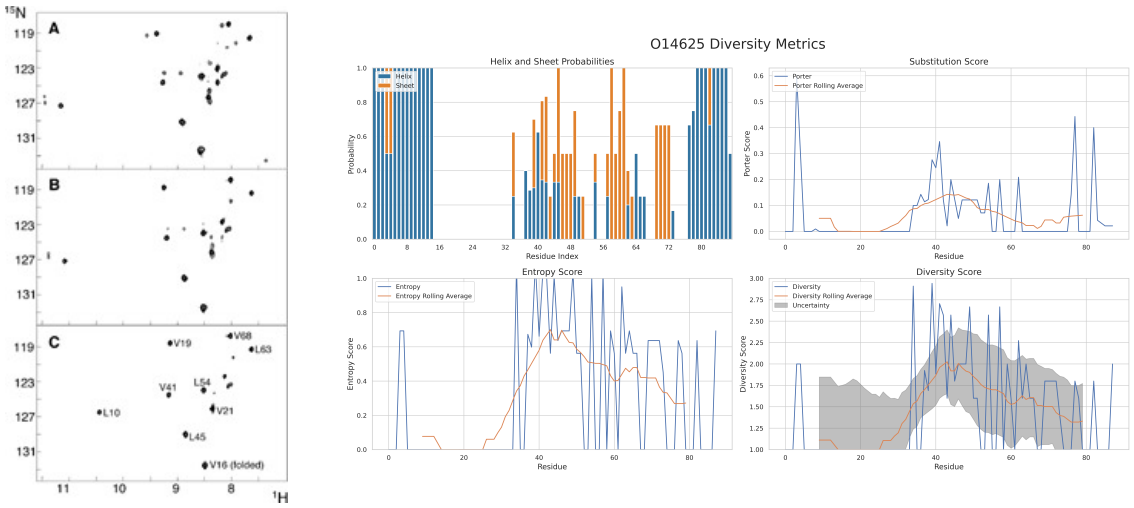

Figure 5.1: **CXCL11:** (left): ^{15}N-^{1}H-HSQC spectra of partially ^{15}N-labeled ITAC1–73 (all L,V ^{15}N-labeled for a total of nine labeled residues) with conditions (A) pH 5.0, 30°C; (B) pH 5.0, 40°C; and (C) pH 4.5, 45°C. Resonance assignments are indicated on C. (right): XCLC11 Diversity metrics plot.

Figure 5.1 (right) shows the diversity metrics of the chemokine protein CXCL11. The top left figure displays the helix and sheet propensity calculated by Morpheus fragment picking. This propensity plot helps us understand the nature and region of the predicted fold-switching. The remaining three subplots show the three diversity metrics: Diversity score, entropy score, subsitution score and uncertainty all labelled respectively in their subplots.

Important Observations through Morpheus:

1. Residues 43, 51 and 52 possessing more than one set of resonances at 45° Celsius seems to be captured by the diversity and entropy metrics giving maximum output near those residues.
2. The propensities for residues 35 to 70 show considerable diversity and mixing in helix, sheet and loop propensities. Existing NMR studies and our fragment-picking analysis suggests that this region might have the ability to switch-folds between multiple secondary structures.

5.2 Scope and Limitations

Typically, the two conformations of metamorphic proteins are experimentally characterized by sophisticated protein structure determination techniques like NMR and cryo-electron microscopy. Large changes in the secondary structure of proteins can be found through circular dichroism, which can be used to check the proteins metamorphic behvaiour Porter *et al.* (2022). These methods can prove to be time-consuming and resource-intensive. In addition, the trigger for fold-switching, which is unique to each metamorphic protein, may not be known beforehand. Thus, making the detection of metamorphic proteins further difficult. Therefore, providing a

way to systematically filter out monomorphic proteins and to search for candidate metamorphic proteins only using the sequence information of the protein is highly valuable. Our work is the first implementation of a proteome-level metamorphic protein predictor only using the primary structure infomration. Morpheus allows for the prediction of proteins whose structure may not be deposited in the Protein Data Bank. Hence, Morpheus contributes to the expansion of known set of metamorphic proteins.

Although using our algorithm, we are able to make high-throughput predictions of metamorphic behaviour in proteins to a good accuracy, our algorithm falls short at making sensitive predictions when the sequences differ by point mutations. Consider E48A in RfaH, this point-mutation has been shown to disrupt the inter-domain contact and thus favour all beta conformation.Burmann *et al.* (2012); Dishman and Volkman (2018) But this behaviour is not captured by Morphues. So while analysing highly similar homologous sequences, of which we would like find those sequences that may be metamorphic, the results obtained from our algorithm may be uncertain. Further, the accuary of Morpheus depends on the amount of protein fragment data it has to work with when making predictions. Moreover, because our algorithm considers the primary and secondary structure information of proteins, and not the tertiary information, it may fail to detect metamorphic proteins that retain their secondary structure but undergo substantial alterations in tertiary contacts and hydrogen bonding contacts.

5.3 Data Availability

The fragment-picking analysis is mainly performed using Python and Java; the code for the same is made available publically on Zenodo
https://zenodo.org/records/14837336. The proteome-level prediction data for the 57 proteomes is also uploaded. The algorithm is deployed on the web-server: http://mbu.iisc.ac.in/∼anand/morpheus.

5.4 Future Work

The shortlisted proteins are good candidates for experimentation to validate their metamorphic behaviour. Since most of the shortlisted proteins are less than 200 residues in length, performing an NMR based structure determination can give us insights into the fold-switching behaviour of the protein if present.

NMR based experiments here can help in searching the protein for structural heterogeneity, but one of the challenges involved in such a venture would be to find the trigger for the fold-switching behaviour of that particular candidate protein. In order to do this, the protein's gene ontology and function needs to be understood for deciphering the trigger.

One of the proteins shortlisted from Morpheus' predictions is a protein in Maize that is involved in circadian rhythm.(refer table 4.2). With proteins that are associated with circadian rhythms, there is a good chance that the fold-switching behaviour is periodic and helps entrain the organism with time of day. Experimentally characterizing this protein and searching for any periodic fold-switch is an avenue for novel research.

References

Abbass J, Nebel JC (2020). Enhancing fragment-based protein structure prediction by customising fragment cardinality according to local secondary structure. BMC Bioinformatics 21(1), 170.

Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, Ronneberger O, Willmore L, Ballard AJ, Bambrick J (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630(8016), 493–500. Publisher: Nature Publishing Group UK London.

Anfinsen CB (1973). Principles that Govern the Folding of Protein Chains. Science 181(4096), 223–230.

Anfinsen CB, Haber E, Sela M, White FH (1961). THE KINETICS OF FORMATION OF NATIVE RIBONUCLEASE DURING OXIDATION OF THE REDUCED POLYPEPTIDE CHAIN. Proceedings of the National Academy of Sciences 47(9), 1309–1314.

Artsimovitch I, Ramírez-Sarmiento CA (2022). Metamorphic proteins under a computational microscope: Lessons from a fold-switching RfaH protein. Computational and Structural Biotechnology Journal 20, 5824–5837.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000). The Protein Data Bank. Nucleic Acids Research 28(1), 235–242.

Booth V, Clark-Lewis I, Sykes BD (2004). NMR structure of CXCR3 binding chemokine CXCL11 (ITAC). Protein Science 13(8), 2022–2028. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1110/ps.04791404.

Burmann BM, Knauer SH, Sevostyanova A, Schweimer K, Mooney RA, Landick R, Artsimovitch I, Rösch P (2012). An α Helix to β Barrel Domain Switch Transforms the Transcription Factor RfaH into a Translation Factor. Cell 150(2), 291–303.

Chakravarty D, Porter LL (2022). Alphafold2 fails to predict protein fold switching. Protein Science 31(6), e4353.

Chakravarty D, Schafer JW, Chen EA, Thole JF, Ronish LA, Lee M, Porter LL (2024). AlphaFold predictions of fold-switched conformations are driven by structure memorization. Nature Communications 15(1), 7296. Publisher: Nature Publishing Group.

Chakravarty D, Schafer JW, Chen EA, Thole JR, Porter LL (2023). AlphaFold2 has more to learn about protein energy landscapes. Pages: 2023.12.12.571380 Section: New Results.

Chang YG, Cohen SE, Phong C, Myers WK, Kim YI, Tseng R, Lin J, Zhang L, Boyd JS, Lee Y, Kang S, Lee D, Li S, Britt RD, Rust MJ, Golden SS, LiWang A (2015). A protein fold switch joins the circadian oscillator to clock output in cyanobacteria. Science 349(6245), 324–328. Publisher: American Association for the Advancement of Science.

Chen N, Das M, LiWang A, Wang LP (2020). Sequence-Based Prediction of Metamorphic Behavior in Proteins. Biophysical Journal 119(7), 1380–1390.

Costello SM, Shoemaker SR, Hobbs HT, Nguyen AW, Hsieh CL, Maynard JA, McLellan JS, Pak JE, Marqusee S (2022). The SARS-CoV-2 spike reversibly samples an open-trimer conformation exposing novel epitopes. Nature Structural & Molecular Biology 29(3), 229–238. Publisher: Nature Publishing Group.

Dishman AF, Volkman BF (2018). Unfolding the Mysteries of Protein Metamorphosis. ACS Chemical Biology 13(6), 1438–1446. Publisher: American Chemical Society.

Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Chiu W, Garner EC, Obradovic Z (2001). Intrinsically disordered protein. Journal of Molecular Graphics and Modelling 19(1), 26–59.

Fox JC, Tyler RC, Guzzo C, Tuinstra RL, Peterson FC, Lusso P, Volkman BF (2015). Engineering Metamorphic Chemokine Lymphotactin/XCL1 into the GAG-Binding, HIV-Inhibitory Dimer Conformation. ACS Chem Biol 10(11), 2580–2588. Publisher: American Chemical Society.

Gront D, Kulp DW, Vernon RM, Strauss CEM, Baker D (2011). Generalized Fragment Picking in Rosetta: Design, Protocols and Applications. PLoS ONE 6(8), e23294.

Göbel U, Sander C, Schneider R, Valencia A (1994). Correlated mutations and residue contacts in proteins. Proteins: Structure, Function, and Bioinformatics 18(4), 309–317.

Hoeffding W (1994). Probability Inequalities for sums of Bounded Random Variables. In NI Fisher, PK Sen, editors, The Collected Works of Wassily Hoeffding, 409–426. Springer New York, New York, NY. Series Title: Springer Series in Statistics.

Jain V, Kikuchi H, Oshima Y, Sharma A, Yogavel M (2014). Structural and functional analysis of the anti-malarial drug target prolyl-tRNA synthetase. Journal of Structural and Functional Genomics 15(4), 181–190.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D (2021). Highly accurate protein structure prediction with AlphaFold. Nature 596(7873), 583–589.

Kabsch W, Sander C (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22(12), 2577–2637.

Khatua P, Ray AJ, Hansmann UHE (2020). Bifurcated Hydrogen Bonds and the Fold Switching of Lymphotactin. The Journal of Physical Chemistry B 124(30), 6555–6564. Publisher: American Chemical Society.

Kim AK, Looger LL, Porter LL (2021). A high-throughput predictive method for sequence-similar fold switchers. Biopolymers 112(10), e23416.

Kim AK, Porter LL (2021). Functional and Regulatory Roles of Fold-Switching Proteins. Structure 29(1), 6–14. Publisher: Elsevier.

Kim YI, Dong G, Carruthers Jr CW, Golden SS, LiWang A (2008). The day/night switch in KaiC, a central oscillator component of the circadian clock of cyanobacteria. Proceedings of the National Academy of Sciences 105(35), 12825–12830. Publisher: National Academy of Sciences.

Lei Y, Takahama Y (2012). XCL1 and XCR1 in the immune system. Microbes and Infection 14(3), 262–267.

Li Bp, Mao Yt, Wang Z, Chen Yy, Wang Y, Zhai Cy, Shi B, Liu Sy, Liu Jl, Chen Jq (2018). CLIC1 Promotes the Progression of Gastric Cancer by Regulating the MAPK/AKT Pathways. Cellular Physiology and Biochemistry 46(3), 907–924.

Mei Q, Dvornyk V (2015). Evolutionary History of the Photolyase/Cryptochrome Superfamily in Eukaryotes. PLOS ONE 10(9), e0135940. Publisher: Public Library of Science.

Mishra S, Looger LL, Porter LL (2021). A sequence-based method for predicting extant fold switchers that undergo α-helix $\leftrightarrow$ β-strand transitions. Biopolymers 112(10), e23471. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bip.23471.

Murzin AG (2008). Metamorphic Proteins. Science 320(5884), 1725–1726.

Mészáros B, Erdős G, Dosztányi Z (2018). IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. Nucleic Acids Research 46(W1), W329–W337.

Osman S (2023). Space exploration: finding new protein conformations using AlphaFold2. Nature Structural & Molecular Biology 30(12), 1835–1835. Publisher: Nature Publishing Group.

Porter LL, Kim AK, Rimal S, Looger LL, Majumdar A, Mensh BD, Starich MR, Strub MP (2022). Many dissimilar NusG protein domains switch between α-helix and β-sheet folds. Nature Communications 13(1), 3802.

Porter LL, Looger LL (2018). Extant fold-switching proteins are widespread. Proceedings of the National Academy of Sciences 115(23), 5968–5973. Publisher: Proceedings of the National Academy of Sciences.

Schafer JW, Porter LL (2023). Evolutionary selection of proteins with two folds. Nature Communications 14(1), 5478.

Solomon TL, He Y, Sari N, Chen Y, Gallagher DT, Bryan PN, Orban J (2023). Reversible switching between two common protein folds in a designed system using only temperature. Proceedings of the National Academy of Sciences 120(4), e2215418120.

Steinegger M, Söding J (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature Biotechnology 35(11), 1026–1028. Publisher: Nature Publishing Group.

Tuinstra RL, Peterson FC, Kutlesa S, Elgin ES, Kron MA, Volkman BF (2008). Interconversion between two unrelated protein folds in the lymphotactin native state. Proceedings of the National Academy of Sciences 105(13), 5057–5062. Publisher: Proceedings of the National Academy of Sciences.

Uversky VN, Oldfield CJ, Dunker AK (2008). Intrinsically Disordered Proteins in Human Diseases: Introducing the D^2 Concept. Annu Rev Biophys 37(1), 215–246.

Wayment-Steele HK, Ojoawo A, Otten R, Apitz JM, Pitsawong W, Hömberger M, Ovchinnikov S, Colwell L, Kern D (2024). Predicting multiple conformations via sequence clustering and AlphaFold2. Nature 625(7996), 832–839. Publisher: Nature Publishing Group.

Werner F (2012). A Nexus for Gene Expression—Molecular Mechanisms of Spt5 and NusG in the Three Domains of Life. Journal of Molecular Biology 417(1), 13–27.

Appendix A

Additional data

A.1 BioRxiv Preprint

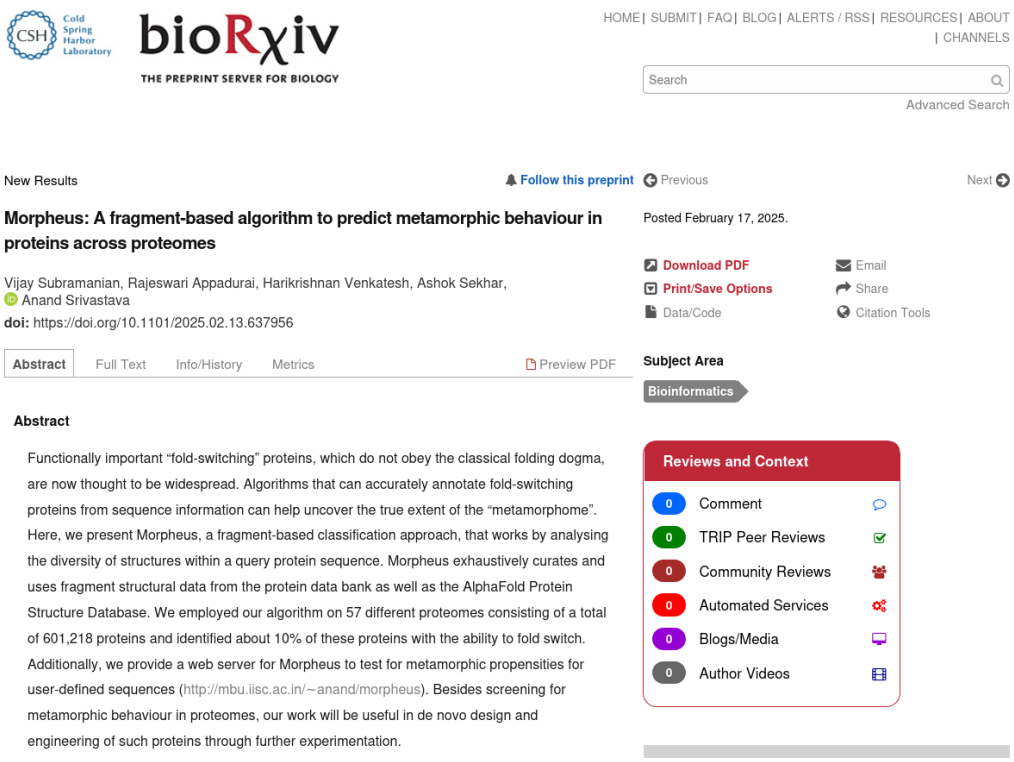

Figure A.1: **BioRxiv preprint:** screenshot shown above. The preprint is publically available using doi: 10.1101/2025.02.13.637956.

A.2 Training Dataset Described in Sec. 3.3

The data used for training the SVM model is borrowed from existing literature on fold-switching proteins and is provided in the tables below. In table A.1, the

columns PDB1 and PDB2 represent the PDB ID along with the chain ID for the two distinct secondary structures that were solved for the fold-switching protein.

A.2.1 Fold-switching Proteins Dataset

PDB1	PDB2	PDB1	PDB2	PDB1	PDB2
1ceeB	2k42A	3ejhA	3m7pA	4qdsA	2qqjA
1g2cF	5tpnA	3ewsB	3g0hA	4qhfA	4qhhA
1h38D	1qlnA	3gmhL	2vfxL	4rmbA	4rmbB
1k0nA	1rk4B	3hdeA	3hdfA	4rr2D	3l9qB
1mnmC	1mnmD	3ifaA	5et5A	4rwnA	4rwqB
1nqdA	1nqjB	3j7wB	3j7vG	4twaA	4ydqB
1ovaA	1jtiB	3j97M	1xtgB	4uv2D	4q79F
1qomB	1nocA	3j9cA	4h2aA	4wsgC	1svfC
1qs8B	1miqB	3jv6A	1zk9A	4y0mJ	4xwsD
1repC	2z9oB	4pyiA	4pyjA	4zrbC	4zrbH
1rkpA	2h44A	3kuyA	5c3iF	4zt0C	4cmqB
1uxmK	2namA	3m1bF	3lowA	5aoeB	5ly6B
1wyyB	5wrgC	3njqA	2pbkB	5b3zA	5bmyA
1x0gA	1x0gD	3o44A	1xezA	5c1vA	5c1vB
1xjtA	1xjuB	3qy2A	1qb3A	5ec5P	3zxgB
1xntA	3lqcA	3t1pA	1kctA	5ejbC	1wp8C
2a73B	3l5nB	3tp2A	5lj3M	5f3kA	5f5rB
2axzA	2grmB	3uyiA	3v0tA	5fhcJ	1eboE
2c1uC	2c1vB	3zwgN	4tsyD	5fluE	2uy7D
2ce7C	3kdsG	4a5wB	3t5oA	5hmgA	3ztjE
2gedB	1nrjB	4aanA	4aalA	5i2mA	5i2sA
2hdmA	2n54B	4ae0A	4ow6B	5ineA	3mkoA
2k0qA	2lelA	4b3oB	3meeA	5jzhA	5jztG
2lejA	2lv1A	4dxtA	4dxrA	5k5gA	2kb8A
2lqwA	2bzyB	4fu4C	4g0dZ	5keqF	1dzlA
2n0aD	2kkwA	4gqcC	4gqcB	5l35D	5l35G
2naoF	1iytA	4hddA	2lepA	7ahlE	4yhdG
2nntA	2mwfA	4j3oF	2jmrA	2KXOA	3RJ9C
2nxqB	1jfkA	4jphB	5hk5H	6Z4UA	7KDTB
2ougC	2lclA	4nc9C	4n9wA	4KSO	
2p3vA	2p3vD	4o0pA	4o01D	1F16	
2qkeE	5jytA	4phqA	2wcdX	8E6Y	

Table A.1: 189 Metamorphic proteins that are used for training the model. PDB1 and PDB2 refer to the two different solved structures of the protein

A.2.2 Monomorphic Proteins Dataset

PDB ID	PDB ID	PDB ID	PDB ID
1O7N.B	5DPJ_1	1VF1.A	1PHN.A
1YCK.A	5DPG_1	2CHH.A	1W6N_1
2HBT.A	2ZMU_1	1FUS_1	2UYZ.A
1ELT.A	1RYP.L	3D80.A	1W0T_1
3RP2.A	1AUN.A	2INC.C	2M34_1
2BV4.A	3MGQ_3	2VML.A	3E2C.A
1NN6.A	2DYR_9	2DK7_1	3E86.A
1IAU.A	3MGQ_2	1KT6.A	2BC3.A
1SLT_1	2QWX.A	1BIO.A	2M7B_1
2W69.A	2JXW_1	1J7D.B	1UZV.A
1RVE.A	4UBP_1	2JE7.A	1KM4.A
2OQA.A	1AB9.B	2DWV_1	2HI3_1
2DYR_7	1A53.A	1ZBF.A	2GDG.A
1FON.A	2AWK_1	1IIU.A	3NAU_1
2HFT.A	2ELJ_1	1O6W_1	2HBA_1
1PMY.A	2JET.B	3DHI.C	4Z4M_1
3MGQ_1	8DFR.A	1AHC.A	4Z4K_1
2J96.A	2YS9_1	1HJ8.A	1GFL.A
1LM4_1	5DTX_1	3MGQ_4	2YSH_1
3EF4.A	1RYP.H	2JKH.A	7FD1.A
2A06.F	3SEB.A	3CHB.D	2GJD.A
2BUR.B	2VZX_1	1JJT.A	3B44.A
1TK7_1	3EIK.A	2VHK.A	2JV4_1
1QZP_1	1Z3Q.A	3D6M.B	1JOT.A
1EJX.B	3D6M.A	2O7M.A	5KEE_1
1KEQ.A	1CPC.A	1JBO.A	4EQP_1
1X8Q.A	1KMV.A	2UX7.A	5JOB_1
1YPH.C	1T32.A	1K55_1	2VI6_1
1K3Y.A	1N5N.A	1UBQ.A	2OV0.A
3ERX.A	2DYR.F	1PAZ.A	2RK3.A
2YSE_1	1WMV_1	2IH3_3	1UGX.A
1LQY.A	1WXC.B	3PCC.B	2VLQ.A
1MIJ_1	2CXQ_1	1CEX.A	2KPZ_1
1MQO.A	1UIZ.A	1BW5_1	1FME_1
2JOF_1	2YSB_1	3L4H_1	3ELN.A
1QMJ.A	4REX_1	1OPS.A	2Z8A.A
1Y71_1	2HPW_1	2JX8_1	2YSD_1
2ZO6_1	1T56.A	1WI3_1	3A03_1
1FIP.A	2KMU_1	1GB1_1	2NN8.A
1XME.B	1MVQ.A	1I5H_2	2YSG_1
2OGQ.A	2BV8.A	2A3D_1	1E85_1
3BWH.A	1F99.A	1KPF_1	2YSF_1
1XEO.A	2PGZ.A	3DR9.A	1I5H_1

2GH9_1	1LMB_3	2FKZ.A	3DJH.A
2DSC.A	5EJU_1	1GH0.A	253L.A
2FDJ.A	2BMO.B	1Y66_1	2DMV_1
5DPI_1	3DUI.A	5DTZ_1	3A02_1
5DPH_1	4HVF_1	5DU0_1	2DKO.B
5EHU_1	2UU8.A	2IB5_1	
5DY6_1	1MRJ.A	1L2H.A	

Table A.2: 198 Monomorphic proteins that are used for training the model. The character after the dot/underscore refers to the chain ID of the protein

A.3 Cross-validation

Figures A.1 to A.6 show how the decision boundary changes, for Morpheus on its training set, when the quadratic SVM model is trained with on a 6 fold cross-validation scheme. On each figure, when that particular partition is used, the corresponding validation matrix is also shown on the right. Figure A.7 shows the parallel coordinate plot for the 4 features of Morpheus and final confusion matrix for Morpheus training dataset.

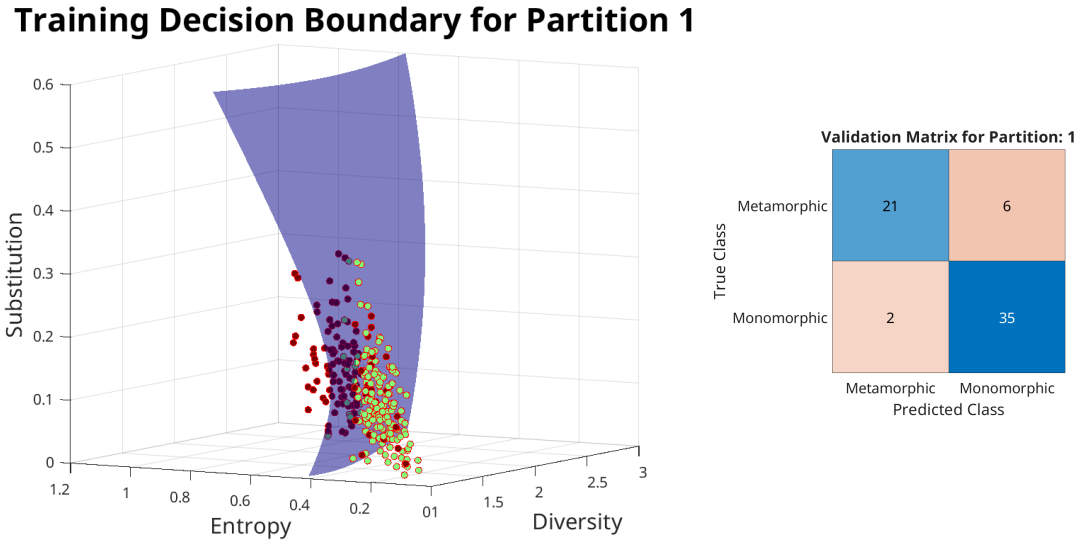

Figure A.1: **First partition split:** The decision boundary and its corresponding validation confusion matrix is shown for the first training split partition

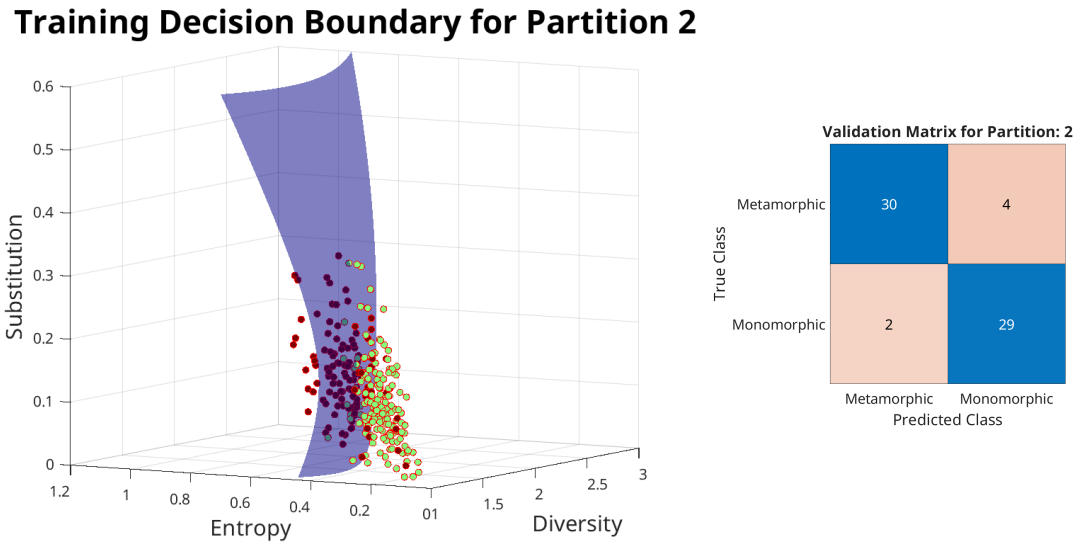

Figure A.2: **Second partition split:** The decision boundary and its corresponding validation confusion matrix is shown for the second training split partition

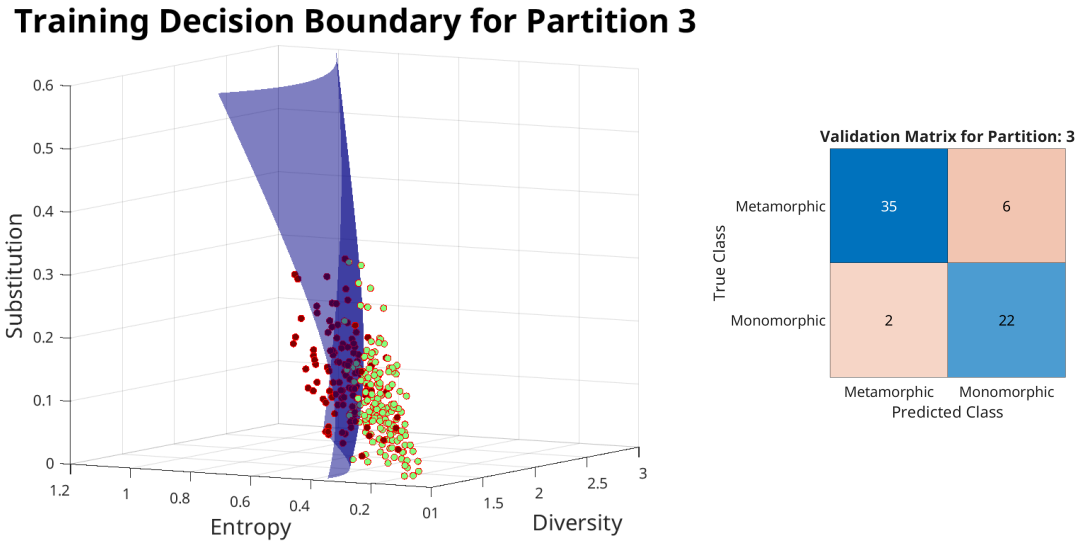

Figure A.3: **Third partition split:**The decision boundary and its corresponding validation confusion matrix is shown for the third training split partition

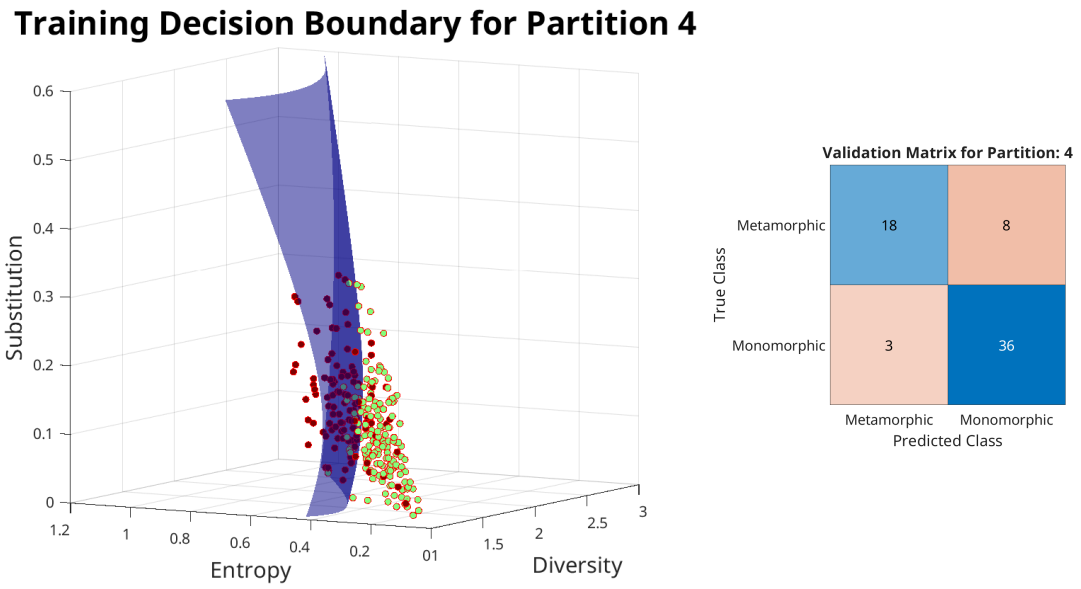

Figure A.4: **Fourth partition split:** The decision boundary and its corresponding validation confusion matrix is shown for the fourth training split partition

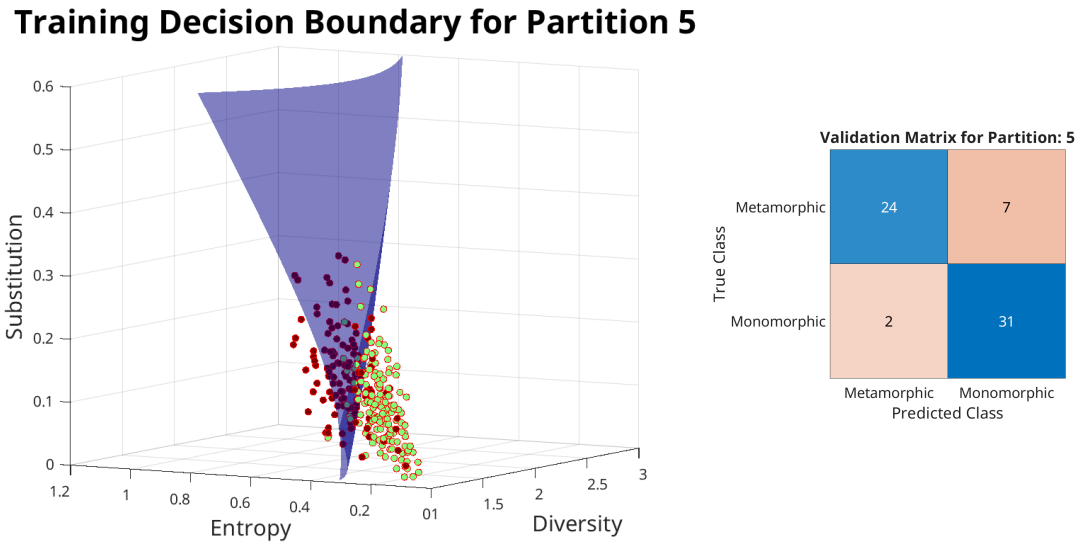

Figure A.5: **Fifth partition split:** The decision boundary and its corresponding validation confusion matrix is shown for the fifth training split partition

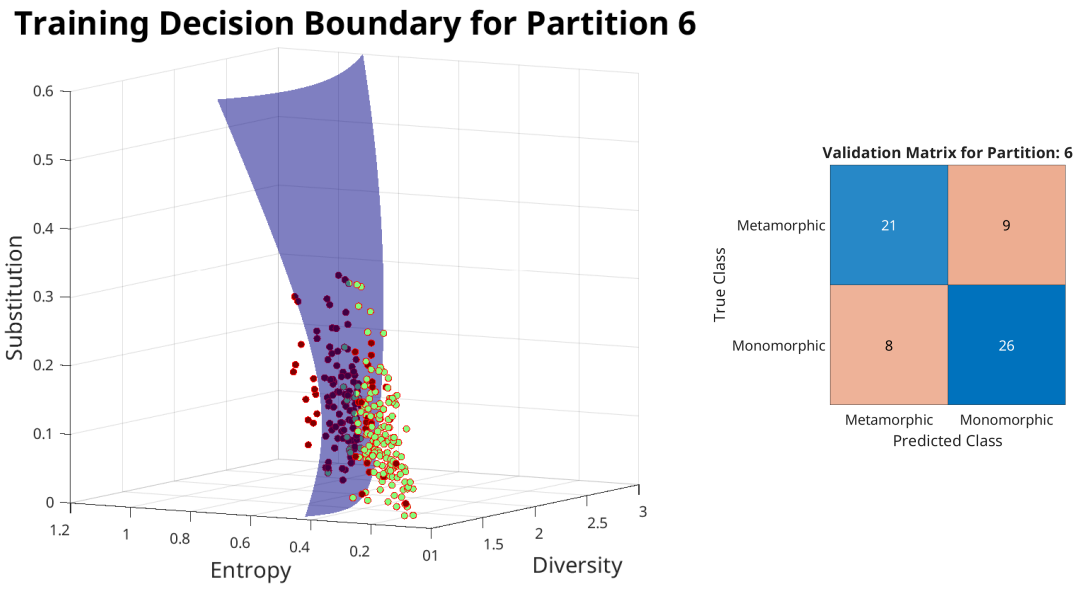

Figure A.6: **Sixth partition split:** The decision boundary and its corresponding validation confusion matrix is shown for the sixth training split partition

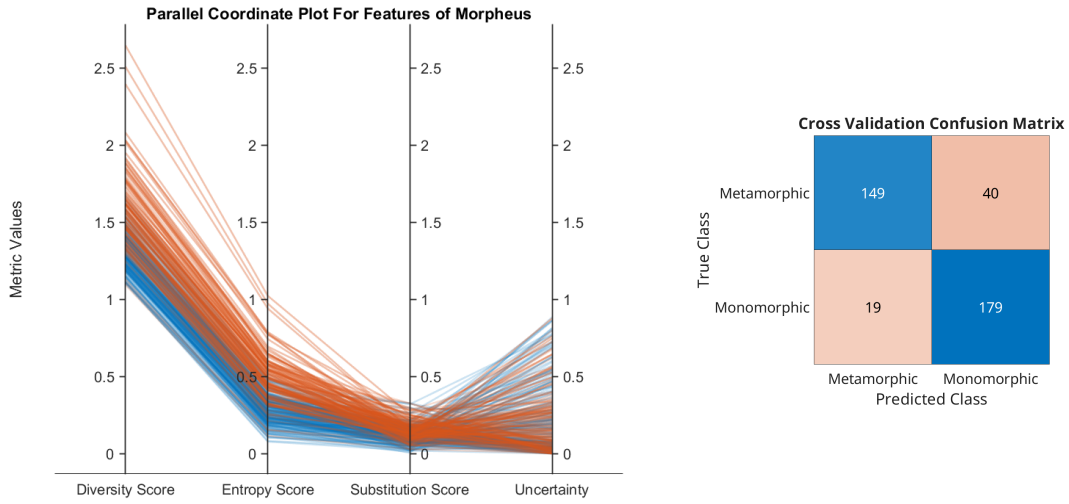

Figure A.7: **Parallel coordinates plot:** The parallel coordinates plot for the four features of Morpheus and the final validation confusion matrix.