Analysing the frustration of disordered-to-ordered protein interactions with machine learning

Jake Jackson[1]

[1]Università di Padova, Dipartimento di Fisica e Astronomia, Padova, 35122, Italia

ABSTRACT

This paper utilizes frustration index data to explore the role of various factors including well type, binding state, residue position, and amino acid pairs in the protein ensemble interaction landscape. The relationship is explored using standard data analysis and machine learning with XGBoost. Classification for DOR bound proteins achieved an accuracy of 0.762, regression was used for frustration index prediction with an RMSE accuracy of 0.446. This was visualized by creating a median frustration plot Figure 4b.

1 Introduction

Understanding protein interactions is vital for understanding biological processes and developing more effective and targeted medicines[1]. However, this is an exceedingly complex and difficult problem. Despite a protein having a well-defined chain of amino acids (AA) there is an enormous range of different ways in which they can fold giving many structures[2]. The folding of a protein is dependent on the interactions its subjected to whether that be to itself or external molecules. Given it is the structure that gives rise to the active sites and therefore function of the proteins. This highlights the importance of understanding the role of these different interactions in nature.

The predominant research focus is on native states, the form in which the protein is considered to be folded properly and at a minima of Gibbs free energy[3]. The native state usually has a well-defined secondary through to quaternary structure. Given the complexity of the interaction landscape found in nature, this paper aims to explore the role of various factors including well type, binding state, residue position, and amino acid pairs in the protein ensemble interaction landscape. This is done through both traditional analysis and machine learning techniques.

This report utilizes frustration index data generated by Fuxreiter et al (2022)[4]. The frustration index gives us a way of quantifying this inability of the particular residue to fulfill all of the competing interactions, at a given instant[5]. The data is split into free and bound. Unlike bound states, free protein states are not bound to other molecules in the cell. The concept of protein disorder is illustrated in Figure 1.

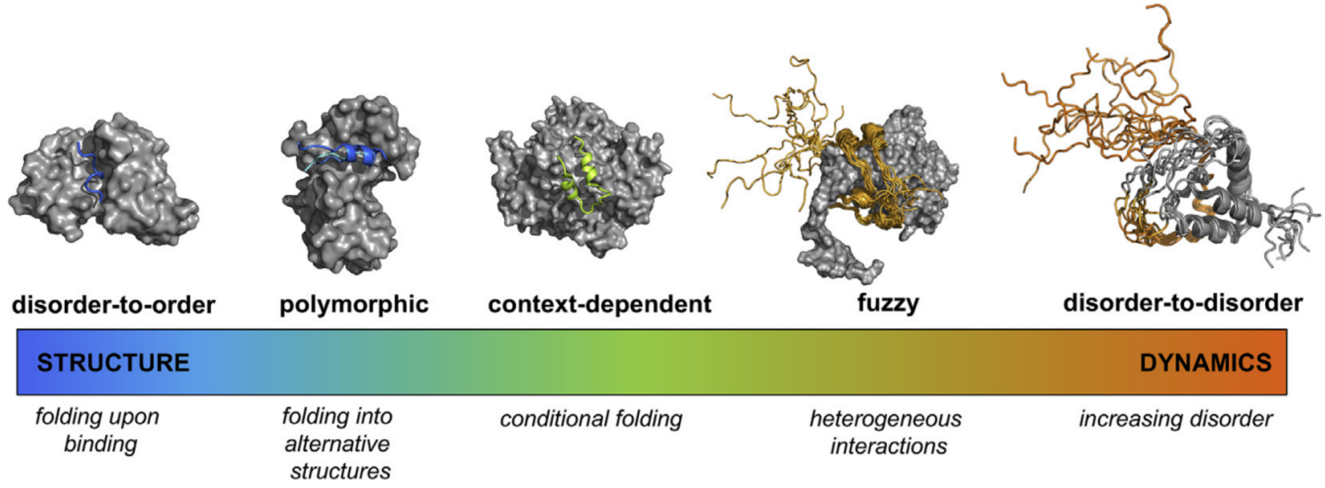


Figure 1. Protein interactions in the context of order: (source[6])

The left of 1 shows disorder-to-order DOR in which a disordered protein binds to an ordered protein. This report will focus on developing an analysis pipeline using DOR, in order to better understand protein interactions. With a near future aim to incorporate a comparative analysis of disorder-to-disorder DDR in the near future.

2 Methods

Before moving to more complex machine learning methods, the data was explored through some basic analysis to see if any fundamental relationships could be determined. The general approach across the project was to start simply and build in computational complexity as required. The following methods were applied to the DOR free and unbound data.

2.1 Data Exploration

The role of residue position on the protein frustration was explored by producing position-based contact maps. This is shown by Figure 7 in the supplementary material. This approach, however, was over complicated and more difficult to interpret. Preliminary machine learning results showed that the amino acids were the most important factors for determining the frustration and therefore the relationship between amino acids in the contacts became the primary focus of the investigation.

The DOR interactions cover 78 unique proteins in the bound case and 82 for free. This corresponds to 167686 and 387924 individual contacts for the free and bound cases, respectively. The data was in the most part explored by comparing how the different properties were linked to the frustration index. This was done by selecting the properties out the data and performing ether an aggregating algorithm like counting or finding the median.

The percentage frequency of each pair of unique amino acids was calculated to visualize the distribution of contacts. The plots were subsequently normalised by the total number of contacts indicated above. A similar method was used however aggregating the median values.

To answer the question, what is the relationship between the well type to frustration index? The data was split into the PDB[7] code and then the well type. Following this, the medians were calculated for each group and this was plotted to produce Figure 3. In a similar fashion this research also asks the question of how the amino acid polarity affects the interactions? In order to do this the amino acids were grouped according to there chemical properties and the data was processed in similar way to other attributes.

2.2 Machine Learning

In order to train the machine learning algorithms the data was split in two different ways. The first is a random shuffle and a percentage of the dataframe rows taken out for testing. The second is the complete removal of a particular protein in order to see how well a model could infer a "new" protein.

For the machine learning analysis the Python package XGBoost[8] was used. XGBoost is an ensemble learning technique combines the predictions of many weak learners in a technique call gradient boosting. This is a very well-optimized process and produces strong predictions. For classification and regression these weak learners are simply trees and regression trees. Like many machine learning algorithms, XGBoost requires all data to be numerical. This problem was overcome by encoding the letters into integers and decoding them back after training using the same encryption. Classification is the simplest machine learning approach. Therefore, the preliminary step was to predict the frustration state. This data only has three labels minimally, neutral, and highly frustrated. The prediction was calculated using XGClassifier. The accuracy of the algorithm was calculated as simply as the fraction of correctly predicted values.

Building upon this the XGRegressor was used to predict a continuous value for the frustration index for each set of contacts. The effectiveness of the algorithm was tested by computing the root mean square error (RMSE) as shown in equation 1 below.

$$RMSE = \sqrt{\frac{\Sigma_{i=1}^{N}(x_i - \hat{x}_i)^2}{N}} \tag{1}$$

Following this the plots were made of the median frustration indices were generated in the same way as previously stated.

In addition, code was wrote to fetch files the PDB data of a the protein databank with PyPDB[7] and then parse the chain sequences using BioPython[9]. This was done with the aim of supplementing the dataset with the local amino acids of each site. Then unsupervised machine learning techniques like K-means clustering[10] could be done to identify sequence motifs. These motifs could then be used for prediction. However this is still in progress, but remains an exciting area for future development.

3 Results

3.1 Data Exploration

Since the data is made of contacts, therefore the % frequencies in the Figures 6a and 6b in the supplementary materials correspond to the most common interactions. Interestingly this appears in the free and bound cases to be L-L (Leucine to

Leucine). This is a non polar to non polar interaction. There is a remarkable similarity between the bound and free states suggesting that the grouping of bound and free proteins doesn't affect the relative frequencies of amino acid interactions. There is also a slight asymmetry when viewing the top left to bottom right symmetry axis.

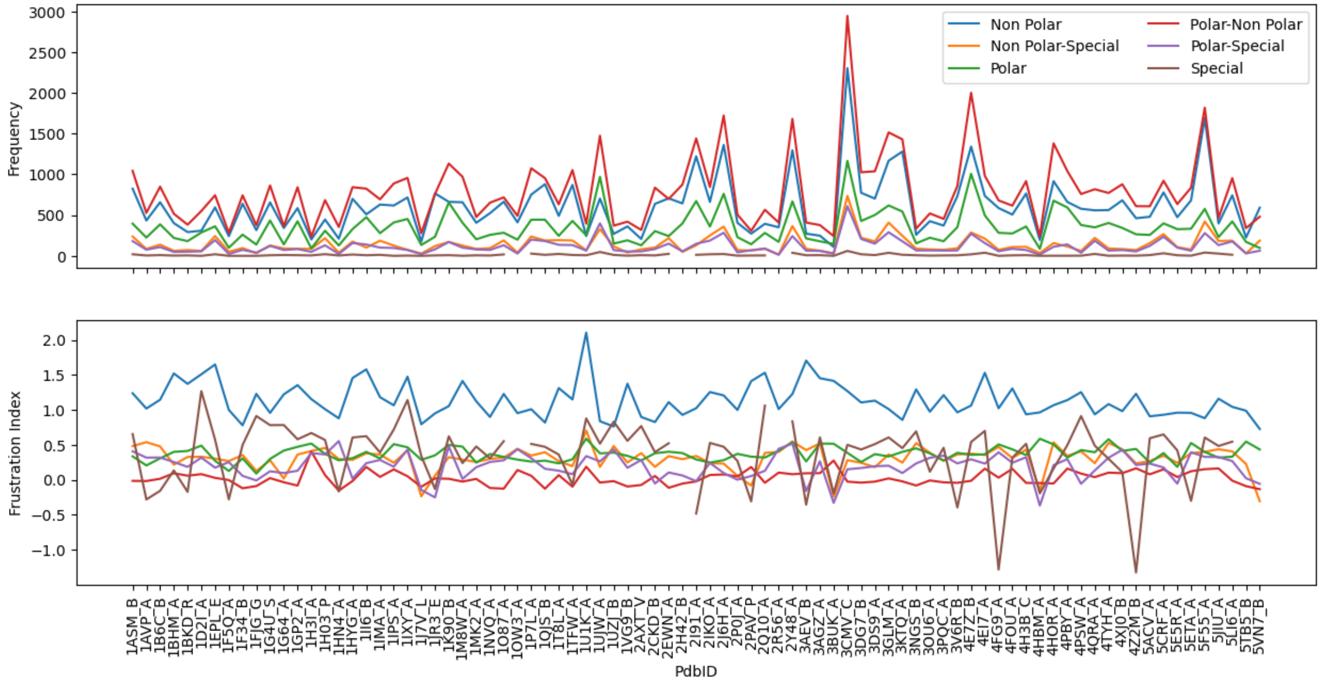


Figure 2. Free polarity groupings

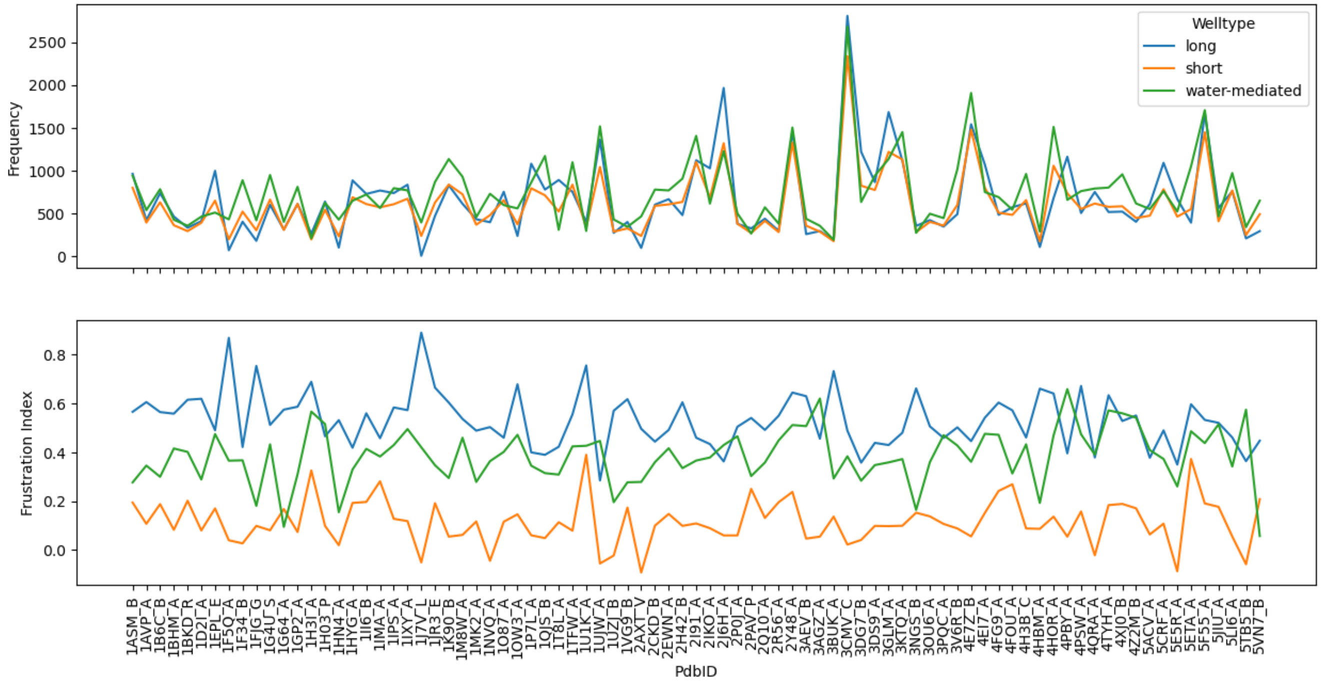


Figure 3. Free well groupings

When examining the interactions there are many interesting results. The minimal frustration states have a higher index. Here we can see that the most common pairing is the polar to non polar connection. This corresponds to a neural frustration.

The most minimally frustrated is the non polar pairs. The special amino acids proline and cysteine, have structural priorities that may explain the frustration index dips in the lower part of Figure 5a. For instance proline has a nitrogen atom covalently locked within a ring, therefore having restricted phi angle[11].

Figure 3 shows that the short wells have the lowest frustration index, followed by the water mediated and then long. Interestingly all well types tend to appear in the data in similar frequencies. This is something that is also seen in the free case (see supplementary material 5b).

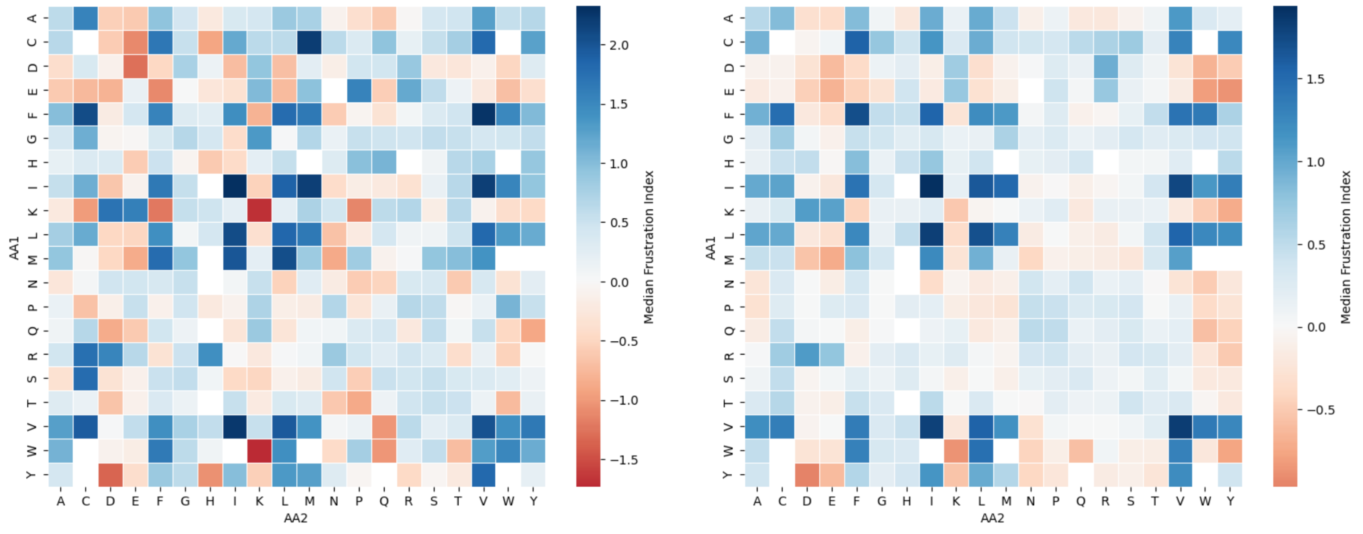
3.2 Machine Learning

Algorithm	Order	Binding	Training Input	Label	Score
XGClassifier	DOR	Free	AA1, AA2	FrstState	0.752
XGClassifier	DOR	Free	Res1, Res2, AA1, AA2	FrstState	0.752
XGClassifier	DOR	Free	Res1, Res2, AA1, AA2 + Interaction	FrstState	0.759
XGClassifier	DOR	Bound	AA1, AA2	FrstState	0.758
XGClassifier	DOR	Bound	Res1, Res2, AA1, AA2	FrstState	0.762
XGRegressor	DOR	Free	AA1, AA2	FrstIndex	0.415
XGRegressor	DOR	Free	Res1, Res2, AA1, AA2	FrstIndex	0.446
XGRegressor	DOR	Bound	AA1, AA2	FrstIndex	0.405
XGRegressor	DOR	Bound	Res1, Res2, AA1, AA2	FrstIndex	0.441

Table 1. Machine learning algorithm scores, where Res denotes the position of amino acid AA along chains 1 and 2. In addition, Frst denotes frustration

Table 1 shows that the XGClassifier had the highest accuracy with 0.762 for DOR bound. The amino acid types in the interaction had the largest impact on the performance. The position only input was tested but gave too low an accuracy to pursue meaningfully.

The fact the positions are far less significant is a interesting result. A potential reason for this maybe that the model cannot retain the complexity required to benefit from positional relationships, which is known to be very complex. Adding the interaction information did actually help gain accuracy producing the best case for free classification.



(a) Test Data for PDB ID: 1ASM_B

(b) Predicted Data for PDB ID: 1ASM_B

Figure 4. XGRegressor frustration prediction for single protein

The XGRegressor had a lower accuracy, however with more time this is likely improvable through parameter tuning such as Grid-Search or using a more specialized algorithm. The bulk properties of the prediction however were very promising. Figures 4b and 4a show a very close match between for the median frustration indexes when predicting an unseen protein. There is a smoothing of the extreme positions for instance K-K and K-W however the positions of extreme values tends to be generally correct. Inspecting the plots look very similar albeit the predicted plot looks slight smoothed.

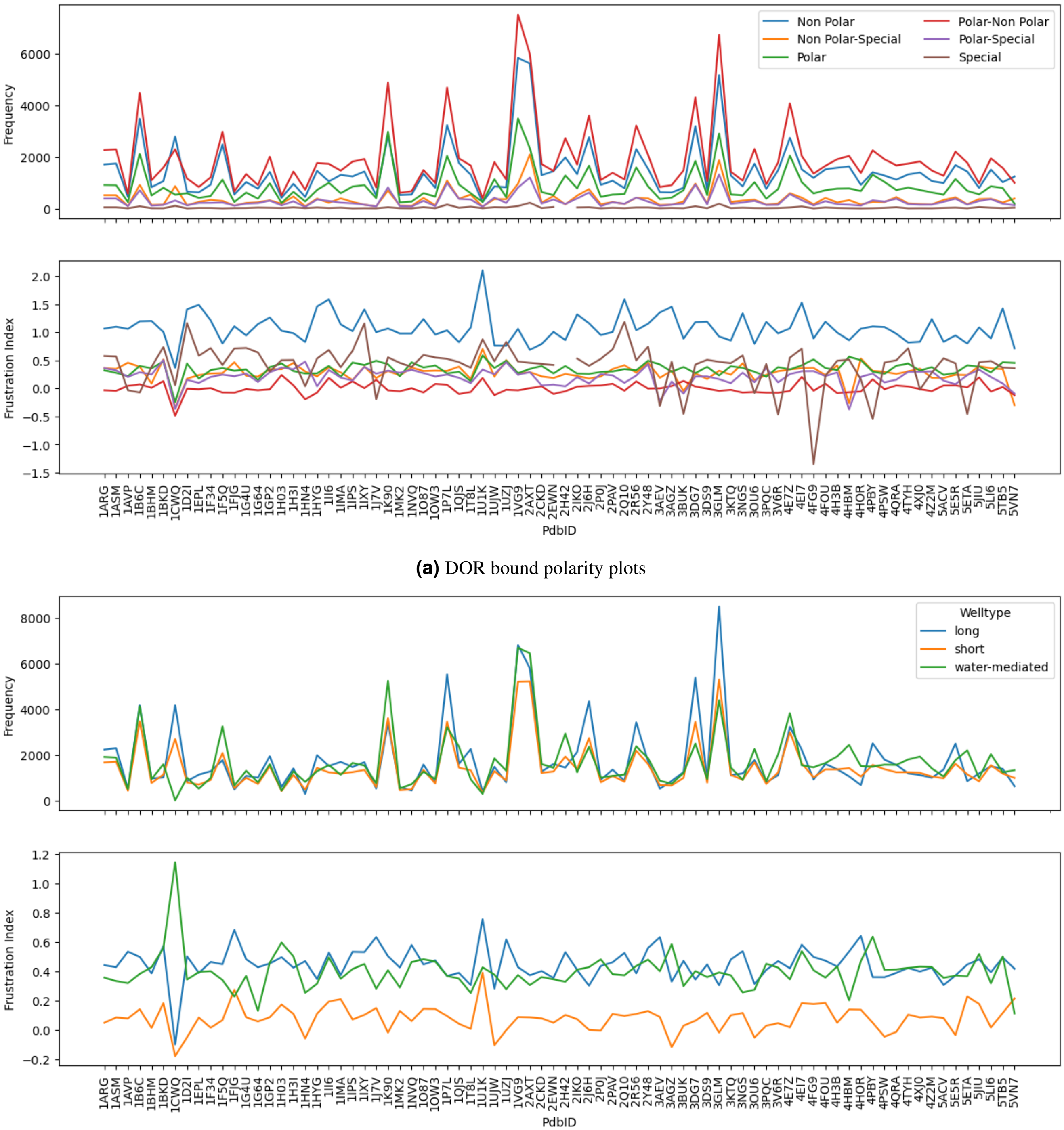
4 Conclusion

Overall this project was successful in its primary objective to explore the protein interaction landscape for DOR contacts. This was visualized to show how the amino acid characteristics interplay with the frustration indices. The application of machine learning achieved an accuracy of 0.762 and 0.446 for classification and regression respectively. This is a promising preliminary investigation. And the technique was still able to produce a reasonable result when taking the median of the predicted frustration indices for the test protein 1ASM_B as shown by Figure 4b.

References

1. Rao, V. S., Srinivas, K., Sujini, G. N. & Kumar, G. N. S. Protein-protein interaction detection: Methods and analysis. *Int. J. Proteomics* **2014**, 147648, DOI: 10.1155/2014/147648 (2014).
2. Dill, K. A., Ozkan, S. B., Shell, M. S. & Weikl, T. R. The protein folding problem. *Annu. Rev. Biophys.* **37**, 289–316, DOI: 10.1146/annurev.biophys.37.092707.153558 (2008). PMID: 18573083, https://doi.org/10.1146/annurev.biophys.37.092707.153558.
3. Shehu, A., Kavraki, L. E. & Clementi, C. On the characterization of protein native state ensembles. *Biophys. J.* **92**, 1503–1511, DOI: https://doi.org/10.1529/biophysj.106.094409 (2007).
4. Monzon, A. M., Piovesan, D. & Fuxreiter, M. Molecular determinants of selectivity in disordered complexes may shed light on specificity in protein condensates. *Biomolecules* **12**, DOI: 10.3390/biom12010092 (2022).
5. Ferreiro, D. U., Komives, E. A. & Wolynes, P. G. Frustration in biomolecules. *Q. Rev. Biophys.* **47**, 285–363, DOI: 10.1017/S0033583514000092 (2014).
6. Miskei, M., Horvath, A., Vendruscolo, M. & Fuxreiter, M. Sequence-based prediction of fuzzy protein interactions. *J. Mol. Biol.* **432**, 2289–2303, DOI: https://doi.org/10.1016/j.jmb.2020.02.017 (2020).
7. Gilpin, W. PyPDB: a Python API for the Protein Data Bank. *Bioinformatics* **32**, 159–160, DOI: 10.1093/bioinformatics/btv543 (2015). https://academic.oup.com/bioinformatics/article-pdf/32/1/159/49016436/bioinformatics_32_1_159.pdf.
8. Chen, T. & Guestrin, C. XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, DOI: 10.1145/2939672.2939785 (ACM, 2016).
9. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423, DOI: 10.1093/bioinformatics/btp163 (2009). https://academic.oup.com/bioinformatics/article-pdf/25/11/1422/48989335/bioinformatics_25_11_1422.pdf.
10. Abu Jamous, B. *Integrative cluster analysis in bioinformatics* (John Wiley Sons Inc., Chichester, West Sussex, United Kingdom, 2015).
11. Morgan, A. A. & Rubenstein, E. Proline: The distribution, frequency, positioning, and common functional roles of proline and polyproline sequences in the human proteome. *PLOS ONE* **8**, 1–9, DOI: 10.1371/journal.pone.0053785 (2013).

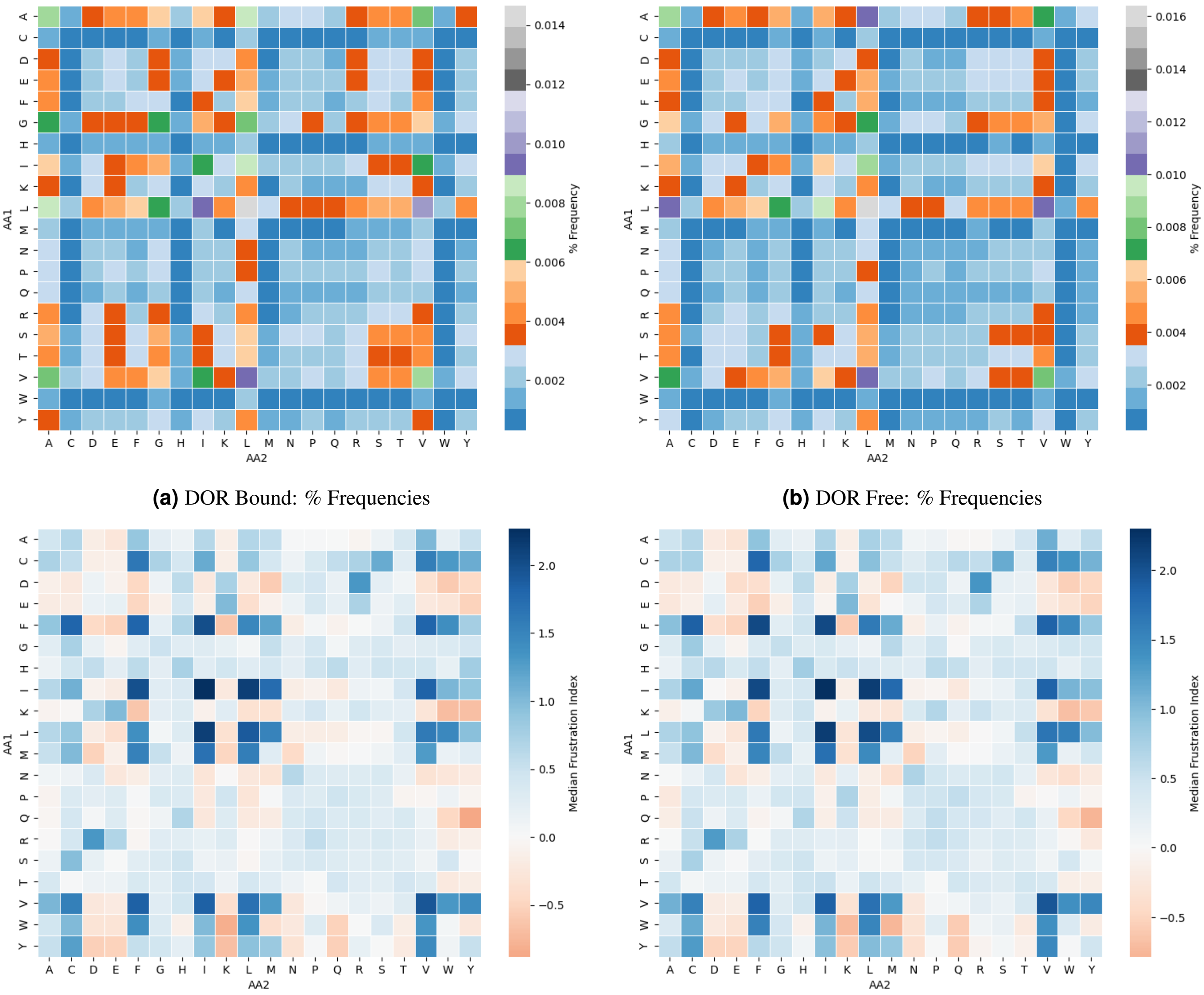
5 Supplementary Material



(a) DOR bound polarity plots

(b) DOR bound well plots

Figure 5. Bound interaction breakdown for well type and interacting amino acid polarities with freqeuncies



(a) DOR Bound: % Frequencies

(b) DOR Free: % Frequencies

(c) DOR Bound: Median Frustration Index

(d) DOR Free: Median Frustration Index

Figure 6. Plots for the frequency and frustrations for residues on interacting residue chains AA1 and AA2.

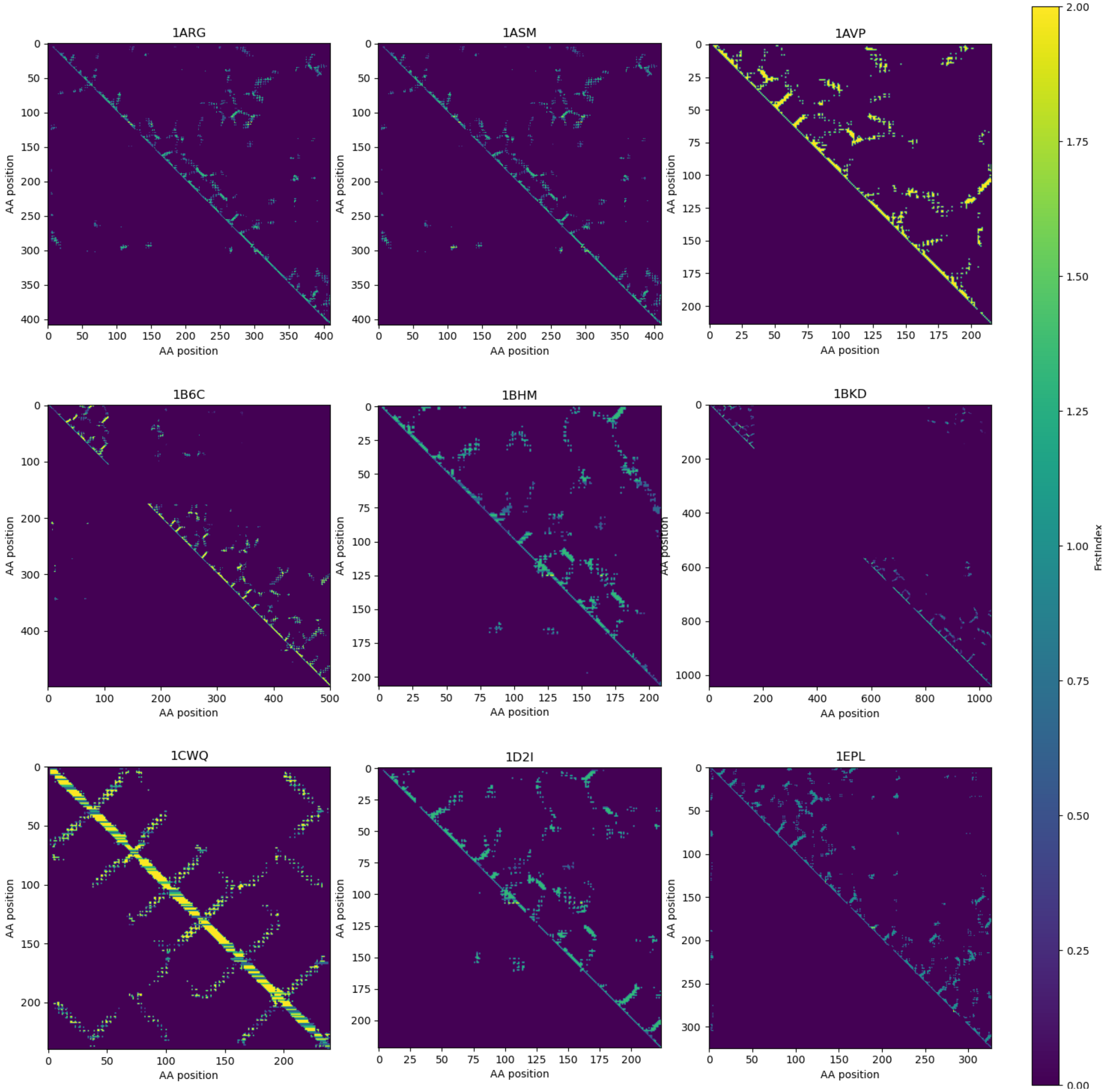


Figure 7. Contact maps for a selection bound DOR proteins, with the frustration in denoted by the colour. For positions with missing data the frustration is plotted as zero.