Training a Support Vector Machine for Protein Secondary Structure Prediction Using Protein Language Model Embeddings

Julia Malec
Department of Computer Science
University of Western Ontario
London, Canada
jmalec@uwo.ca

Jonathan Oxman
Department of Computer Science
University of Western Ontario
London, Canada
joxman@uwo.ca

Charles Schepanowski
Department of Computer Science
University of Western Ontario
London, Canada
cschepan@uwo.ca

***Abstract*—Proteins are the basic molecule on which all life is based, and so their study is of great interest in the medical, pharmacological, and general life sciences. A protein's function is directly determined by its chemical properties, which follow from the physical structure of the molecule. Originally studied via biochemical experimentation, protein structure is now most commonly determined through bioinformatics methods, often utilizing machine learning techniques. Proteins are most easily represented through their protein sequence, which does not capture the full structural information of the protein. Thus, the prediction of protein structure from the sequence is a key problem in this field. The final goal of protein structure prediction is usually to determine the protein's tertiary structure, which encodes the full structural information: this is achieved with high accuracy by contemporary models such as AlphaFold. An often-used intermediate step in this problem is the protein secondary structure, which encodes local substructures of the molecule. Secondary structure is also useful in passing additional information to other bioinformatics algorithms, and so remains of interest in the field despite AlphaFold's successes. The most recent development in secondary structure prediction is Porter 6, released on December 27, 2024, which achieved an 86.6% classification accuracy using a combination of protein language embedding models and deep neural networks. We consider a similar approach, utilizing ProtT5 to encode protein data and a support vector machine (SVM) classifier with a sigmoid kernel to achieve 84% classification accuracy on an extremely limited dataset both in terms of sample size and quality. Despite the simpler model architecture, our model's performance is competitive with the state-of-the-art Porter 6, and outclasses most other contemporary secondary structure prediction methods.**

***Index Terms*—Protein Sequence, Machine Learning, Support Vector Machine, Natural Language**

I. Introduction

Proteins are the fundamental building blocks of life, and are responsible for nearly all biological functions in living organisms. As such, the study of proteins is critical in nearly every aspect of life sciences. One of the key ways to understand proteins and their properties is by considering their molecular structure. This structure determines the form and function of any protein, so in understanding it we can begin to potentially unlock the mysteries of what function that protein serves and how it interacts with surrounding proteins and other molecules.

The structure of proteins can be divided into four categories, ordered by layer of complexity: primary, secondary, tertiary, and quaternary protein structure. Proteins are comprised of amino acids attached to one another in a chain and arranged in a particular order. Primary protein structure merely encodes this order. The higher levels of protein structure encode more complex information regarding the physical form the amino acid chain takes in three-dimensional space.

Traditionally, protein structure has been determined through biochemical experimentation. However, this is a slow process, and in recent years various bioinformatics techniques have been leveraged to try to predict the structure of a protein through computational algorithms rather than physical experiments. Typically, the final result of interest is the tertiary or quaternary structure. However, the secondary structure remains useful either as an intermediate step to assist in this end goal, or on its own merits. Tertiary structure fundamentally encodes more information than primary and secondary structure due to the extra global spacial data. Hence, the relative simplicity of secondary structure suggests its prediction may be achieved with high accuracy and low computational cost via a more basic approach than tertiary structure. Indeed, secondary structure is still being researched, with improvements being made as late as December 27, 2024 [1].

Ideally, we could create an algorithm that can predict secondary structure efficiently and with near-perfect accuracy. However, in practice, this is not possible either with a traditional algorithm or a machine learning approach, due to factors such as the complexity of the problem, a lack of quality training data, and computational constraints [2]. Instead, we adopt the humbler goal of attempting to classify the secondary structure on any input protein with high accuracy, low computational cost at prediction time, and reasonable simplicity. For this, we consider a support vector machine approach, which is typically well-suited to this type of classification problem with limited training data due to its resistance to overfitting. To capture the underlying complexity of the problem without violating our need for model simplicity, we employ a nonlinear kernel.

II. Background

A. Proteins & Amino Acids

Proteins are complex biomolecules important for a broad range of tasks in organisms, including catalyzing biochemical reactions, providing structural support, facilitating cell signaling, transporting molecules, and defending against pathogens. Proteins are composed of smaller molecules, called *amino acids*, that are linked together in a chain. This chain folds into a specific three-dimensional shape, and the protein's shape directly determines its function.

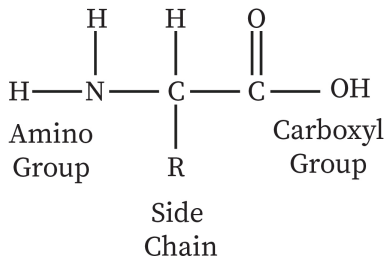

Fig. 1: The general chemical structure of an amino acid [3].

There are 20 distinct common amino acids, each sharing a common structure (Figure 1): an amino group (NH_2), a central carbon atom, and a carboxyl group (COOH). What differentiates amino acids is their unique *side chain*, or R group, attached to the central carbon. For example, alanine's side chain is a simple methyl group (CH_3), consisting of one carbon atom and three hydrogen atoms. In contrast, phenylalanine has a bulkier side chain, consisting of a carbon atom bonded to and a six-carbon ring. (See Figure 2.)

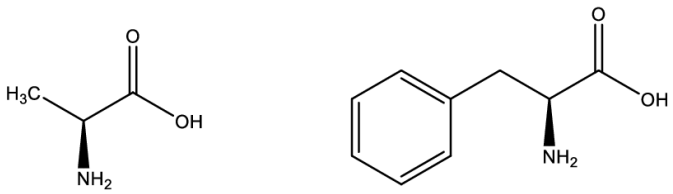

Fig. 2: Chemical structures of alanine and phenylalanine [4].

B. Protein Primary Structure

Amino acids are linked through dehydration synthesis reactions, a type of chemical reaction that occurs between the amino group of one amino acid and the carboxyl group of another. Through a series of such reactions, a chain of amino acids forms with the following structure: a repeating sequence of nitrogen (N) and two carbon (C) atoms, referred to as the protein's *backbone*, where each amino acid's side chain (R group) extends outward from the backbone. In protein synthesis, amino acids are linked in a specific order determined by genetic information stored within DNA. The sequence of amino acids in the chain is known as the *primary structure* or *primary sequence* of a protein. Each amino acid can be represented by a single-letter code (Figure 3), making it easy to encode protein sequences as strings of letters in a computer. When amino acids are linked in this manner, they are commonly referred to as *residues*.

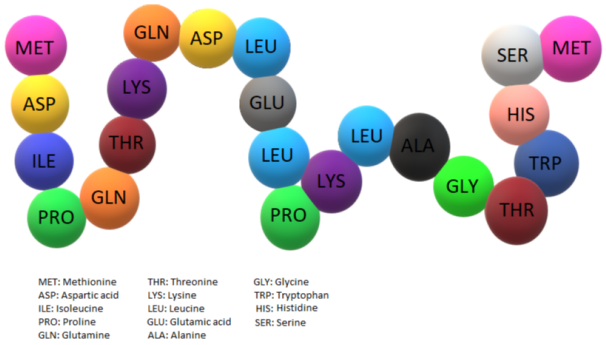

Fig. 3: An example of a chain of amino acids [5]. The corresponding encoded sequence is given by MDIPQTKQDLELPKLAGTWHSM.

C. Homology

Protein *homology* refers to the similarity between protein sequences due to shared ancestry. Although homology is a qualitative concept that cannot be easily quantified, certain properties are widely accepted as evidence of a homologous relationship between proteins. One such property is protein sequence similarity, which measures the degree of resemblance between proteins based on their amino acid composition [6]. Protein sequence similarity is computed via *sequence alignment*, a computational method that aligns protein sequences to maximize matches between them. This process accounts for substitutions, insertions or deletions in either sequence to identify regions of similarity. Conserved regions, in which sequences are highly similar, often indicate critical functional or structural roles [7]. Pairwise sequence alignment algorithms are reasonably fast, with a worst-case time complexity of $O(nm)$ [8] where n and m denote the lengths of the protein sequences. On the other hand, aligning more than two sequences, known as *multiple sequence alignment (MSA)*, is generally computationally expensive and relies on approximation techniques rather than fully correct algorithms [9].

D. Protein Secondary Structure

The order of the amino acids in a protein determines how the chain will fold and the overall structure of a protein dictates its function. As the protein is synthesized by joining amino acids, the backbone starts to fold into local substructures known as *secondary structures*. Secondary structures can be classified into several categories, the two most common of which are the α-helix, where the carbon backbone forms a spiral, and the β-sheet, where the carbon backbone folds into a zigzag

pattern (Figure 4) comprised of individual substructures known as β-strands. Regions that do not fall into either class will be referred to as *coils*. Protein structure is dictated by the principle of *energetic favourability*: the chain of amino acids folds in such a way as to render the resulting structure the most energetically stable [10]. Due to the physical and chemical properties of each amino acid, the specific order of amino acids in the chain determines which secondary structures form during the folding process, since it is more energetically favourable for certain consecutive amino acids to adopt a particular secondary structure over others.

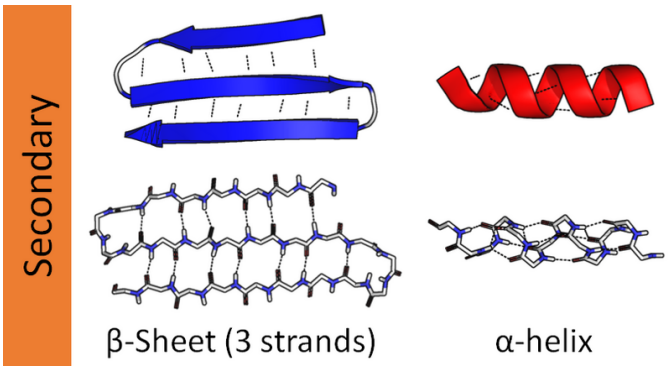

Fig. 4: Common secondary structures in proteins [11].

E. Protein Tertiary and Quaternary Structure

The functional form of many proteins is the *tertiary structure*. The tertiary structure can be thought of as an arrangement of secondary structures positioned in space relative to each other (see Figure 5). While the formation of secondary structures occurs within local regions of the amino acid chain, for example, a short segment of five amino acids spiralling into an α-helix, this next stage considers long-range interactions. Any atom in any amino acid could potentially come into close proximity with any atom in another amino acid, regardless of their distance in the primary sequence. These interactions can either be energetically favourable or unfavourable. For instance, side chains with similar charges (both positive or both negative) will repel each other, much like the repulsion between similarly charged magnets. On the other hand, oppositely charged side chains can attract, forming stabilizing interactions. Additionally, to be functional, some proteins require a quaternary structure, which is a larger complex formed by the combination of multiple independently folded tertiary structures. Once experimentally determined, structures are usually deposited in public databases like the Protein Data Bank (PDB) [12], making them accessible to the scientific community.

F. Insights from Protein Structure

The physical structure of proteins is not only interesting from a general scientific point of view, but extremely useful in many fields. For instance, understanding a protein's structure is crucial in drug discovery, where drug design often involves creating molecules that bind to specific sites on harmful proteins to block or reduce their activity [13], [14]. Another application is the detection of dangerous mutations in proteins. A mutation is defined as a change in the amino acid sequence of a protein, such as the incorrect substitution of one amino acid with another during protein synthesis. Such mutations can have varying effects on the protein: if a mutation occurs in a region of the protein that is key in determining the overall structure, it can alter the protein's structure in a way as to render the protein non-functional. In humans, this manifests itself as many diseases including Alzheimer's disease, diabetes, and some cancers [15], [16]. One method of diagnosing such diseases is by examining specific proteins in the body and comparing their structure with the known healthy structure of the protein. Thus, understanding the structure of proteins offers great utility in both identification and treatment of diseases.

Although predicting the full structure of a protein is the most useful in terms of applications, protein secondary structure has its own uses. Since tertiary structure is composed of secondary structures, the secondary structure plays a key role in determining it. As such, understanding the secondary structures a protein forms is an important step in determining the tertiary structure, and many algorithms designed to recover the tertiary structure of a protein rely on the secondary structure as part of the input information. In addition to this intermediate role, secondary structure can also be used directly in bioinformatics applications such as PRALINE [17], an MSA tool which takes advantage of the additional information provided by the secondary structure.

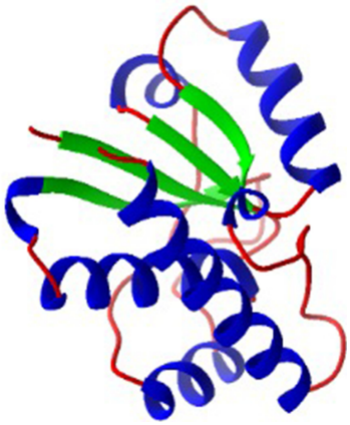

Fig. 5: A visualization of a tertiary structure [18]. Blue: α-helix. Green: β-sheet. Red: coil.

G. Protein Structure Prediction

Given the significance of the information encoded in a protein's structure, it is natural to try to catalog the structure of every protein. While it is possible to determine the protein structure via biochemical experimentation, this process is time-consuming and expensive [19], with less than 250,000 protein structures determined through this process as of 2024

[20]. In contrast, determining the order of amino acids in a protein is relatively straight forward [21], highlighting the value of having a computational model to predict protein structure directly from the primary sequence. Protein structure prediction is the computational process of determining the three-dimensional structure of a protein based on incomplete data such as the primary structure.

H. Classification of Secondary Structure

Experimentally resolved structures usually correspond to the tertiary level of protein folding. To derive secondary structure labels for individual amino acids, indicating the type of secondary structure each residue belongs to, algorithms such as Define Secondary Structure of Proteins (DSSP) [22] are commonly used. For a protein sequence of length n, its corresponding *labeled sequence* consists of two strings of length n: one being the protein sequence itself, and the other indicating the corresponding secondary structure labels. Secondary structure labels are typically assigned using either a three-state or eight-state classification system. In the three-state system, each amino acid is categorized as part of a helical structure, sheet/strand structure, or coil (a catch-all for any regions not falling into either of the first two classes). Three-state classification systems are typically evaluated by Q3 accuracy, which measures the percentage of correctly predicted amino acid positions in the three categories: helix, sheet/strand, and coil. The eight-state system offers a more detailed classification by subdividing the three main secondary structure types into further subcategories, and is similarly evaluated with Q8 accuracy. We present an example of a Three-state classification of a protein, where H denotes a helical structure, E denotes a sheet/strand structure, and C denotes a coil:

- **Protein Sequence:**
  `CSCSSLMDKECVYFCHLDIIW`
- **Label Sequence:**
  `CCCCCCCHHHCCCCCCEEECC`

I. Support Vector Machine (SVM)

Support vector machines (SVM) [23] are a type of supervised machine learning classification algorithm. Given a dataset $S \subset \mathbb{R}^n$, the core concept of the SVM algorithm is finding an $n-1$-dimensional hyperplane $H \subset \mathbb{R}^n$ that separates vectors with different labels. Though this may appear to be an intrinsically linear model and thus lack the complexity required to capture nonlinear relationships in the data, this obstacle is easily surmounted by means of applying a nonlinear kernel to map the data from its original space into a high-dimensional space in which we hope the image of the data becomes linearly separable. Indeed, with the correct (sufficiently complex) kernel, almost any relationship can be captured [24]. However, in practice, it is not feasible to find this 'perfect' kernel, and so in general SVM algorithms settle for a kernel that offers a desired level of performance without taking too many computational resources to be found. The most commonly used kernel types include polynomial, sigmoid, and radial basis function (RBF) kernels [25]. While the above discussion applies only to binary classification, it can be easily extended to multiclass classification problems by separately training a classifier for each pair of classes, commonly known as a one-vs-one approach. SVMs have seen success when applied in many binary and multiclass classification problems [26]. However, one restriction appears in that SVMs are not particularly well-suited to large datasets, due to their poor (quadratic) time complexity in the number of training samples [23]: this can be addressed with approximation techniques [27].

J. Protein Language Models

Proteins carry an enormous amount of biological information that is difficult to capture through conventional algorithmic techniques [10]. One parallel example of such a problem is that of language processing: similarly to the amino acids within a protein sequence, the meaning of words in a sentence is determined not just by a dictionary definition of the word, but the context the word appears in. One approach that has successfully tackled this problem is language embedding models, which encode each word (or, in more sophisticated models, each token, defined as a group of characters that often appears in sequence [28], [29]) into a vector in a high-dimensional vector space. These encodings, or *embeddings*, capture not only the intrinsic meaning of the encoded symbol, but also its surrounding context.

One such model which has recently seen particular success in protein sequencing applications is ProtT5 [30], a state-of-the-art protein embedding model based on a transformer architecture and pre-trained on large protein sequence datasets using unsupervised learning techniques. Through ProtT5, each amino acid in a protein sequence is embedded in a 1024-dimensional vector space. Past empirical results indicate that these protein embeddings capture many complex relationships within the amino acid sequence, including information about the protein structure [31].

III. Related Work

A. Past Approaches to Secondary Structure Prediction

Secondary structure prediction can be framed as a classification problem, where each amino acid in a protein is assigned to a class within either the three-state or eight-state system. Secondary structure machine learning models learn patterns from labeled data to identify which consecutive stretches of amino acids are most likely to adopt specific secondary structures. The first application of a neural network to protein secondary structure prediction was in 1989 [32], leading to the development of an ensemble model based on a perceptron with a single hidden layer, which achieved a Q3 accuracy of 63%. This unimpressive performance can be partially attributed to the small size of the training dataset, which consisted of only 48 proteins, as well as the computational limitations of the time which restricted the model to a shallow feed-forward architecture.

SVMs were first applied to the secondary structure prediction problem in 2001 [33], achieving a Q3 accuracy of 73.5%. This was done using a radial basis kernel function with a kernel coefficient (γ) of 0.01 and a regularization parameter (C) of 1.5. In both methods, the protein sequence was encoded as a sequence of one-hot vectors utilizing a sliding window approach. An input to the models is constructed for each amino acid as follows: given the amino acid A in position k, let S_1 denote the subsequence of amino acids in positions $k-n, \ldots, k-1$, and let S_2 denote the subsequence of amino acids in positions $k+1, \ldots, k+n$ (here, n varies depending on the method, but must be consistent in the definitions of both S_i). Let S denote the combined sequence S_1AS_2. The input to the model corresponding to A is then the matrix obtained by appending the one-hot vectors corresponding to the amino acids in S in order: in other words, the matrix representing a window of amino acids centred on A. There are twenty distinct encodings of amino acids, and an additional one-hot vector is required to represent the null case in which part of the sliding window lies past the end of the protein sequence. Thus, the one-hot vectors involved are 21-dimensional. In 2003 a similar approach was tried [34], replacing the one-hot encodings with PSI-BLAST profile vectors [35] and employing a quadratic kernel function, with a Q3 accuracy of 77%. Unlike the arbitrary one-hot encodings, PSI-BLAST profiles, which are generated through a standard BLAST search [36] followed by an alignment of similar protein sequences, provide a detailed evolutionary representation of a protein sequence and allow for the passing of additional biological information to the model.

B. Modern Methods

With the advancement of computational resources and the growing availability of experimentally determined protein structures, machine learning approaches for protein structure prediction have significantly improved. Porter, a widely used computational tool for secondary structure prediction, released its sixth version, Porter 6 [1], on December 27, 2024: this recent paper evaluates various deep learning models trained on several distinct sequence embeddings, including one-hot vectors, ProtT5, and ESM-2 [37]. These embeddings were tested with deep learning architectures including convolutional neural networks (CNN), recurrent neural networks (RNN), and long short-term memory (LSTM) networks. Among these, the convolutional bidirectional recurrent neural network (CBRNN), which integrates both CNN and RNN components, demonstrated the best performance. Using ESM-2 embeddings, the CBRNN achieved the highest accuracy, with a Q3 accuracy of 86.6% and a Q8 accuracy of 76.43% on a 2022 test set, and a Q3 accuracy of 84.56% and a Q8 accuracy of 74.18% on a 2024 test set. These results either match or surpass state-of-the-art methods in the field. Recently, protein language model embeddings have been used alongside CNNs and RNNs as alternatives to MSAs for secondary structure prediction [38], [39]. The theoretical upper limit of three-state (Q3) accuracy, approximately 91% [2], has yet to be reached, highlighting the importance of continued research to develop advanced, efficient and accurate secondary structure prediction models.

C. Objective

A key limitation of the PSI-BLAST approach is that obtaining PSI-BLAST profiles is computationally expensive, due to repeated large-scale database searches, the necessity for multiple sequence alignments (MSA), and the iterative refinement of profile vectors, all of which contribute to the high computational cost. This highlights a key limitation of many structure prediction models, including traditional secondary structure prediction models [40] and advanced state-of-the-art tertiary structure prediction models like AlphaFold2 [41]: their dependence on MSAs and a supply of homologous sequences. Not only do the computational costs associated with computing an MSA impact the scalability of these models; it is also not always possible to find sufficient homologous sequences for a given protein, a requirement for generating a reliable and informative MSA. These issues make MSA-dependent methods impractical for certain applications [42], underscoring the need for further research into accurate and efficient protein language model based approaches.

One of the most notable breakthroughs in protein structure prediction was the development of Alphafold [41] in 2020, which achieved near-experimental accuracy in predicting tertiary protein structures. Despite these advancements in tertiary structure prediction, protein secondary structure prediction remains an active area of research, as secondary structure prediction can be thought of as an intermediate step in determining the overall fold and can serve as input for tertiary structure prediction models such as ROSETTA [43] and MODELLER [44]. As such, a model that can accurately predict secondary structure remains useful in the current state of the field. We aim to construct such a model by leveraging novel machine learning methods. Alongside high accuracy, our target is model simplicity and low computational cost, particularly at prediction time. Ideally, we would like to do away with the necessity of computing an MSA to retrieve a secondary structure prediction.

IV. Approach

At their core, PSI-BLAST profile vectors encode a position-specific probability distribution of amino acid residues at each position of a query sequence based on a collection of similar sequences. Using this encoding instead of one-hot vectors improved accuracy in previous approaches, and so our approach is inspired by the potential to further extend the performance of secondary structure prediction methods with even more informative input vectors. Protein embedding vectors are more informative than PSI-BLAST profile vectors because they consider the entire protein sequence to capture semantic properties of the protein when generating each vector. Additionally, they offer the added advantage of bypassing the MSA computation.

With the conversion of the amino acids in each protein into vectors, the problem then presents itself as a simple three-

class classification problem on points in a vector space, for which an SVM approach is natural. Although we have seen a comprehensive evaluation of many deep learning models to identify the most effective strategy for protein secondary structure prediction [1], a research gap remains in the application of SVMs to this problem. The SVM architecture has the advantages of simplicity, which, along with regularization techniques, allows it to handle relatively small datasets without risk of overfitting. With this in mind, the overall prediction pipeline we consider is as follows:

1) Each input protein sequence is converted into a set of 1024-dimensional embedding vectors via ProtT5.
2) Each embedding vector is passed through an SVM which classifies its secondary structure.

V. Methods

A. Data

We consider a pre-cleaned dataset [45] compiled from three benchmark datasets: CASP12 [46], CB513 [47], and TS115 [48].

1) CASP12: Protein structure prediction is a critical challenge in bioinformatics, and to evaluate the performance of computational methods in this field, a biannual community-wide experiment, known as the Critical Assessment of Structure Prediction (CASP), is conducted. CASP provides an objective benchmark by comparing the predictions of participating groups to experimentally determined protein structures, which remain unpublished during the assessment period. As the gold standard for assessing progress in computational protein structure prediction, CASP highlights the strengths and weaknesses of current methods. The CASP12 dataset is sourced from the CASP assessment held in 2016.

2) CB513: The CB513 dataset is a well-known benchmark in the field of protein secondary structure prediction, consisting of 513 non-redundant protein chains. The CB513 dataset maintains a low sequence identity among its protein chains and includes experimentally determined secondary structures, providing a reliable and diverse set of examples for training and testing.

3) TS115: TS115 is a recognized dataset frequently employed to measure the performance of secondary structure prediction models, comprising 115 non-redundant protein chains annotated with secondary structure labels. Its sequences like CB513 maintain low sequence identity. TS115 is often used alongside other datasets like CB513 due to its small size.

The data consists of distinct protein sequences, each paired with its corresponding sequence of three-state labels for the secondary structure. Data is split into a training set consisting of 8,678 protein sequences and a test set of 649 sequences. Each sequence is approximately 260 amino acids in length on average, ranging from a length of 20 to 1632. A subset of both the training and test protein sequence data is passed through ProtT5 to convert each amino acid in each sequence into an embedding vector: this procedure yields a training and test set of approxmiately 894,286 and 180,000 individual embedding vectors and their corresponding labels, respectively. The ratio of the class labels among each set is as follows:

- **Training Set:** Helix: 0.35, Sheet/Strand: 0.21, Coil: 0.42
- **Test Set:** Helix: 0.21, Sheet/Strand: 0.36, Coil: 0.43

B. Model Construction

We handle the more complicated multi-class SVM classification problem by means of a one-vs-one decision function. Given the complexity of the problem, it is almost certain that the data is not linearly separable. Hence, we employ a nonlinear kernel in the SVM. The following commonly-used kernels are considered: radial basis function (RBF), polynomial, and sigmoid. For each of these kernels, a grid search with five-fold cross-validation is employed to determine the optimal parameters for the model from the following parameter grid:

- **Regularization Parameter (C):** $\{0.9, 0.95, 1, 1.05, 1.1\}$
- **Kernel Coefficient (γ):** $\{1, 0.5, 0.1, 0.01, 0.001\}$
- **Degree (for a polynomial kernel):** $\{2, 3, 4\}$

Although the most rigorous approach for hyperparameter tuning requires testing each set of hyperparameters on the complete training set, this is not feasible due to computational constraints. Instead, we first determine the optimal hyperparameters on a randomly chosen subset of 10,000 training samples, and, afterwards, use them to train the full model, with a training set of 150,000 training samples and 25,000 test samples. The final model is then evaluated on the reserved test set: since there is no particular practical disadvantage to misclassifying one structure over another, we use Q3 accuracy as the performance metric.

C. Visualization

For visualization of the results, we present for each class the summary statistics of the classification on the test set, including precision, recall, f1-score, and support. For evaluation of the model selection process, we generate a learning curve to visually demonstrate the generalization error.

VI. Results

A. Model Selection

Running the cross-validated grid search on training sets of 10,000 samples yielded promising results, particularly for the RBF and sigmoid kernels. (The polynomial kernel, for which it was determined that a degree of 2 was optimal, lags in performance). We present the results of the hyperparameter search in Table I. The resulting best model was configured with a sigmoid kernel, a regularization coefficient of 1, and a kernel coefficient of 0.01. The learning curve of this model, trained on increasing dataset sizes of 100 samples up to a maximum of 10,000 and tested on a separate validation set of 10,000 samples, is displayed in Figure 6. For smaller training set sizes,

the model clearly overfits, as indicated by the accuracy being higher on the training set compared to the test set. However, as the training set size increases, the performance gap between the training and validation sets diminishes, indicating better generalization of the model to unseen data.

TABLE I: Each kernel's optimal hyperparameters & accuracy

Kernel Type	Accuracy	Best Parameters (C, γ)
POLY	0.8240	1.05, 1.0
RBF	0.8364	1.1, 1.0
SIGMOID	0.8365	1.0, 0.01

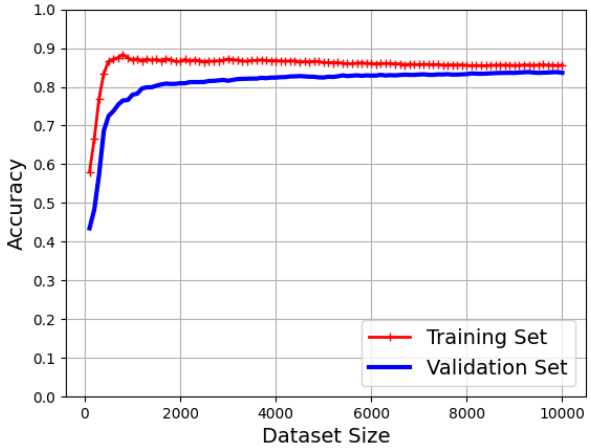

Fig. 6: The learning curve for the best small model identified during the model selection process.

B. Full Model

Based on the previous results, the full model was trained on a set of 150,000 samples using a sigmoid kernel with a regularization coefficient of 1 and a kernel coefficient of 0.01. The model achieved an accuracy of 0.84 on the test set of 25,000 samples, with the per-class statistics being presented in Table II.

TABLE II: The performance of the full model (trained on 150,000 samples), evaluated on the test set of 25,000 samples. The class labels are as follows: **C:** coil. **E:** sheet/strand. **H:** helix.

Class	Precision	Recall	f1-score	support
C	0.81	0.86	0.83	10998
E	0.85	0.77	0.81	5424
H	0.89	0.87	0.88	8578

VII. Discussion

A. Upper Limit on Accuracy

Secondary structure prediction is only tractable as a problem insofar as the secondary structure is well-defined between similar proteins. It is sometimes the case that two protein sequences are particularly similar in many respects, but the secondary structure differs in seemingly similar regions of residues. The most pertinent case is that of homologous proteins: proteins within different homology groups with seemingly close protein sequences can have substantially different secondary structures [49]. This is of particular interest, since homology information has been shown to be encoded in ProtT5's amino acid embeddings [50]. Results based on analysis of the secondary structure of homologous proteins as determined via protein sequence alignment and protein structure alignment [51] yield an upper bound of approximately 91.4% ($\pm$0.8%) accuracy for 3-class secondary structure prediction even in the general, non-homologous case [2]. This limit can be considered as a hard cap on performance for all models attempting to predict secondary structure which we can only aspire to, as it stems from intrinsic structural differences between similar proteins.

B. Interpretation of Results

The model performs well on the training and test data reaching an overall accuracy of 84% on the test set. With these results, we achieve a substantial improvement over previous SVM methods: this accuracy is in fact nearly competitive with the results of the state-of-the-art Porter 6 model. However, it is worth noting the Porter 6 model was trained on a more recent and larger dataset than what was used in our experiments. The lower recall of 0.77 for sheet/strand structures in particular may be partially explained by the imbalance in the datasets: in the training set, this class is rarer, while in the test set, it is significantly more common. Overall, given the impressive performance of the model, particularly under the non-ideal circumstances elaborated on further in the following section, we may consider the approach a success.

VIII. Limitations

Although the approach and results presented above are mostly sound overall, there are some limitations with regard to both the interpretation of the results and the generalization of the approach. Most of these issues stem from computational constraints and data quality.

A. Class Imbalance

The dataset used in our analysis has a substantial class imbalance, with helices and sheet/strand structures appearing at a rate of 2 to 1 in the training set. Notably, the test set presents a substantially different distribution of classes, with the rate of helices and sheet/strands being almost reversed compared to the training set. Although the class imbalance may add bias to the model, the low generalization error achieved in the results indicates that the model is robust with regard to this issue: however, we would still expect this factor to enact a reduction in the performance on the model.

B. Scaling Training Set

The dataset used in our experiments is fairly small, consisting of under 10,000 proteins. This was further cut-down by selecting a mere 150,000 embedding vectors to fit and test the model on. Based on the results seen for the various data sample sizes, there is a clear upward trend in the performance as the sample size increases, though the improvement does slow throughout this increase. However, it may not be tractable to push the performance to its limits due to the time complexity of the optimization problem solved in the process of fitting an SVM model scaling quadratically with the number of samples. The fundamental issue at hand is that the basic SVM algorithm is not well-suited to large datasets: we consider a workaround in the next section.

C. Non-I.I.D Random Variables

One concern with the validity of the model is the violation of the independent identically distributed (i.i.d) assumption intrinsic in much of machine learning theory. The problem stems from the source of the vectors passed into the SVM. The vector corresponding to each amino acid is generated via an embedding model which takes the source protein into account as part of the embedding algorithm. Note that ProtT5 offers no guarantee of a length-agnostic embedding (indeed, such an embedding would necessarily lose much of the context we wish to capture, given that it would be restricted to only considering amino acid interactions within a range of $\frac{n}{2}$, where n is the length of the shortest protein sequence in the training data for the embedding model). Therefore, it is logical to conclude that the distribution of the embedding vectors varies at least with respect to the length of the source protein. We offer two solutions to this uncertainty, each with its own disadvantages:

- **Source embedding vectors from many distinct protein sequences.** One way to eliminate potential bias in the model stemming from this type of problem is to introduce many different source proteins and select the embedding vectors in such a way as to take very few vectors from each protein. Taking M such proteins, let $f_i (1 \leq i \leq M)$ denote the PDF corresponding to the distribution of the embedding vectors sourced from each protein. The PDF of the distribution of individual samples of embedding vectors taken from a dataset sourced by taking n samples from each distribution is then given by that of the mixture distribution [52, Chapter 1] $f(x) = \frac{1}{M} \sum_1^M f_i(x)$. Taking embedding vectors from such a distribution yields identically distributed samples: if n is set to 1, we also achieve independence and thus the full i.i.d condition. Though theoretically sound, this approach fails in that it forces us to obtain labeled training data for a prohibitive quantity of proteins, given that to be fully rigorous we require one protein per embedding vector. Even if we are willing to settle for 'near-independence' by allowing $n > 1$, we are still restricted by the length of the shortest protein sequence we consider. These lengths can be as low as 20, and so we are restricted to at most 20 samples per protein, still far too few to produce a reasonably sized dataset.)
- **Construct a specific model for each length of protein separately.** This approach allows us to remove the differences in distribution stemming from the distinct lengths completely. However, it faces the same issue as the previous approach with respect to independence and taking multiple samples from the same protein. Moreover, we are now faced with a critical shortage of data, as some lengths of protein are quite rare and would not be feasible to build a model for at all.

Note that the above only serves to answer a very specific issue regarding the potential distribution shift within the model's training data. There are also likely to be other differences in distribution between proteins unrelated to the length, based on factors such as the frequency of certain amino acids within the sequence. These factors may contribute to the gap between the performance of our model and the generally accepted 91% upper limit. On the other hand, it should be noted that the violation of the i.i.d assumption outlined above is common among any model trained on protein embeddings. For example, the sliding window approach commonly used in similar problems, as employed by the two SVM methods discussed in the Related Work section, suffers from the same limitation. The reason for the repetition of the violation is clear: the model must have access to contextual information for each amino acid so as to be able to encode anything useful at all. As such, given the overall performance relative to other models, we should not ascribe too much weight to this mainly theoretical issue.

IX. Future Work

A. Scaling Training Set

Training an SVM on large datasets can be computationally expensive and memory intensive. However, once trained, an SVM can be used for protein secondary structure prediction with reduced computational overhead, ultimately leading to more efficient and scalable methods for protein structure prediction, and integration in other bioinformatics pipelines. Therefore, it is worth exploring avenues into expanding the training dataset further. Since extra training samples yield diminishing returns in terms of performance, there is a critical point beyond which adding more data is no longer worth the investment. Based on the rate of increase in the learning curves presented in the previous sections, we have not reached this point yet. One method particularly worth exploring is approximate extreme point training [27], in which a carefully selected subset of the samples is used for the support vector optimization rather than the full dataset: this could potentially drastically reduce the training time of the model without compromising the performance.

B. Data Quality Improvement

A central limitation surrounding our methods stems from the small and imbalanced dataset used in the construction of

the model. Recent approaches such as Porter 6 have used datasets that have been curated to yield far higher quality training and test sets than what was used in our experiments. Due to computational and time constraints, we were unable to generate a better-suited dataset for our investigation. The gap in performance between our model and Porter 6's state-of-the-art performance is small, and we hypothesize that the margin may be closed by employing techniques such as the application of redundancy reduction thresholds to construct a larger, more balanced, and representative dataset for effective comparison against state-of-the-art methods in protein secondary structure prediction.

C. Eight-State Classification

Finally, given the success of our approach when applied to the easier three-state secondary structure classification problem, a natural way to extend the results is by applying the same methods (perhaps with the enhancements outlined above) to the more difficult eight-state classification problem. The long-term objective is to present a simple, accurate, and theoretically well-understood model to the protein secondary structure prediction problem in all its forms.

APPENDIX A
CODE AVAILABILITY AND DATA

- Github repository with all code available
- Dataset used in investigation

REFERENCES

[1] W. Alanazi, D. Meng, and G. Pollastri. Porter 6: Protein secondary structure prediction by leveraging pre-trained language models (plms). *International Journal of Molecular Sciences*, 26(1):130, 2025. doi:10.3390/ijms26010130.

[2] Chia-Tzu Ho, Yu-Wei Huang, Teng-Ruei Chen, Chia-Hua Lo, and Wei-Cheng Lo. Discovering the ultimate limits of protein secondary structure prediction. *Biomolecules*, 11(11), November 2021.

[3] Bartelby. Amino acids which do not have any charge on them are neutral amino acids. https://www.bartleby.com/subject/science/chemistry/concepts/neutral-amino-acids, 2021. [Accessed 31-12-2024].

[4] Ahmad Alali. *Investigations on cyclization steps and third ring oxygenation in rishirilide biosynthesis*. PhD thesis, 01 2020.

[5] Wikimedia Commons. File:22 amino acid sequence (n-terminal).png — wikimedia commons, the free media repository, 2023. [Online; accessed 31-December-2024].

[6] R. F. Doolittle. Similar amino acid sequences: Chance or common ancestry? *Science*, 214(4517):149–159, 1981. doi:10.1126/science.7280696.

[7] J. Pei and N. V. Grishin. Al2co: Calculation of positional conservation in a protein sequence alignment. *Bioinformatics*, 17(8):700–712, 2001. doi:10.1093/bioinformatics/17.8.700.

[8] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970. doi:10.1016/0022-2836(70)90057-4.

[9] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990. doi:10.1016/S0022-2836(05)80360-2.

[10] Z. Li and H. A. Scheraga. Monte carlo-minimization approach to the multiple-minima problem in protein folding. *Proceedings of the National Academy of Sciences*, 84(12):4000–4004, 1987. doi:10.1073/pnas.84.12.4000.

[11] Wikimedia Commons. File:alpha beta structure (full).png — wikimedia commons, the free media repository, 2023. [Online; accessed 31-December-2024].

[12] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000. doi:10.1093/nar/28.1.235.

[13] Longxing Cao, Brian Coventry, Inna Goreshnik, Buwei Huang, William Sheffler, Joon Sung Park, Kevin M. Jude, Iva Marković, Rameshwar U. Kadam, Koen H. G. Verschueren, Kenneth Verstraete, Scott Thomas Russell Walsh, Nathaniel Bennett, Ashish Phal, Aerin Yang, Lisa Kozodoy, Michelle DeWitt, Lora Picton, Lauren Miller, Eva-Maria Strauch, Nicholas D. DeBouver, Allison Pires, Asim K. Bera, Samer Halabiya, Bradley Hammerson, Wei Yang, Steffen Bernard, Lance Stewart, Ian A. Wilson, Hannele Ruohola-Baker, Joseph Schlessinger, Sangwon Lee, Savvas N. Savvides, K. Christopher Garcia, and David Baker. Design of protein-binding proteins from the target structure alone. *Nature*, 605(7910):551–560, May 2022.

[14] W. L. Jorgensen. The many roles of computation in drug discovery. science. *New York, N. Y.)*, 303(5665):1813–1818, 2004. doi:10.1126/science.1096361.

[15] Alaina S DeToma, Samer Salamekh, Ayyalusamy Ramamoorthy, and Mi Hee Lim. Misfolded proteins in alzheimer's disease and type ii diabetes. *Chemical Society Reviews*, 41(2):608–621, 2012.

[16] K. S. Bhullar, N. O. Lagarón, E. M. McGowan, I. Parmar, A. Jha, B. P. Hubbard, and H. P. V. Rupasinghe. Kinase-targeted cancer therapies: progress, challenges and future directions. *Molecular cancer*, 17(1):48, 2018. doi:10.1186/s12943-018-0804-2.

[17] Victor A Simossis and Jaap Heringa. Praline: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic acids research*, 33(suppl_2):W289–W294, 2005.

[18] Jisna Vellara Antony, Roosafeed Koya, Pulinthanathu Narayanan Pournami, Gopakumar Gopalakrishnan Nair, and Jayaraj Pottekkattuvalappil Balakrishnan. Protein secondary structure assignment using residual networks. *Journal of Molecular Modeling*, 28(9):269, Aug 2022.

[19] B. E. Shakhnovich, E. Deeds, C. Delisi, and E. Shakhnovich. Protein structure and evolutionary history determine sequence space topology. *Genome research*, 15(3):385–392, 2005. doi:10.1101/gr.3133605.

[20] RCSB Protein Data Bank. Pdb statistics: Overall growth of released structures per year.

[21] R. Aebersold and M. Mann. Protein sequence analysis: The mass spectrometry approach. *Nature*, 422(6928):198–207, 2003. doi:10.1038/nature01517.

[22] W. Kabsch and C. Sander. Dssp: A software for the description of secondary structure of proteins. *Journal of Applied Crystallography*, 16(5):778–785, 1983. doi:10.1107/S0021889883018333.

[23] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28, 1998.

[24] M. N. Murty and Rashmi Raghava. *Kernel-Based SVM*, pages 57–67. Springer International Publishing, Cham, 2016.

[25] Arti Patle and Deepak Singh Chouhan. Svm kernel functions for classification. In *2013 International Conference on Advances in Technology and Engineering (ICATE)*, pages 1–9, 2013.

[26] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, 408:189–215, 2020. doi:10.1016/j.neucom.2019.10.118.

[27] Manu Nandan, Pramod P Khargonekar, and Sachin S Talathi. Fast svm training using approximate extreme points. *The Journal of Machine Learning Research*, 15(1):59–98, 2014.

[28] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 1532–1543, 2014. https://aclanthology.org/D14-1162.

[29] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2019)*, pages 4171–4186, 2019. doi:10.18653/v1/N19-1423.

[30] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2022.

[31] Konstantin Weissenow, Michael Heinzinger, and Burkhard Rost. Protein language-model embeddings for fast, accurate, and alignment-free protein structure prediction. *Structure*, 30(8):1169–1177.e4, May 2022.

[32] L. H. Holley and M. Karplus. Protein secondary structure prediction with a neural network. *Proceedings of the National Academy of Sciences of the United States of America*, 86(1):152–156, 1989.

[33] S. Hua and Z. Sun. A novel method of protein secondary structure prediction with high segment overlap measure: Support vector machine approach. *Journal of Molecular Biology*, 308(2):397–407, 2001. doi:10.1006/jmbi.2001.4614.

[34] J. J. Ward, L. J. McGuffin, B. F. Buxton, and D. T. Jones. Secondary structure prediction with support vector machines. *Bioinformatics*, 19(13):1650–1655, 2003. doi:10.1093/bioinformatics/btg212.

[35] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast: A new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997. doi:10.1093/nar/25.17.3389.

[36] David W Mount. Using the basic local alignment search tool (blast). *Cold spring harbor Protocols*, 2007(7):pdb–top17, 2007.

[37] R. M. Rao, J. Meier, and J. Liu. et al. Technical report, ESM-2: A general-purpose protein sequence model for downstream tasks. bioRxiv, 2022. doi:10.1101/2022.04.29.489189.

[38] J. Singh, K. Paliwal, T. Litfin, J. Singh, and Y. Zhou. Reaching alignment-profile-based accuracy in predicting protein secondary and tertiary structural properties without alignment. *Scientific reports*, 12(1):7607, 2022. doi:10.1038/s41598-022-11684-w.

[39] M. S. Klausen, M. C. Jespersen, H. Nielsen, K. K. Jensen, V. I. Jurtz, C. K. Sønderby, M. O. A. Sommer, O. Winther, M. Nielsen, B. Petersen, and P. Marcatili. Netsurfp-2.0: Improved prediction of protein structural features by integrated deep learning. proteins: Structure, function. *and Bioinformatics*, 87(6):520–527, 2019. doi:10.1002/prot.25674.

[40] M. Torrisi, M. Kaleel, and G. Pollastri. Deeper profiles and cascaded recurrent and convolutional neural networks for state-of-the- art protein secondary structure prediction. *Sci. Rep.*, 9:12374, 2019.

[41] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, M. Zielinski, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021. doi:10.1038/s41586-020-03049-6.

[42] L. M. F. Bertoline, A. N. Lima, J. E. Krieger, and S. K. Teixeira. Before and after alphafold2: An overview of protein structure prediction. *Frontiers in bioinformatics*, 3, 2023. doi:10.3389/fbinf.2023.1120370. Article 1120370.

[43] C. A. Rohl, C. E. M. Strauss, K. M. S. Misura, and D. Baker. Protein structure prediction using rosetta. *Methods in Enzymology*, 383:66–93, 2004. doi:10.1016/S0076-6879(04)83004-0.

[44] A. Sali and T. L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 234(3):779–815, 1993. doi:10.1006/jmbi.1993.1626.

[45] Tamzid Hasan. Protein secondary structure casp12 cb513 ts115, 2021. Retrieved 1/12/2024. https://www.kaggle.com/datasets/tamzidhasan/protein-secondary-structure-casp12-cb513-ts115.

[46] Home - CASP12 — predictioncenter.org. https://www.predictioncenter.org/casp12/index.cgi. [Accessed 30-12-2024].

[47] J A Cuff and G J Barton. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins*, 34(4):508–519, March 1999.

[48] Yuedong Yang, Jianzhao Gao, Jihua Wang, Rhys Heffernan, Jack Hanson, Kuldip Paliwal, and Yaoqi Zhou. Sixty-five years of the long march in protein secondary structure prediction: the final stretch? *Brief Bioinform*, 19(3):482–494, May 2018.

[49] Olga Bagrova, Ksenia Lapshina, Alla Sidorova, Denis Shpigun, Aleksey Lutsenko, and Ekaterina Belova. Secondary structure analysis of proteins within the same topology group. *Biochemical and Biophysical Research Communications*, 734:150613, 2024.

[50] Kamil Kaminski, Jan Ludwiczak, Kamil Pawlicki, Vikram Alva, and Stanislaw Dunin-Horkawicz. plm-blast: distant homology detection based on direct comparison of sequence representations from protein language models. *Bioinformatics*, 39(10):btad579, 09 2023.

[51] P. E. Bourne and I. N. Shindyalov. Structure comparison and alignment. In P. E. Bourne and H. Weissig, editors, *Structural Bioinformatics*. Wiley-Liss, Hoboken NJ, 2003. ISBN: 0-471-20200-2.

[52] Sylvia Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Springer New York, 2006.