Springboard Capstone 2: Predicting features of Circular Dichroism spectra from a protein’s amino acid sequence

Diana Koulechova | January 2024

Background & Project Motivation

Proteins are the “molecular machines” of the body, and just like with the machines we encounter in our daily lives, form determines functions. The form, or structure, that a protein adopts is dictated by the identities of the amino acids that are strung together like beads on a string. Here we will focus on the localized interactions between these amino acids, especially α-helices and β-sheets (the two main types), illustrated below in Figure 1:

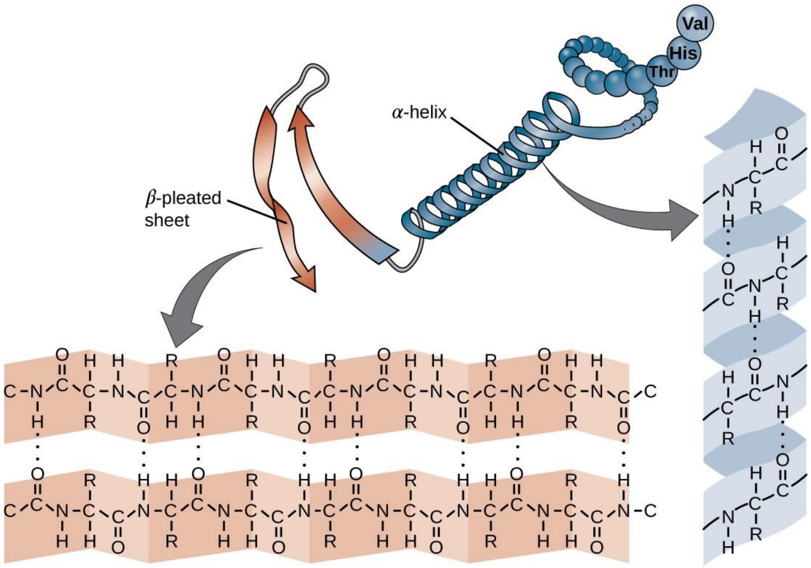


Figure 1. Illustration of protein secondary structures from https://www.nursinghero.com/study-guides/microbiology/proteins.

Circular Dichroism (CD) is a laboratory technique that can be used to detect the presence of α-helices and β-sheets. The Protein Circular Dichroism Data Bank (PCDDB) is a public repository CD spectra that contains datasets used for the development of a number of the prominent deconvolution algorithms, which take a CD spectrum as input and estimate the proportion of the protein composed of α-helices or β-sheets.

Here, we were interested in starting with the amino acid sequence and seeing if we could predict spectral features.

Feature Engineering

The entire PCDDB database has fewer than 1,000 entries. This means that I was unable to model an output as complex as the entire CD spectrum. Instead, I classified the spectra as indicating the presence of α-helices if they had a minimum between 220 – 224 nm or β-sheet if they had a minimum between 217 – 219 nm. Spectra were also classified as to whether they had negative values at 195 nm, indicative of a disordered polypeptide chain. While those results will be shown, there are very few examples of the positive class and thus any results for random coil should be taken with a very large grain of salt.

The amino acid sequence was similarly too complex to use as an explanatory variable. Instead, I used AAIndex (details are described at https://www.genome.jp/aaindex/) to convert the amino acid sequences to numerical vectors using indices corresponding to certain physiochemical properties. The initial subset of indices chosen is detailed in Table 1.

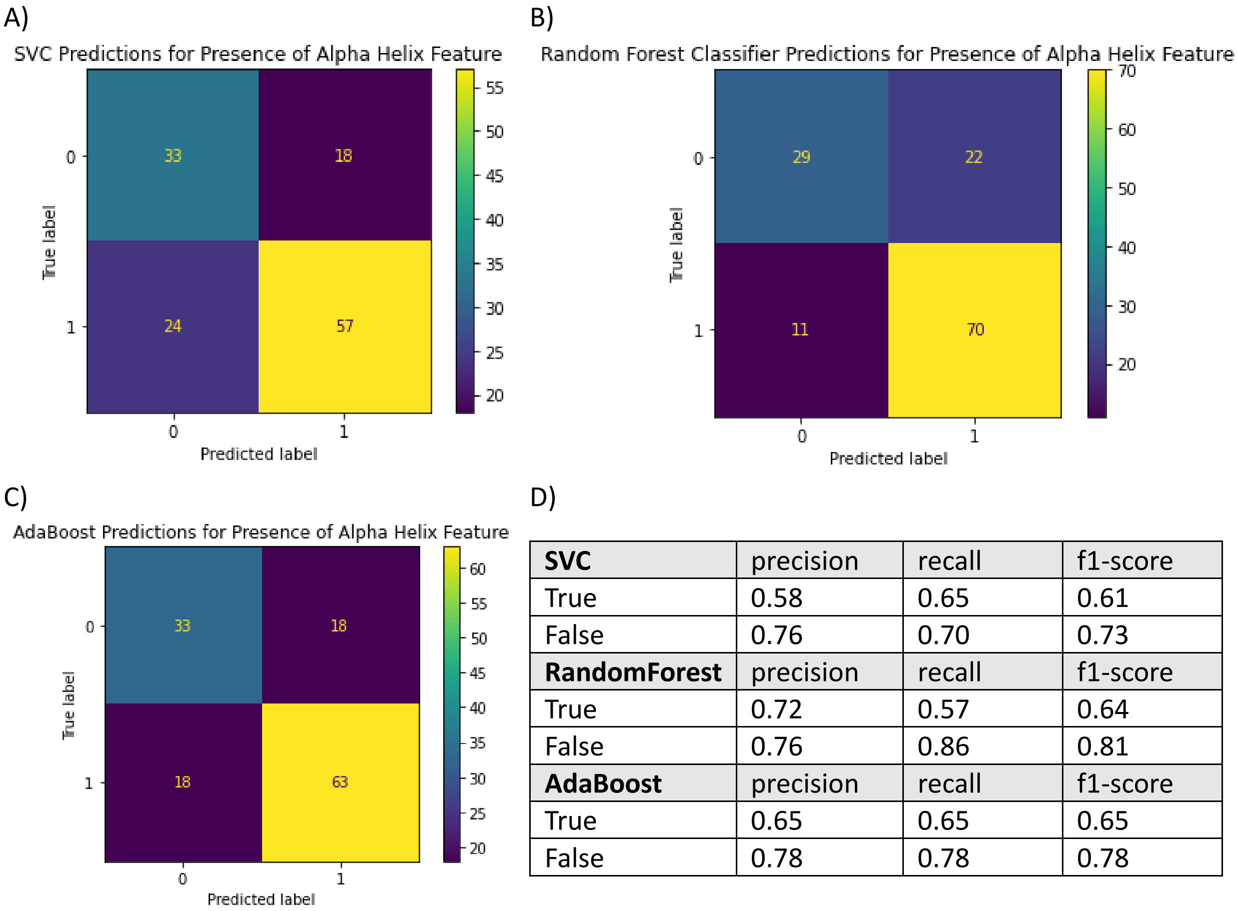
Table 1. AAIndex indices chosen for numerical conversion of protein sequences.

Property Name	Property Description	AAIndex Access ID	Reference
size	Residue volume	BIGC670101	Bigelow, 1967
partition_coeff	Partition coefficient	GARJ730101	Garel et al., 1973
hydrophobicity	Hydrophobicity index	ARGP820101	Argos et al., 1982
flexibility	Average flexibility indices	BHAR880101	Bhaskaran-Ponnuswamy, 1988
chemical_shifts	alpha-CH chemical shifts	ANDN920101	Andersen et al., 1992
K_helixcoil	Helix-coil equilibrium constant	FINA770101	Finkelstein-Ptitsyn, 1977
phi	Side chain torsion angle phi (AAAR)	LEVM760104	Levitt, 1976

After converting each primary sequence into a numerical vector, they were further reduced to a single number by finding either the mean or sum of the numbers.

Identifying a reasonable model and hyperparameters

During exploratory data analysis, three classifiers were found to show the most promise: Support Vector, Random Forest, and AdaBoost. Using either a random or grid search, depending on the number of hyperparameters in question, the best hyperparameters were identified for each and the three models then compared side by side for their ability to accurately predict whether a protein's spectrum would exhibit α-helical indicators. Results are summarized in Figure 2. All three appear quite similar; Random Forest was chosen as the model to proceed further as it may have a slight advantage over the other two.



SVC	precision	recall	f1-score
True	0.58	0.65	0.61
False	0.76	0.70	0.73
RandomForest	precision	recall	f1-score
True	0.72	0.57	0.64
False	0.76	0.86	0.81
AdaBoost	precision	recall	f1-score
True	0.65	0.65	0.65
False	0.78	0.78	0.78

Figure 2. Confusion matrices (A – C) and model metrics (D) for three classifiers.

Can the model lend insights into the physiochemical determinants of secondary structure?

The first question to ask is how reducing the sequence numerical vectors to a single number by averaging compares to doing so by adding. If we extract the feature importance values for our best model, plotted in Figure 3, we find that, with some minor and one glaring (the helix-coil equilibrium constant) exceptions, the sum seems to matter slightly more than the mean. Importantly, all of the values for feature importance are relatively small and roughly evenly distributed among the different indices.

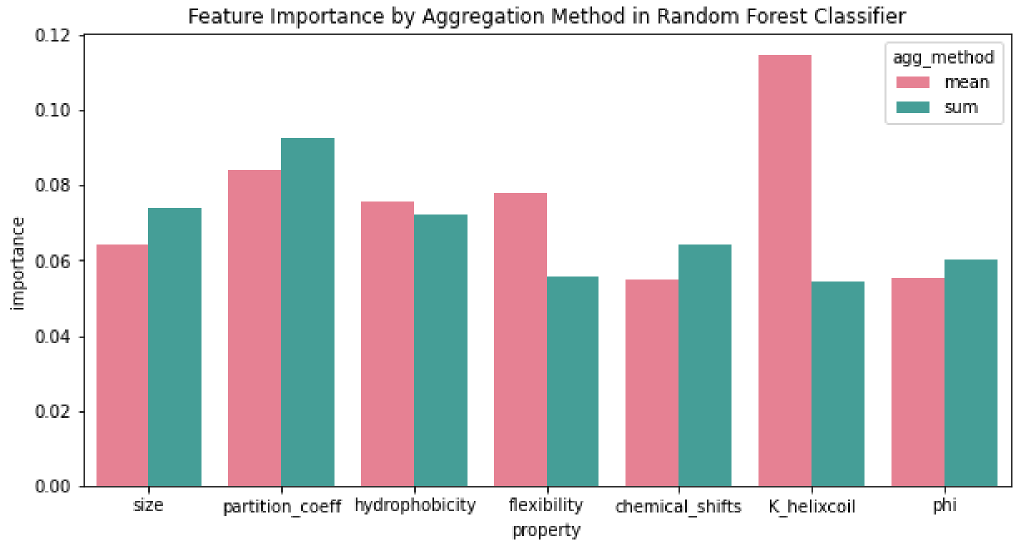


Figure 3. Comparing the contribution to the model of the mean and sum of each amino acid index conversion.

While there is no apparent single winner for explanatory variable when looking at predicting α-helices, the next step was to see if this would be true for all three of the secondary structure elements being assessed. These results are summarized in Figure 4.

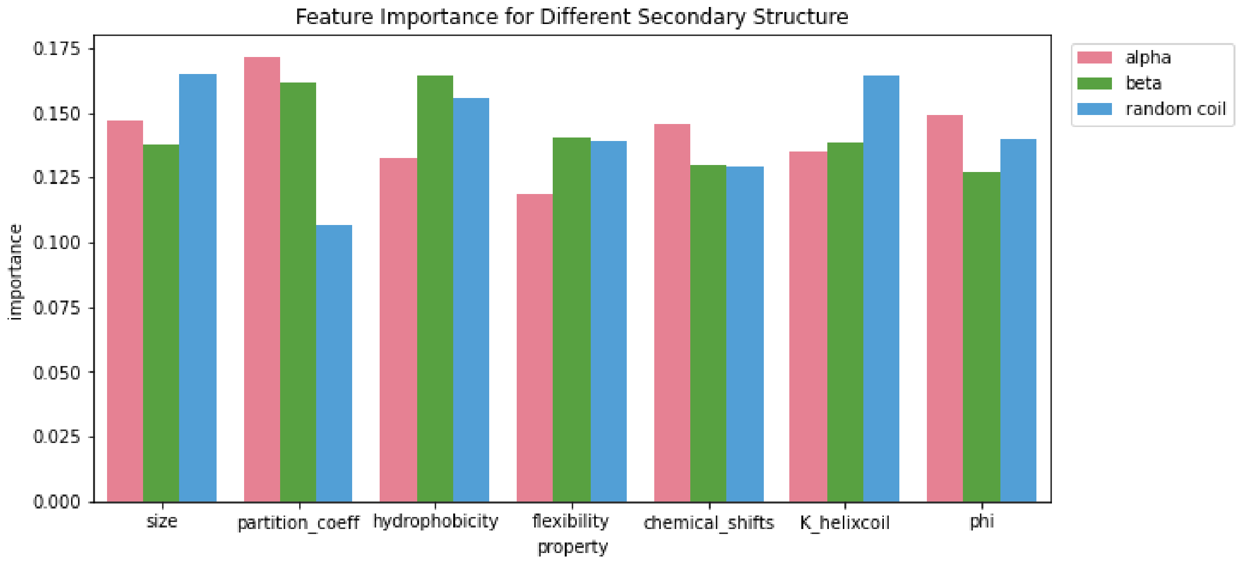
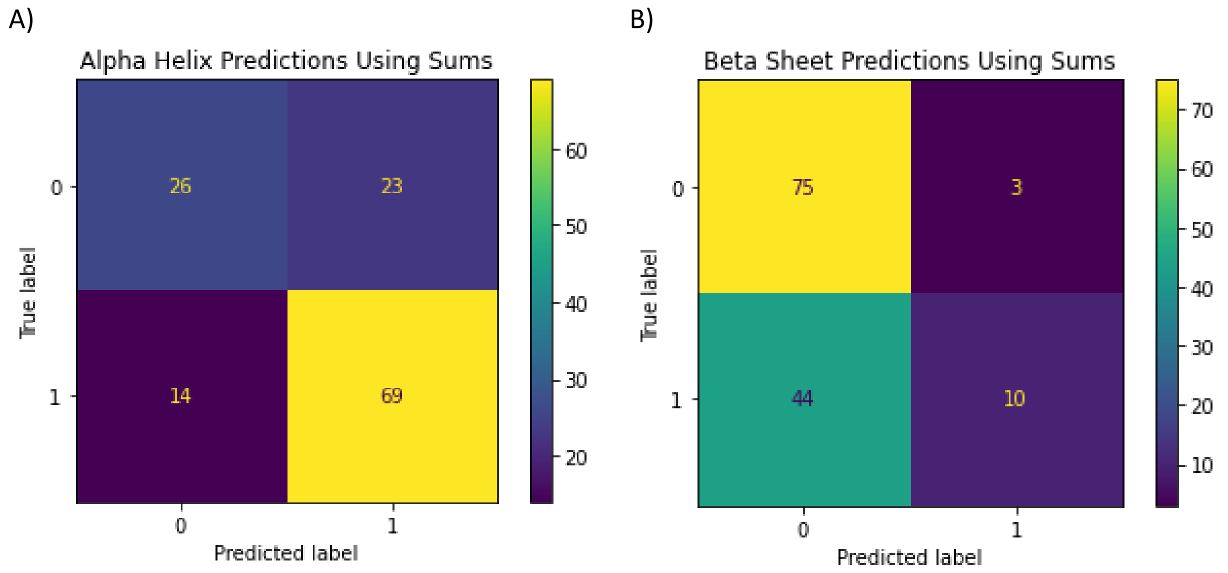


Figure 4. Comparing the contributions of different physiochemical properties to the prediction of three types of secondary structure elements.

While there is no one clear winner, it does seem like hydrophobicity and the closely related partition coefficient are relatively strong across types. As a final evaluation, the predictive power of the sum of the hydrophobicity index values of a protein's amino acid sequence was assessed. As shown in Figure 5A,C below, the confusion matrix for α-helical features looks almost identical to that achieved when using all of the physiochemical indices as predictors (Figure 2B). The metrics for β-sheet prediction look significantly less robust (Figure 5B,C), although the sizes of the classes are comparable between the two secondary structure types. While of-course disappointing, this is consistent with the larger and more qualitatively obvious signal observed during examination of individual spectra for α helices relative to β sheets.



C)

α-helix	precision	recall	f1-score
False	0.65	0.53	0.58
True	0.75	0.83	0.79
β-sheet	precision	recall	f1-score
False	0.63	0.96	0.76
True	0.77	0.19	0.30

Figure 5. Confusion matrices for model outcomes when sum of hydrophobicity index-converted amino acid sequence is used as sole explanatory variable for the presence of spectral elements indicating the presence of α-helix (A) or β-sheet (B), along with the metrics for both (C).

Next Steps

For general improvement of any CD-related algorithms, more – and more varied – spectral data is the clear necessity. A more targeted – and thus less experimentally intensive – approach can be taken by using the predictions and observations here to generate protein variants comprising hydrophobic sum series and measuring their CD spectra.

If additional laboratory data are either not an option or not desired, comparing different hydrophobic indices, comparing them with varying partition coefficient indices, or doing a deep dive with the mean of helix-coil constants (which appears to be an outlier in Figure 3) are all potential algorithmic avenues.