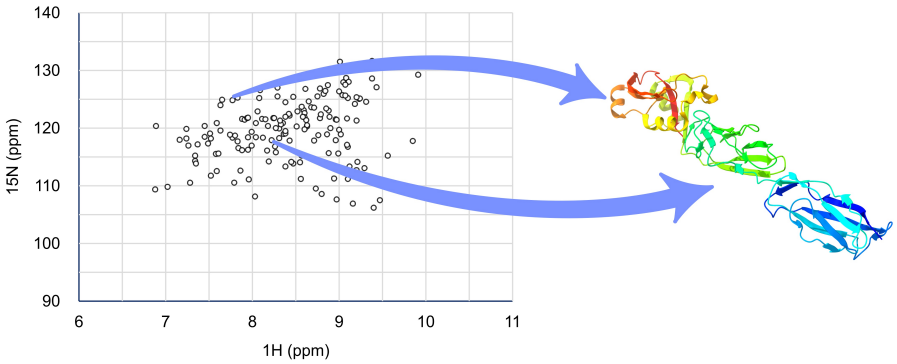
Testing the Potential of Secondary Structure Content Prediction of Proteins from N-HSQC Spectra



Jonas Dietrich
jonas.dietrich@uni-jena.de

Reasearch Internship in MC3.1.1 Analytical Chemistry

First Supervisor **Prof. Dr. Steinbeck**
Research Group **Steinbeck group**
Departement **Analytical Chemistry**
University **Friedrich Schiller University Jena**

Second Supervisor **Dr. Peter Bellstedt**
Research Group **Bellstedt lab**
Departement **Clinical institute**
University **Zurich University**

March 13, 2023

Contents

1 Introduction

Secondary structure prediction is a crucial step in understanding the structure of proteins and their function. Current de novo models based on primary structure inputs (such as Alpha Fold [1]) are not suitable for all structure types [2] or dynamic and environment-dependent conformations [3]. The most accurate results can be obtained by spectroscopic determination of protein structure, which is a time-consuming process.

A widely used spectroscopic method for protein structure analysis is nuclear magnetic resonance (NMR) spectroscopy. In NMR-based protein structure determination, various types of NMR spectra (HSQC, NOESY, HMBC, etc.) are measured, from which the protein structure can be reconstructed [4] [5]. This procedure is not only very time consuming, but also requires a high level of expertise on the part of the researchers. A particularly useful spectrum for NMR determination is the N-HSQC spectrum, which is relatively easy to measure and allows the study of proteins ranging from small size to whole protein complexes [6]. In addition, the N-HSQC spectrum provides a unique pattern of peaks for each individual protein and is therefore also known as a fingerprint spectrum. These two advantages make N-HSQC spectra a perfect tool for secondary structure prediction based on real measurement data. The use of measured data for structure prediction is expected to increase robustness on the one hand and is also suitable for the prediction of dynamic, environment-dependent as well as unknown protein structure types on the other hand [7] .

The earliest article found implementing this idea was published in 2006 by V.H. Moreau et. al. under the name "Prediction of the amount of secondary structure of proteins using unassigned NMR spectra: a tool for target selection in structural proteomics" in the journal Genetics and Molecular Biology [8]. Here, the authors used unassigned HSQC spectra to predict the three-state secondary structure elements of proteins. Based on 72 proteins from the Protein Data Bank (PDB) [9], spectra obtained from the Biological Magnetic Resonance Bank (BMRB) [10] were divided into 10x10 grids. The number of peaks measured in each quadrant was used to predict the secondary structure by a multiple linear regression model. To improve the prediction performance, only quadrants that had an absolute correlation higher than 0.3 were used as input to the model. Using only N-HSQC spectra resulted in 70% accuracy for the alpha helix and 71% for the beta sheet, while the coil content was calculated by taking the difference of the sum of the two secondary structure predictions to 1. In non-mathematical language, the authors argue that if the secondary structure is neither an alpha helix nor a beta sheet, it must be

a coil.

Much progress has been made in the field of structure prediction since the findings of Moreau et. al. in 2006 [11], and the size of NMR and protein databases has grown steadily since then, currently (11/2022) standing at about 8257 PDB-to-BMRB links. Over the past decade, protein structure prediction used template or template free based predictions methods [12] [13]. Advances in machine learning and deep learning combined with huge amounts of data collected by crystallographic (and NMR) measurements have made it possible to obtain very good results [14]. However, there is still a lack of an updated version of a structure elucidation for protein secondary structures that incorporates both statistical methods and protein measurements to combine the best of both worlds by using the accuracy of NMR experiments and the speed of statistical models.

The goals of this project are twofold. First, to reproduce the results of Moreau et. al. and to investigate the influence of the amount of data on the predictive performance. The completed model can therefore be considered an important update after 16 years and can also be used as a basis for predicting the three-state secondary structure by N-HSQC spectra. The second goal is to demonstrate the potential of predicting structures based on N-HSQC and to open the door for future follow-up projects that extend the models implemented here with more advanced machine learning methods as well as Deep Learning or even the prediction of eight-state secondary structures. By confirming the results of Moreau et. al. or improving them with the extended dataset, this project could demonstrate the potential of protein secondary structure prediction and promote the use of prediction models based on measured data to improve the robustness of protein structure predictions.

2 Material and Methods

2.1 Data acquisition

The goal of data collection was to create three files containing the peak list, secondary structure information, and measurement conditions.

Protein IDs were downloaded from the file "Matched submitted BMRB-PDB entries in one deposition session" (`https:/bmrb.io/search/`, On GitHub: BMRB_PDB_ids.csv) from the BMRB website. 8257 entries were found (11/2022). Duplicate BMRB IDs were removed such that only the BMRB ID that occurred first remained in the dataset. With the new dataset without duplicates, all N-HSQC backbone spectra were downloaded from the BMRB website. Next, the secondary structure information of the proteins was downloaded from a server running the program DSSP [15] [16]. The final step of data acquisition was to download the measurement conditions (temperature, pressure, pH) from the NMRStar file or PDB database. On first review, the NMRStar file showed clearer results and more consistency in the use of units and was therefore used first for the measurement conditions. In addition to NMRStar2.0, NMRStar3.0 files were also used. The missing measurement conditions were filled by the found conditions from the PDB database. Before the actual filtering step, all entries without N-HSQC spectra, secondary structure information, or not all three measurement conditions were removed from the data set in advance.

2.2 Data Filtering

Because the 72 proteins used in the reference article were not specified with PDB ID, I used the submission date of the paper as a filter so that all entries with a submission date in the BMRB database before 2005/09/25 were also saved in a separate file to create a second data set.

The data were filtered based on five criteria: chemical shifts, temperature, pressure, pH, and secondary structure. According to the figures in the Moreau et. al. publication, no protein has peaks with N-shift greater than 140 ppm and less than 90 ppm. The upper limit for H-shifts was 11 ppm and the lower limit was 6 ppm. All proteins with peaks outside these ranges were removed from the data set. The temperature, pressure, and pH properties were not mentioned in the article, but apparently reasonable limits were set to ensure high quality of the data set (Tab. 1). The temperature range was from 273 K to 310 K, and thus from nearly frozen to body temperature, ensuring that no denatured proteins were included [17]. For pressure, only measurements at atomspheric pressure (ATM) were used to ensure the same measurement

conditions for all samples. The pH range was from 5 to 8, again to have the proteins in their usual environment and highest stability [18] [19].

Only the cases where all filters were successfully applied were included in the data set.

Table 1: Used filter setting for the data preprocessing step.

Filter	Lower Limit	Upper Limit
Temperature	273 K	310 K
Pressure	1 bar	1 bar
pH	5	8
N-Shift	90 ppm	140 ppm
H-Shift	6 ppm	11 ppm

A special filter was used for the secondary structure elements. The three-level secondary structure consisted of 4 different structures: helix, sheet, coil and polyproline helix. The DSSP translation table for the eight-level to the three-level secondary structure can be found in Tab. 2.

Table 2: Translation table from eight to three-state secondary structure.

Appearance overall (sum = 651135)	Appearance proteins (sum = 5803)	8(9)-state	3(4)-state
162591	5801	-	C
12053	2396	G	H
67781	5592	T	C
117178	3903	E	E
79357	5657	S	C
196269	4748	H	H
5807	2195	B	E
7073	1828	P	P
3026	470	I	C

Since secondary structure calculation tools are not mentioned in Moreau et al. but the prediction of the three-level secondary structures helix, sheet, and coil, all proteins with polyproline helix were filtered out from the dataset. After the filtering process, the number of entries in the database was 2841 and 401 with the deposition date filter applied.

2.3 Binning, Secondary Structure and Quadrant Correlation

The binning was done using the previously established filter limits for N-shifts and H-shifts of 90 ppm to 140 ppm and 6 ppm to 11 ppm, respectively. The spectra were then divided into 10x10 equally spaced quadrants, with each quadrant covering a range of 5 ppm by 0.5 ppm. Within each quadrant, the number of peaks was counted and stored in a matrix.

For secondary structure content, the percent occurrence of helix, sheet, and coil was calculated for each protein.

For the quadrant correlation calculation, 100 new data sets were created, with each data set corresponding to a specific quadrant position within the original 10x10 matrices. For each new data set and corresponding secondary structure content, the correlation coefficient was calculated, resulting in a correlation coefficient for each quadrant in the N-HSQC spectra.

2.4 Model training

Secondary structure prediction was performed using two different approaches and two different training data sets, corresponding to four different linear regression models. One approach used as input only quadrants that had a correlation coefficient higher than 0.3 or lower than -0.3 for any secondary structure element, as done by Moreau et. al. In the second approach, all quadrants were used regardless of their correlation coefficient. For both approaches, the number of peaks in the quadrant was normalized. Then, a 100 times 10-fold cross-validation was used to create 1000 models, each of which predicted the test data set with 488 protein entries. An additional change was made to the models trained on the small data set consisting only of protein entries prior to 2006. Instead of using all 401 protein entries, during cross-validation 72 proteins were randomly selected for the model training to mimic the model from the reference article.

Model creation, filtering, and downloading were set up and performed using the Python libraries sklearn for linear regression and cross-validation [20] and scipy.stats for metrics calculation [21]. All downloaded data are freely available online in the BMRB and PDB databases. Due to request problems on the DSSP online server, a separate server was initialized and used to download the secondary structure content. The codes used to carry out this work is availably on github at joaldi2208/HelixWizard.

3 Results

3.1 Data Sets

Three data sets were created based on the filtered 2841 data entries. These include a small data set to replicate the results from the reference article, a large data set with all available data to compare the effects of an increasing amount of data, and a test data set to test the linear models.

First, the small data set was formed using the deposition date of the spectra in the BMRB database. All spectra with a deposition date before 2005/09/25, the date of publication of the reference article, were included in the data set. The number of entries in the small data set was 401 proteins. The remaining 2440 were split into training and test data using the sklearn train_test_split function in a ratio of 80/20. The 20% portion represented the test data set with 488 protein entries. The large training dataset was formed by merging the 80% portion and the small training dataset. This data set contained 2353 protein entries. This method attempted to maximize the number of entries and diversity of the small data set because the proteins used are not known from the reference article.

3.1.1 Measurement conditions

The H-shift distributions show a narrow distribution for both the small and test data sets (Figure 1). While the test data set shows a symmetric distribution, the small data set shows a small shoulder around 9 ppm and has a higher density at about 8.3 ppm, which is the mean for both the small and test data sets. Compared to the two smaller data sets, the large data set is slightly shifted to the lower field and has a wider distribution. The mean value is 8.8 ppm and the distribution is also skewed to the right, indicating that the density of values above 8.8 ppm is higher.

For the N-shift distribution, again the small and test data sets show a narrow distribution compared to the large data set, while the broad distribution from the large data set is right skewed and, as with the H-shift distribution, has a higher density in the low field (Figure 2). The mean value is 125 ppm. For the small data set, one can see a shoulder in the low field around 128 ppm, while the test data set shows a flat area in the high field around 105 ppm. Both data sets have a similar mean around 122 ppm.

The temperature distribution shows mainly entries around the three most frequent temperature values (Fig. 3). First, most of the data points are at room temperature (297 K), second and third at 293 K and 303 K, respectively.

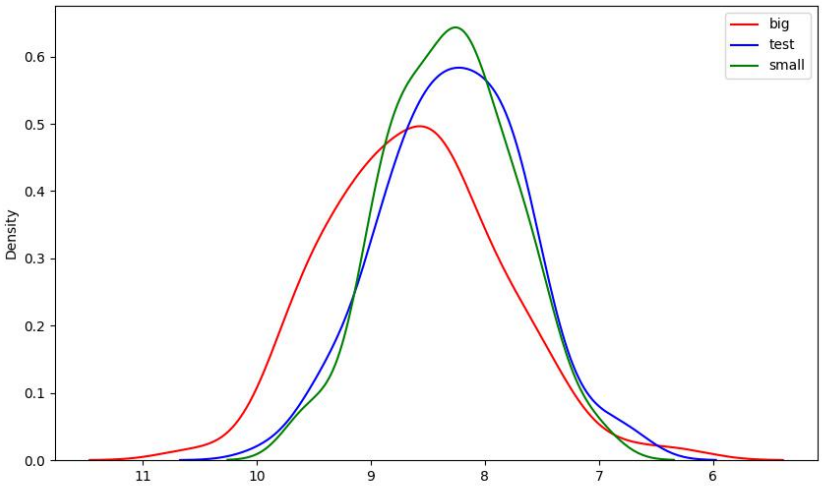


Figure 1: H-shift distribution for the N-HSQC spectra for the two training sets (red = 2353 proteins, green = 401 proteins) and the test set (blue = 488 proteins).

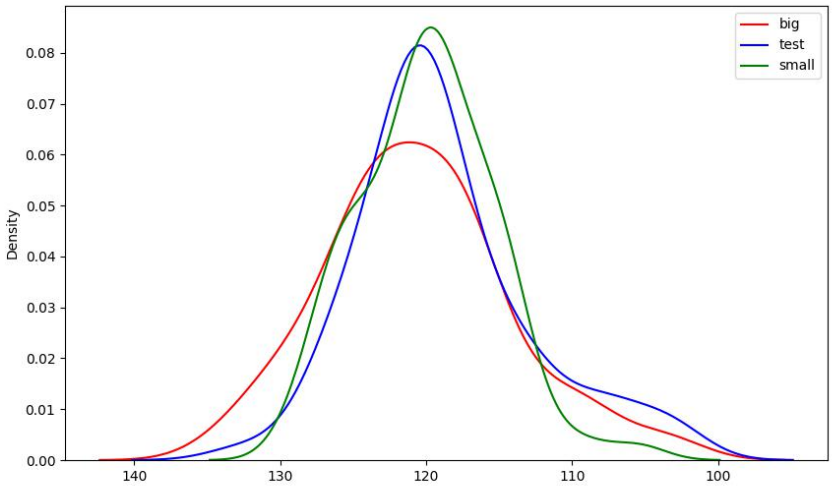


Figure 2: N-shift distribution for the N-HSQC spectra for the two training sets and the test set.

Overall, the distribution does not show much difference between the three data sets.

Next, the pH distribution shows a similar distribution between the small and large data set, but the distribution of pH values from the test data set is more broadly distributed across mainly three pH values (6.0, 6.5, and 7.5) (Fig. 4). Overall, the test data set shows a higher density toward the small pH values compared to the other two data sets. The small and large data sets are very similar in their distribution. The main density is at pH 6.5 and 7.0, while the small data set is even more concentrated in the even numbers than the large data set, which is more distributed around one value.

In summary, all three data sets show a comparable distribution of their measurement conditions and chemical shifts. Especially considering that all figures shown are density plots and therefore, especially if the large data set differs somewhat from the small and test data set due to its four to five times size, all important information for a prediction should be included in the large data set.

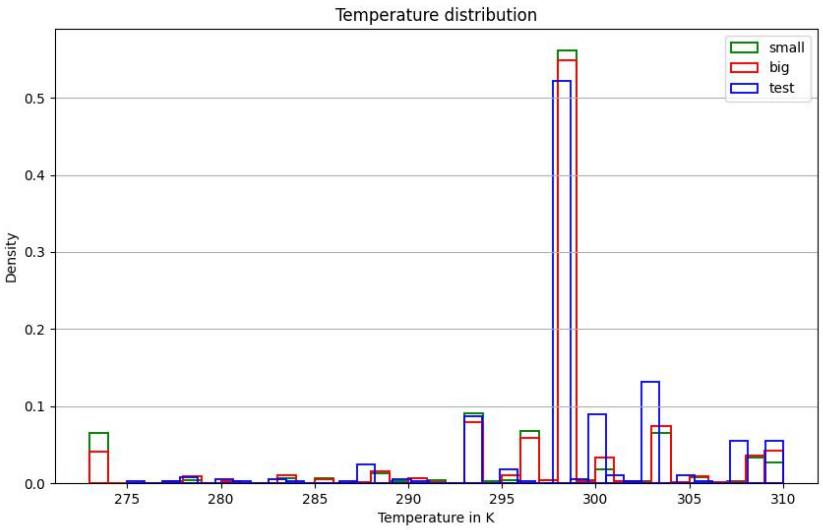


Figure 3: Temperature distribtution for all three data sets.

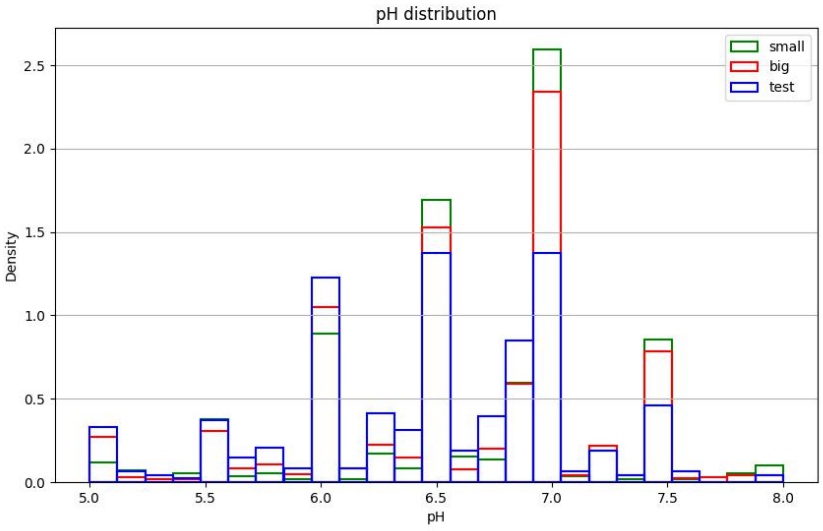


Figure 4: PH distribution for all three data sets.

3.1.2 Secondary structure content

Next, the distribution of the three secondary structures across all three data sets is shown (Figure 5).

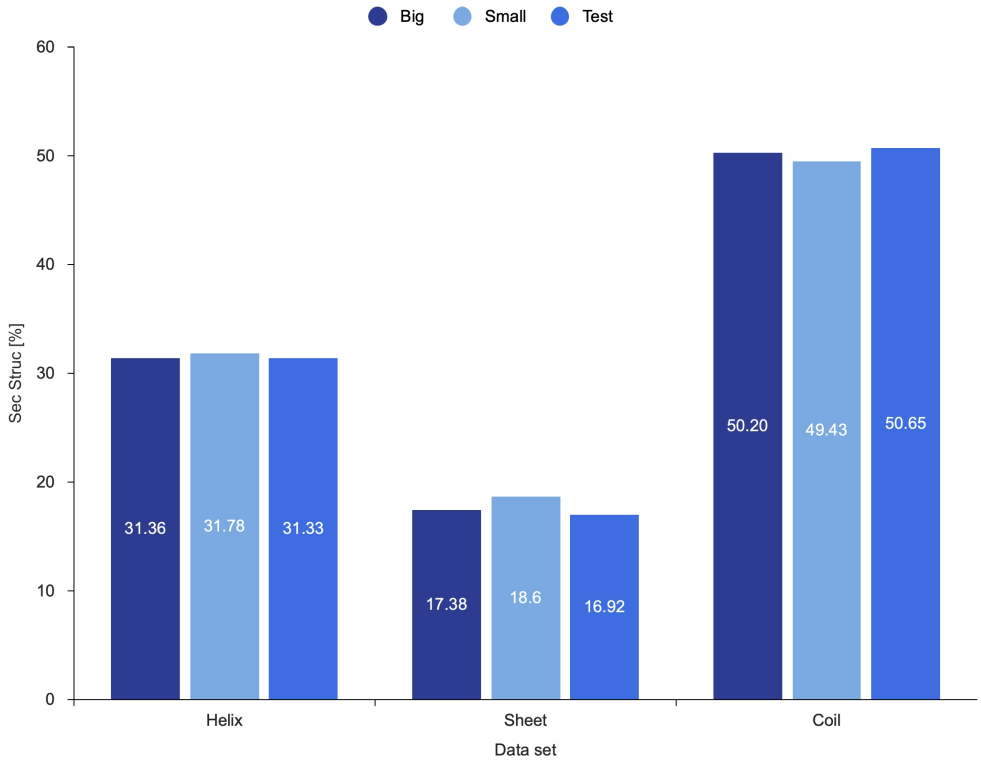


Figure 5: Comparison between the amount of secondary structure content for all three data sets.

The content of secondary structures is almost the same in all three data sets, with the small data set having slightly more sheet and helix structures and thus a lower proportion of coils. Overall, however, the distribution is the same, with the coil accounting for the largest proportion of secondary structures in the data set at about 50%, followed by the helix at about 32%. The least common secondary structure is the sheet, which accounts for nearly 18% of all secondary structures in the dataset.

When analyzing the PDB database, the highest content of secondary structures is the helix, followed by the sheet and coil, and then the left-handed alpha helix [22]. However, in our case, the distribution shows a majority of coils followed by helix and sheet. The reason for this is that the 3-state

calculation of coil contains many other structures such as the left-handed alpha helix. Looking at the 8-state secondary structure, the proportion of alpha helix is higher than that of coil, followed by beta sheet(Table 2).

3.2 Quadrants with high correlation

First, we compare and examine the correlation coefficients of the quadrants in the 10x10 grid subdivided N-HSQC spectra. Three different correlation plots were compared based on the large data set, the small data set, and the data from the reference article. As stated in the article, only correlation coefficients below -0.3 or above 0.3 are considered to follow the authors' intention to include only quadrants with at least a moderate correlation coefficient. In Fig. 6, the quadrants colored blue show a positive correlation above 0.3, while the quadrants colored red show a negative correlation below -0.3.

Comparing the three different data sets, we can see that they all show positive and negative correlations in similar regions of the N-HSQC spectra for the respective secondary structure. Based on the helix correlation coefficients, the positive correlation is represented by the same quadrants in all three data sets, while the negative one undergoes a slight adjustment as the amount of data increases. The quadrants with the highest negative correlation form a corner structure that is mirrored almost opposite to the corner structure of the positively correlated quadrants. Overall, there are more quadrants with negative correlation than with positive correlation.

In contrast to the helix correlation, more positively correlated quadrants than negatively correlated quadrants are generally found in the sheet heatmaps. It is noticeable that between the sheet and helix correlation heatmaps, the position of the positively and negatively correlated quadrants has changed. Previously, the left region of the N-HSQC spectrum showed a negative correlation for the helix, while it showed a positive correlation for the sheet. Also, the mentioned corner structure with high negative correlation for the helix content now shows the highest positive correlation for the sheet content. Based on this finding, it should be possible to distinguish between the sheet and helix contents in the spectra, as they show an inverse behavior in the N-HSQC spectra.

From the reference data set with 72 proteins to the large data set with 2353 proteins, the quadrants with positive high correlation are shifted more to the middle of the spectra and the total number of quadrants with positive correlation decreases. The negatively correlated quadrants for the sheet content, overlap the positively correlated quadrants of the helix content. Starting from the reference dataset, two outlying quadrants are removed

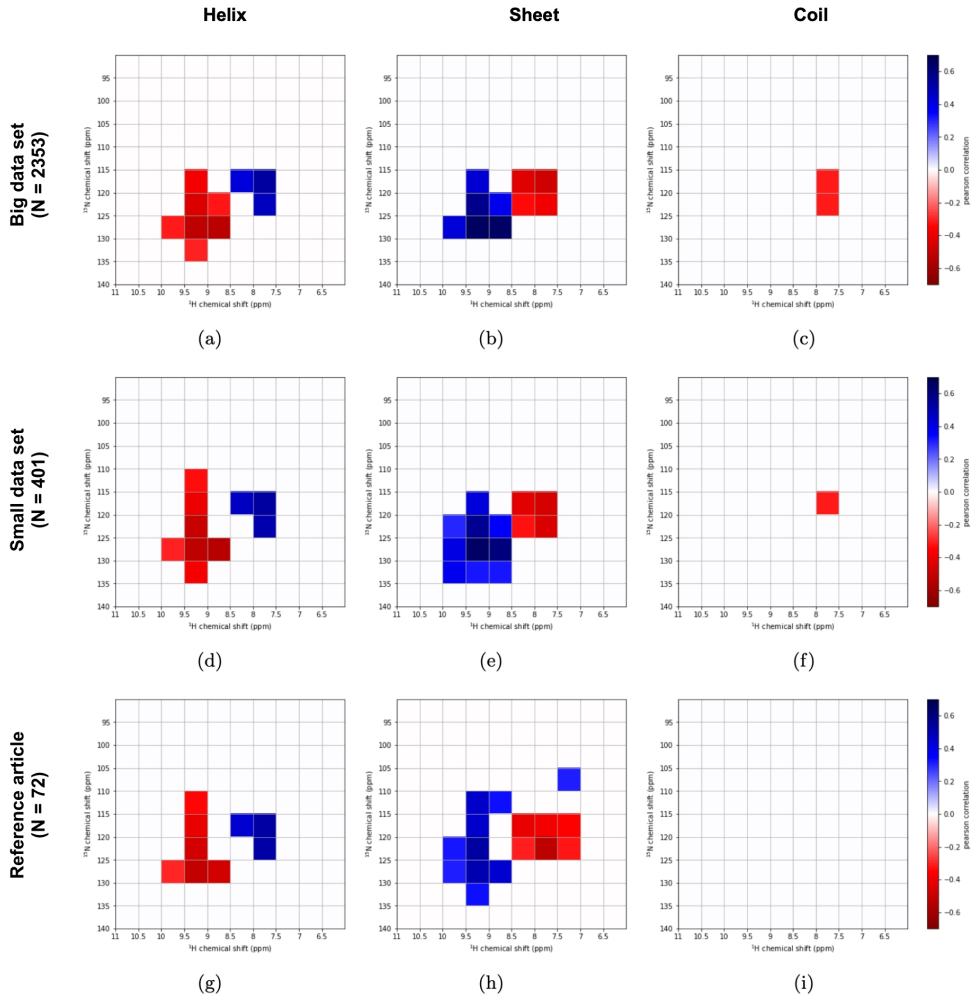


Figure 6: Comparison of moderate to high correlation quadrants for all three secondary structures across the big and small data set as well as the reference article.

compared to the small dataset, while there is no difference between the small and large datasets for the selected negative quadrants.

In the heatmaps of the coil correlation, one can notice that no correlation was found in the whole N-HSQC spectra in the reference article. Therefore, the authors decided to calculate the coil content as the content that was neither helix nor sheet. In our small and large data sets, we find one and two negatively correlated quadrants, respectively, but no positively correlated quadrants at all. This could indicate difficulties in predicting the coil content. In addition, the negatively correlated quadrants found for coil content overlap with the negatively correlated quadrants for sheet content, which also makes it difficult to distinguish between sheet and coil content.

The combination of the colored quadrants for the helix, sheet, and coil content of the large data set was used to create the *2353subgroup* model, and the same combination was used for the small data set to create the *401subgroup* model.

3.3 Statistical Analysis

To evaluate the models and draw conclusions about the impact of the amount of data used, one must compare the predicted secondary structure content of the test set with the actual secondary structure content calculated by the DSSP program. Two common methods for evaluating such predictions are the Pearson correlation coefficient and the root mean square error (RMSE).

As can be seen in Fig. 7, one can observe significant differences between the Pearson correlation coefficients of all three secondary structure contents as well as the different model types. In general, all tested models show a similar pattern in the prediction results: While the sheet content is predicted with the highest accuracy, closely followed by the helix content, the prediction of the coil content shows by far the worst accuracy.

Starting with the *401all* model, the Pearson correlation coefficient only reaches values of 0.41 for the helix content, 0.43 for the sheet content and 0.12 for the coil content. Thus, the *401all* model is the worst performing model in the test set and indicates that using only 72 proteins for the training process is not sufficient to efficiently find the relationship between all 100 quadrants.

Using the *401subgroup* model, there is a significant improvement in the prediction of helix and sheet content with corresponding Pearson correlation coefficients of 0.72 and 0.82. Interestingly, the prediction of coil content is 0.07 worse compared to the *401all* model. Therefore, it can be assumed that by using quadrants that have a high correlation coefficient for secondary structure content, the complexity of the prediction can be reduced. Secondary structure

contents with well-defined quadrants with positive or negative correlation such as sheet or helix benefit from this method, while secondary structure contents such as coil with only a few and vague correlations lose prediction accuracy.

Compared to the *401subgroup* model, the *2353subgroup* model shows a Pearson correlation coefficient of 0.76 and 0.82 for the helix and sheet content, respectively, showing a slight improvement over the smaller model. At the same time, the prediction accuracy for the coil content drops even further to 0.03. Since the difference between the *401subgroup-* and the *2353subgroup* model is only in the amount of data used to calculate the correlation coefficients for the quadrants (401 vs. 2351) and the training process (72 vs. 2000), the difference in performance between the models is due to the difference in the size of the data set. Either the selection of better correlated quadrants or the more diverse training data, but presumably both have their impact on improving the predictive accuracy of the *2353subgroup* model. In contrast, the coil prediction does not seem to benefit from the additional training data and shows no correlation at all.

Finally, comparing the Pearson correlation coefficients of the *2353all* model, we can see the best overall performance with a Pearson correlation coefficient of 0.8 for the helix content, 0.87 for the sheet content, and 0.5 for the coil content. This is an unexpected result considering the performance differences between the *401subgroup* and *2353subgroup* models and the *401all* model. Based on the previously discussed models, the use of high correlation quadrants appeared to improve the helix and sheet content immensely by nearly doubling the prediction accuracy while losing the predictive ability of the coil content. However, when the training data is more than five times larger than the *401all* model, the prediction accuracy in predicting the helix and sheet content has doubled or even more than doubled. In addition, the prediction accuracy of the coils has further increased compared to the *401all* model, which already showed better performance than the other two models.

Overall, the expansion of the data set has a large impact on model performance, especially when all quadrants are used for prediction. Another finding is the consistent order of predictability of the secondary structure elements within the models: prediction of the sheet content showed the highest accuracy across all models, followed by prediction of the helix content, and lastly prediction of the coil content.

In addition to the Pearson correlation coefficient, the RMSE value is also used to compare the secondary structure prediction for the four models, as it allows us to calculate the actual distance between the predicted and measured data, rather than their trends.

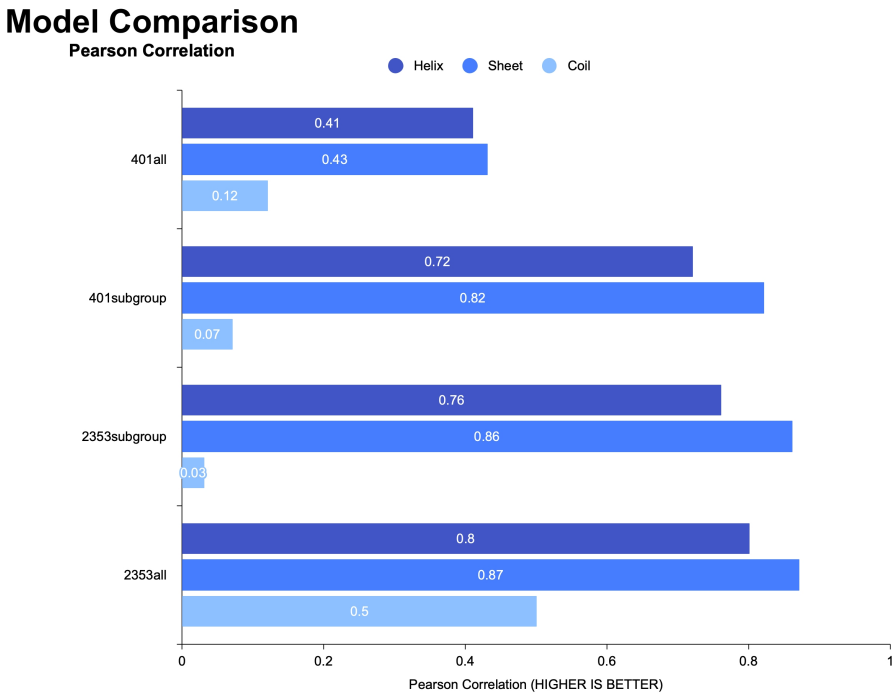


Figure 7: Comparison of the four different models based on the pearson correlation coefficient of the prediction of the three secondary structure elements helix, sheet and coil. The *2353all*-model (bottom) shows overall the best performance.

The result shown in Fig. 8 confirms the findings from the Pearson correlation coefficient discussed earlier. Starting with the *401all* model, the RMSE values are highest throughout the comparison with values of 0.44, 0.28, and 0.45 for the helix, sheet, and coil contents, respectively.

As expected from the above discussion, the *401subgroup* model shows significant improvement in predicting the helix and sheet content, but this time the prediction of the coil content is also improved from 0.45 to 0.2. The improvement of the *401subgroup* model seems to be due to the fact that the majority of the data points lie on the horizontal used for secondary structure content prediction (no correlation) (Fig. 9 f).

Staying with the *2353subgroup* model, there is a slight improvement from 0.18 to 0.15 for helix content and from 0.09 to 0.08 for sheet content compared to the *401subgroup* model. As mentioned earlier, the accuracy of the coil content prediction is based on a high density of proteins with a coil content of about 0.5, so the horizontal line around 0.48 of the *2353subgroup* model falsely suggests good accuracy without finding any correlation between prediction and measurement. The marginal improvement over the *401subgroup* model is given by the even more horizontal line crossing the center of the peak cloud more effectively (Fig. 9 c).

As with the Pearson correlation coefficient comparison, the *2353all* model shows the best overall RMSE values. Unlike the previously discussed model, the RMSE value for coil content of 0.14 in the scatter plot also shows correlation between predictions and measurements (Fig. 9 i). Elsewhere, RMSE values of 0.14 for helix content and 0.08 for sheet content show only minimal improvement in the competing models discussed previously.

To summarize an important finding from the data: When comparing the replicated model (*401subgroup* model) with the model from the reference article, we were able to show that with similarly found correlating quadrants we obtain similar correlation values (helix prediction), while minor changes in the selected quadrants can lead to a large improvement or a deterioration (sheet prediction). This emphasizes the importance of proper quadrant selection for prediction.

For the sake of clarity, the aforementioned scatter plots in Figure 9 show the linear correlation between the secondary structure content predicted by the models and the measured secondary structure content. Unlike the previous methods, where the Pearson correlation coefficient and RMSE value were calculated based on 488 000 predictions (test set size · 10x100 cross-validation), the scatter plots used the mean predictions of 1000 models from the cross-validation process, resulting in 488 points in the scatter plot.

Overall, the R^2 values are very low with a maximum value of 0.64 for the

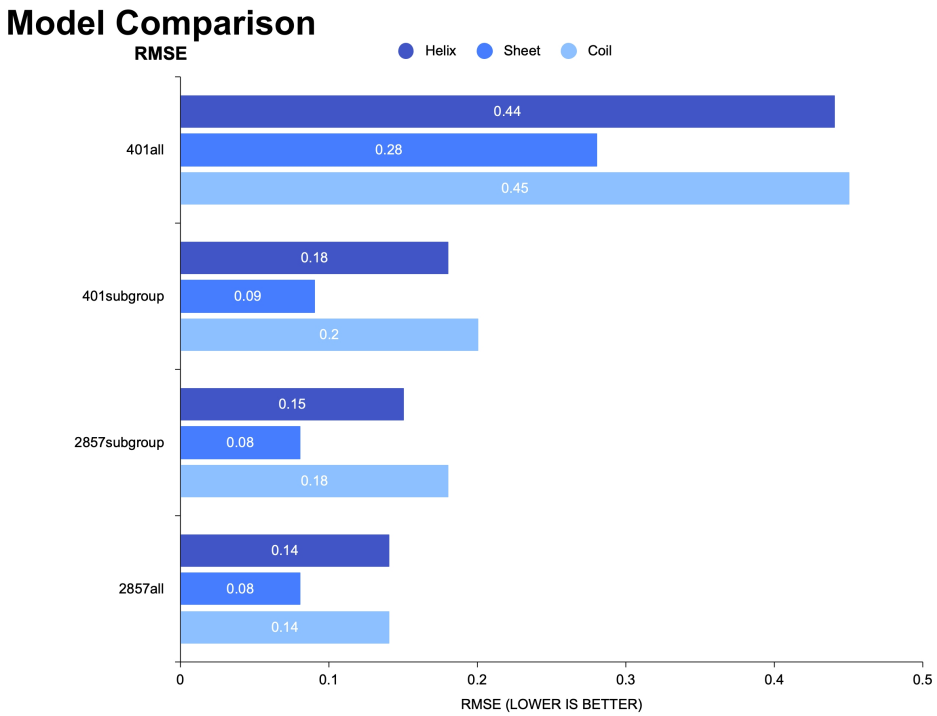


Figure 8: Comparison of the four different linear models secondary structure elements predictions by using the RMSE score. The *2353all*-model (bottom) shows the best performance.

helix content, 0.76 for the sheet content, and only 0.25 for the coil content, all from the *2353all* model. Even though the RMSE values (Figure 8) are on average wrong by 8% to 9% for sheet content and 14% to 18% for helix content, the three best models are not able to effectively explain the variation in secondary structure content. In particular, the coil content prediction suffers from being unable to explain any variation with R^2 values ranging from 0.25 for the *2353all* model to as low as -0.21 for the *2353subgroup* model.

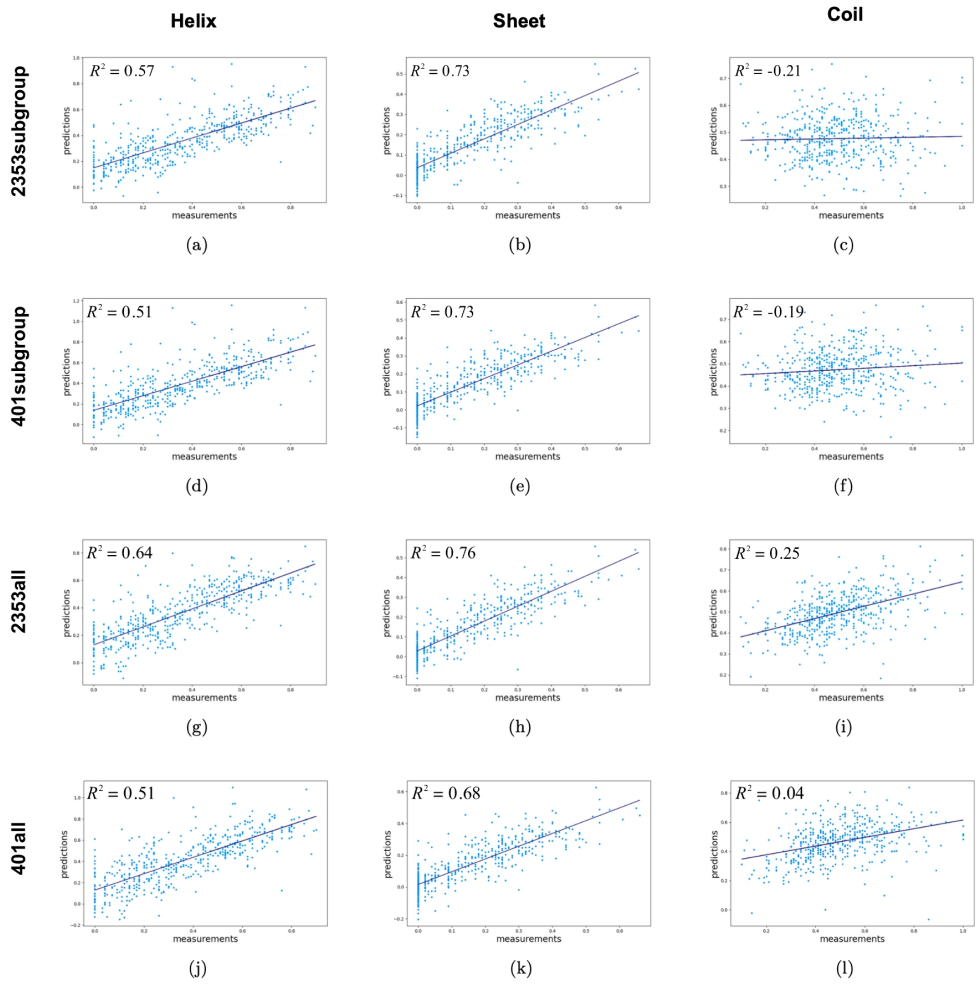


Figure 9: Correlation of the predicted and measured secondary structure content for helix, sheet and coil for all four trained models.

3.4 Interplay of the Helix, Sheet and Coil Model

Finally, we examined the distribution of prediction losses and how the three different models (helix, sheet and coil model) for the secondary structure elements complement each other by combining all three predictions. As shown in Figure. 10 and 11, the right panel shows the distribution of prediction losses for all three secondary structure elements. Calculating the sum of the three secondary structure predictions for all 488 000 predictions, we obtain the distribution centered around one in the figure on the left.

Starting with Figure 10, which shows the models trained on the larger dataset, one can observe the narrowest distribution for the sheet content and a wider distribution for helix and coil content in the *2353all* model, as expected. Surprisingly, the distribution of the combined prediction, when summed, shows a near perfect distribution centered at one. In comparison, the *2353subgroup* model shows a wider distribution for three secondary structure elements, which also leads to a wider distribution when combined. Both models show a slightly skewed right-hand distribution for the helix component in the individual prediction error plots, in contrast to the left-skewed distribution for the sheet and coil components. Since the error calculation was performed using the formula $measurement - prediction$, the right-skewed helix content distributions show a slight overestimation of helix content by the models, while the opposite is true for the sheet and coil content.

Continuing with the models trained on the small data set in Figure 11, the *401sugroup* model shows the same relative distribution as the models discussed previously: The distribution of the sheet content is the most narrowed distribution, followed by the distribution of the helix content and the coil content, which have almost the same distribution and differ only in the opposite skewness. As seen previously, the sheet prediction error histogram also shows a slightly left skewed distribution.

As expected from the previous analysis, the *401all* model shows the widest distribution for all secondary structure elements, which also leads to the widest distribution when the individual secondary structure elements are combined. The pattern found in the skewed distributions for all three different secondary structure elements is more pronounced due to the wider distribution. Therefore, the overestimation of the helix component and the underestimation of the sheet and coil components are the highest. Overall, the models trained with the small data set show a flatter distribution than the models trained with the larger data set.

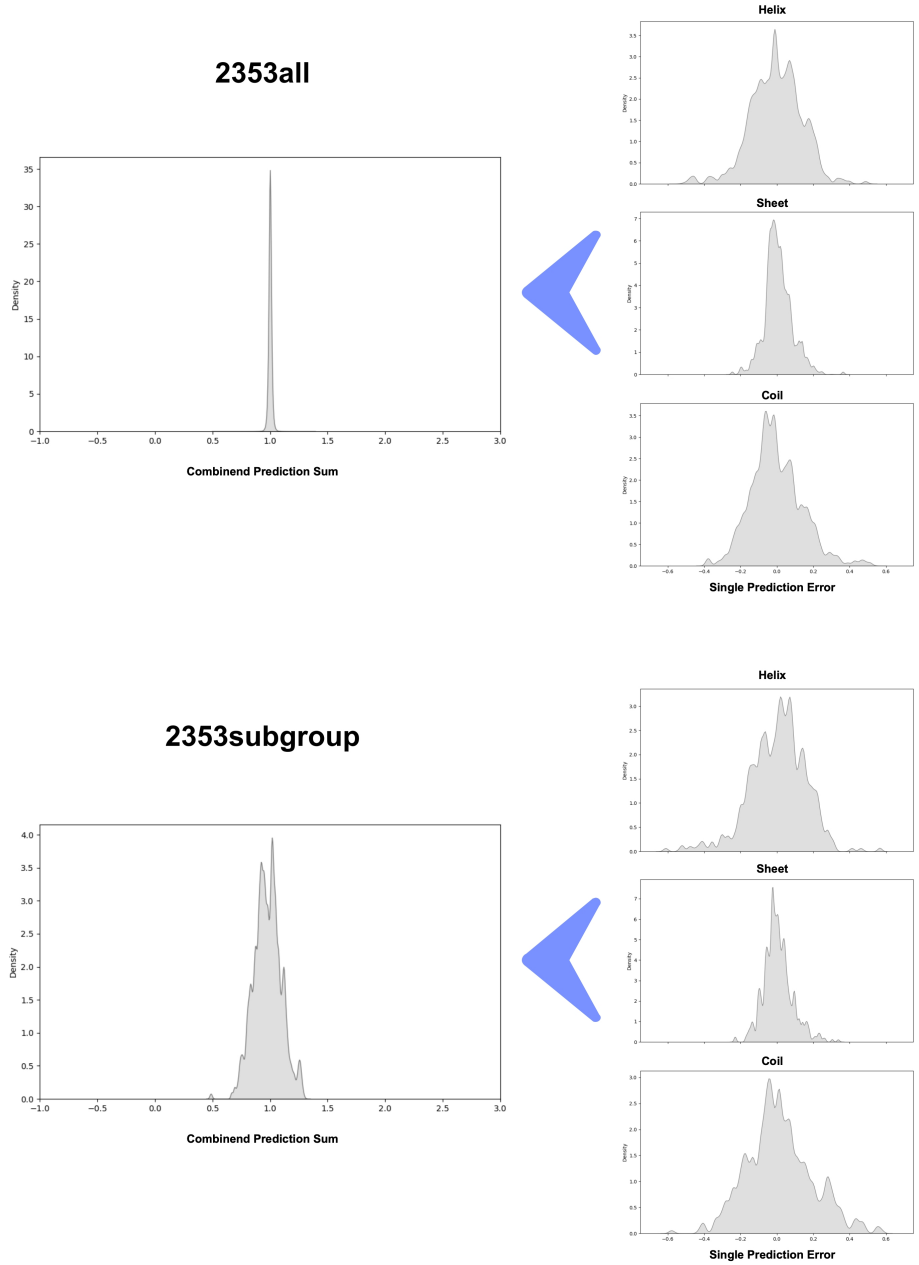


Figure 10: Correlation of the predicted and measured secondary structure content for helix, sheet and coil for all four trained models.

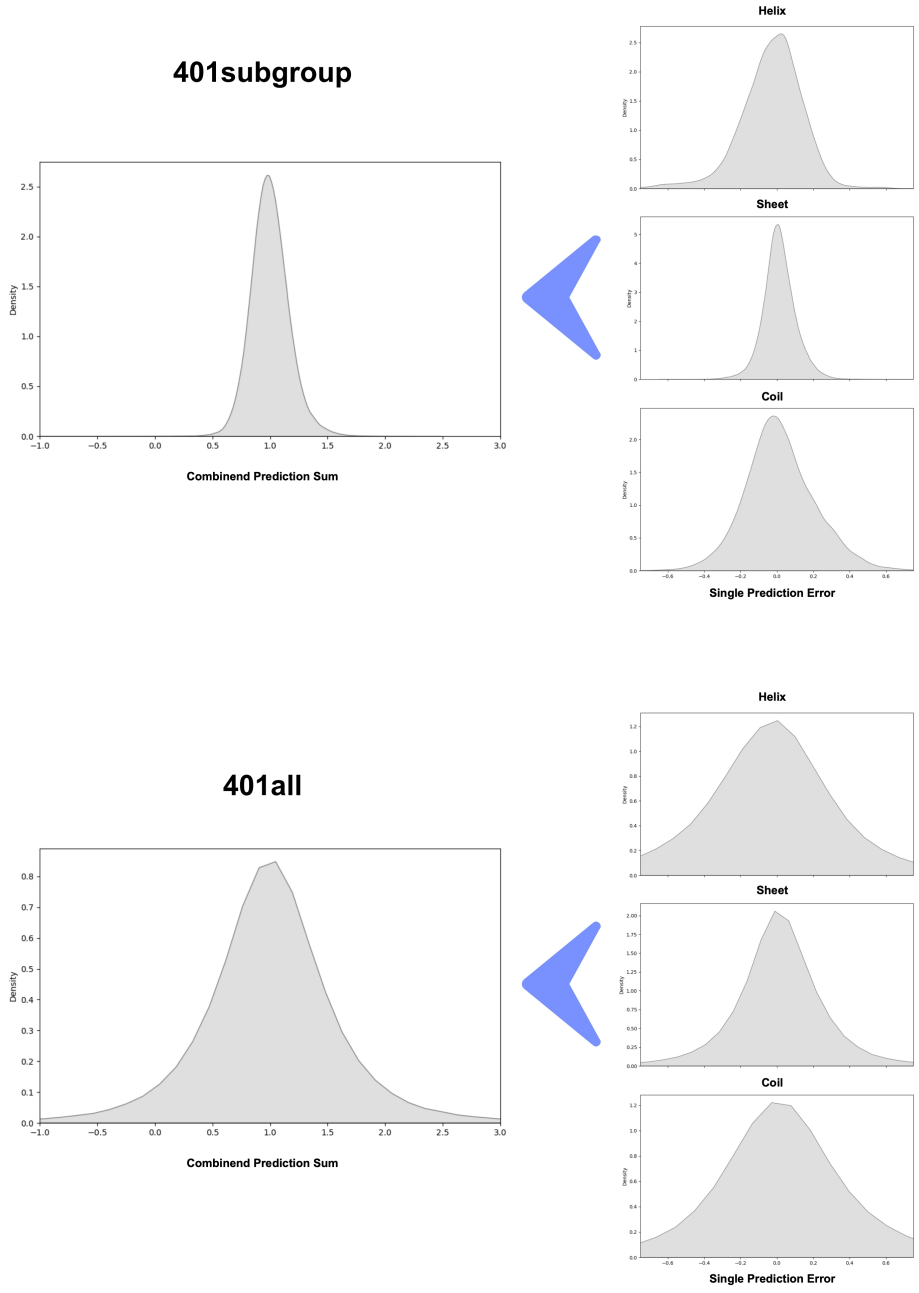


Figure 11: Correlation of the predicted and measured secondary structure content for helix, sheet and coil for all four trained models.

4 Discussion

Using a acquired dataset of 2841 proteins, we reproduce and improve on the results of an article published in Genetics and Molecular Biology in 2006 by Moreau et. al. that predicts the secondary structure content of proteins using their N-HSQC spectra. By using a subset of the dataset containing only entries prior to the publication of the reference article, I was able to find similar correlating quadrants to those used in the article.

We also explore the potential of N-HSQC spectra for predicting secondary structure content by comparing four models that differ in the size of their training data set and the given spectra information. All in all, we improved the paper result from 70% and 71% accuracy for helix and sheet content by 10 percentage points and even 16 percentage points to 80% and 87%, respectively. In addition, we were also able to predict the coil content to a certain extent.

When comparing the models with a small training data set, the use of quadrants with high correlation for a secondary structure element leads to a significant improvement in prediction. Based on this result, we can understand why the authors of "Prediction of the amount of secondary structure of proteins using unassigned NMR spectra: a tool for target selection in structural proteomics" used this approach.

As expected, the prediction accuracy increases with the size of the data set when comparing models trained with a small data set and models trained with a large data set. This time, however, the model that includes all 100 quadrants shows better performance than the model that uses only quadrants with high correlation for a secondary structure element.

Interestingly, for all models, the prediction accuracy decreased from sheet to helix to coil, even though the amount of secondary structure content represented in all datasets is exactly the opposite. This phenomenon can be partially explained by the relationship between the eight-state secondary structure and the three-state secondary structure. The program DSSP used calculates the eight-state secondary structure elements and forms the three-state secondary structure elements by a certain combination of these structure elements. While only two different eight-state secondary structure elements were combined for the three-state secondary structure of helix and sheet, the three-state secondary structure of coil consisted of four different eight-state secondary structure elements. Assuming that each eight-member secondary structure element can be identified by a specific pattern in the N-HSQC spectra, the more different eight-state secondary structure elements are reduced to a single three-state secondary structure element, the more difficult

it is to find the corresponding patterns, since different combinatorial overlaps may occur. In support of this assumption, the comparison of correlating quadrants showed little to no correlation found for coil content, indicating more complex patterns possibly due to the combination of four different eight-state secondary structures. One way to test this explanation is to predict the eight-state secondary structure elements. If the eight-state secondary structure elements contained in the three-state secondary structure elements of the coil have better prediction accuracy, this assumption seems to hold. Of course, not every secondary structure element may have an equally unique pattern, which will also play an important role in prediction accuracy.

While the difference between coil and helix/sheet is rather more obvious based on the number of quadrants found with high correlations in the spectra, the sheet and helix content show the same number of quadrants, at least for the large data set. Here, especially the number of positive correlations seems to be a main factor for the prediction accuracy. Therefore, the finding of at least twice as many positive correlations for sheet content than for helix content may explain the higher prediction accuracy of sheet content.

Although the information from the quadrants with high correlation is able to explain the helix and sheet content to some extent, the best model uses all 100 quadrants, suggesting that there is more important underlying information in the spectra than the quadrants with high correlation for a secondary structure element.

5 Outlook

N-HSQC spectra provide a unique pattern for each molecule and encode information about secondary structure content. This information can be used to predict the secondary structure content of proteins. Despite the confirmation and improvement of Moreau et. al.'s results, the method needs further investigation to realize its full potential.

In the age of artificial intelligence, it is obvious to explore the potential of the method with machine learning and deep learning tools, using more data and including the prediction of eight-state secondary structures in addition to three-state secondary structures. On the road to exploring the use of N-HSQC spectra for secondary structure prediction, some challenges lie ahead, especially in the prediction of secondary structures related to proline, as its secondary amine does not provide a signal in the N-HSQC spectra.

By solving these problems and using more and more data from the ever-growing databases, this method could be applied in protein structure prediction tools used for initial inspection of proteins in the drug discovery stage and as an aid in general protein research. In contrast to currently used prediction models based on primary protein structure, the blind spot of protein dynamics can be investigated by using actual structural information measured by NMR. Therefore, N-HSQC-based prediction methods could be used as a constraint for tertiary protein structure prediction methods and provide us with new insights into the protein folding process.

6 References

[1] John Jumper et al. "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596.7873 (2021), pp. 583–589.

[2] Fodil Azzaz et al. "The Epigenetic Dimension of Protein Structure Is an Intrinsic Weakness of the AlphaFold Program". In: *Biomolecules* 12.10 (2022), p. 1527.

[3] Peter B Moore et al. "The protein-folding problem: Not yet solved". In: *Science* 375.6580 (2022), pp. 507–507.

[4] Kurt Wüthrich. "Protein structure determination in solution by NMR spectroscopy." In: *Journal of Biological Chemistry* 265.36 (1990), pp. 22059–22062.

[5] Paul Guerry and Torsten Herrmann. "Advances in automated NMR protein structure determination". In: *Quarterly reviews of biophysics* 44.3 (2011), pp. 257–309.

[6] Yves Aubin, Geneviève Gingras, and Simon Sauvé. "Assessment of the three-dimensional structure of recombinant protein therapeutics by NMR fingerprinting: demonstration on recombinant human granulocyte macrophage-colony stimulation factor". In: *Analytical chemistry* 80.7 (2008), pp. 2623–2627.

[7] Erik RP Zuiderweg. "Insights into Protein Dynamics from 15 N-1 H HSQC". In: *Magnetic Resonance Discussions* (2021), pp. 1–35.

[8] Vitor Hugo Moreau, Ana Paula Valente, and Fábio CL Almeida. "Prediction of the amount of secondary structure of proteins using unassigned NMR spectra: a tool for target selection in structural proteomics". In: *Genetics and Molecular Biology* 29 (2006), pp. 762–770.

[9] Helen M Berman et al. "The protein data bank". In: *Nucleic acids research* 28.1 (2000), pp. 235–242.

[10] Jeffrey C Hoch et al. "Biological Magnetic Resonance Data Bank". In: *Nucleic Acids Research* 51.D1 (2023), pp. D368–D376.

[11] Brian Kuhlman and Philip Bradley. "Advances in protein structure prediction and design". In: *Nature Reviews Molecular Cell Biology* 20.11 (2019), pp. 681–697.

[12] Yaoqi Zhou et al. "Trends in template/fragment-free protein structure prediction". In: *Theoretical chemistry accounts* 128 (2011), pp. 3–16.

[13] Mohammed AlQuraishi. "Machine learning in protein structure prediction". In: *Current opinion in chemical biology* 65 (2021), pp. 1–8.

[14] Shi Dong, Ping Wang, and Khushnood Abbas. "A survey on deep learning and its applications". In: *Computer Science Review* 40 (2021), p. 100379.

[15] Wouter G Touw et al. "A series of PDB-related databanks for everyday needs". In: *Nucleic acids research* 43.D1 (2015), pp. D364–D368.

[16] Wolfgang Kabsch and Christian Sander. "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features". In: *Biopolymers: Original Research on Biomolecules* 22.12 (1983), pp. 2577–2637.

[17] Nicola J Baxter and Michael P Williamson. "Temperature dependence of 1 H chemical shifts in proteins". In: *Journal of biomolecular NMR* 9 (1997), pp. 359–369.

[18] An-Suei Yang and Barry Honig. "On the pH dependence of protein stability". In: *Journal of molecular biology* 231.2 (1993), pp. 459–474.

[19] Kemper Talley and Emil Alexov. "On the pH-optimum of activity and stability of proteins". In: *Proteins: Structure, Function, and Bioinformatics* 78.12 (2010), pp. 2699–2706.

[20] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[21] Pauli Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[22] Sven Hovmöller, Tuping Zhou, and Tomas Ohlson. "Conformations of amino acids in proteins". In: *Acta Crystallographica Section D: Biological Crystallography* 58.5 (2002), pp. 768–776.

7 Appendix

Table 3: Correlation coefficients for quadrants with a higher absolute correlation than 0.3 in the reference article

quadrant	helix	sheet	coil
32	-	0.31	-
45	-	0.33	-
46	-0.34	0.45	-
52	-	-0.35	-
53	0.53	-0.38	-
54	0.44	-0.40	-
56	-0.41	0.45	-
62	-	-0.32	-
63	0.52	-0.53	-
64	-	-0.31	-
66	-0.47	0.53	-
67	-	0.32	-
75	-0.47	0.44	-
76	-0.51	0.50	-
77	-0.30	0.31	-
86	-	0.33	-

Table 4: Correlation coefficients for quadrants with a higher absolute correlation than 0.3 in the big data set

quadrant	helix	sheet	coil
53	0.53	-0.48	-0.31
54	0.42	-0.43	-
56	-0.37	0.43	-
63	0.47	-0.39	-0.31
64	-	-0.34	-
65	-0.32	0.38	-
66	-0.44	0.57	-
75	-0.53	0.65	-
76	-0.53	0.66	-
77	-0.32	0.43	-
86	-0.30	-	-

Table 5: Correlation coefficients for quadrants with a higher absolute correlation than 0.3 in the small data set

quadrant	helix	sheet	coil
46	-0.33	-	-
53	0.53	-0.46	-0.32
54	0.47	-0.42	-
56	-0.4	0.42	-
63	0.51	-0.45	-
64	-	-0.33	-
65	-	0.36	-
66	-0.50	0.56	-
67	-	0.3	-
75	-0.54	0.6	-
76	-0.53	0.65	-
77	-0.30	0.4	-
85	-	0.32	-
86	-0.37	0.32	-
87	-	0.38	-

8 Acknowledgement

There will be an acknowledgement in the Master Thesis.