Computational and Structural Biotechnology Journal 19 (2021) 4919–4930

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Interpreting a black box predictor to gain insights into early folding mechanisms

Isel Grau [a], Ann Nowé [a,b], Wim Vranken [a,b,c,d,*]

[a] Artificial Intelligence Laboratory, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium
[b] Interuniversity Institute of Bioinformatics in Brussels, ULB/VUB, Triomflaan, 1050 Brussels, Belgium
[c] Structural Biology Brussels, Vrije Universiteit Brussel, Brussels 1050, Belgium
[d] VIB Structural Biology Research Centre, Brussels 1050, Belgium

ARTICLE INFO

Article history:
Received 15 March 2021
Received in revised form 23 August 2021
Accepted 26 August 2021
Available online 27 August 2021

ABSTRACT

Protein folding and function are closely connected, but the exact mechanisms by which proteins fold remain elusive. Early folding residues (EFRs) are amino acids within a particular protein that induce the very first stages of the folding process. High-resolution EFR data are only available for few proteins, which has previously enabled the training of a protein sequence-based machine learning 'black box' predictor (EFoldMine). Such a black box approach does not allow a direct extraction of the 'early folding rules' embedded in the protein sequence, whilst such interpretation is essential to improve our understanding of how the folding process works. We here apply and investigate a novel 'grey box' approach to the prediction of EFRs from protein sequence to gain mechanistic residue-level insights into the sequence determinants of EFRs in proteins. We interpret the rule set for three datasets, a default set comprised of natural proteins, a scrambled set comprised of the scrambled default set sequences, and a set of *de novo* designed proteins. Finally, we relate these data to the secondary structure adopted in the folded protein and provide all information online via http://xefoldmine.bio2byte.be/, as a resource to help understand and steer early protein folding.

1. Introduction

Proteins perform a multitude of essential functions in nature. A well-defined three-dimensional organization of their constituent atoms is often required to enable their function, with many proteins folding independently into such stable structures [12]. Other proteins might need chaperone help [33]or interaction partner(s) [22] to fold or might not fold at all [35]. In all cases, the protein sequence encodes its behavior and, by extension, the environmental context that is required for the protein to fold and/or function. From the different theories about how proteins fold independently [10,12,13,24], the concept of initial 'foldon' formation is now strongly supported by hydrogen–deuterium exchange (HDX) based mass spectrometry (MS) experiments [13,20]. Foldons essentially form through favorable interactions between amino acids close to each other in the sequence (early folding residues or EFRs), so further creating local structural elements that provide the right context for other residues in the protein to fold. They so help to determine the initial conformational states in the pathway towards the native fold, the end product of the folding process. Many studies have focused on determining which pathways lead to the final protein fold, with approaches describing the folding funnel, energy landscape and minimal frustration [3,4,9,11,14,17,23,26]. These processes are often directly related to the final protein fold, which is not necessarily the case for the early folding sites we predict here; they essentially describe the sites where local protein conformations are most likely to be restricted, which can influence the folding pathways that are preferentially followed by the protein. They therefore act prior to the formation of more complex local structures, and the interactions between those structures.

To gain insights in the early folding residues that drive very first stage of protein folding and the subsequent formation of foldons, the Start2Fold database was created [28]. This database collects data from pulsed labelling and related HDX experiments [28]. Based on the Start2Fold per-residue information from NMR experiments for a set of 30 proteins, the EFoldMine predictor [30] was developed to detect the location of likely early folding residues in a protein sequence. It does not, however, use the protein sequence directly, but instead depends on a 5-residue window of biophysical characteristics predicted from the sequence (backbone and side-chain dynamics, and helix, sheet and coil propensities). A support

* Corresponding author at: VUB, Pleinlaan 2, 1050 Brussels, Belgium.
E-mail address: wim.vranken@vub.be (W. Vranken).

https://doi.org/10.1016/j.csbj.2021.08.041

vector machine with a radial basis function kernel then uses the resulting 25 features to predict the location of EFRs from in a single protein sequence. Although support vector machines are known to be highly accurate classifiers based on strong mathematical foundations, the resulting model in multi-dimensional space is difficult, if not impossible, to understand by humans. This restricts the extraction of further knowledge about the determinants of early folding in proteins. An attempt was already made to obtain an interpretable machine learning predictor for early folding residues through the use of a Generalized Matrix Learning Vector Quantization (GMLVQ) algorithm [2]. The GMLVQ is a supervised technique, which can only learn from labeled data and obtains a matrix of relevant correlations between features that lead to the identification of classes, with the interpretability of this technique being a form of (paired) feature attribution. An alternative is the use of more intrinsically interpretable machine learning techniques such as decision trees or rule-based algorithms. However, these algorithms are generally less attractive in terms of performance compared to black boxes, and typically require large amounts of training data. Since EFR data, especially from NMR, are costly and time-consuming to obtain experimentally, it is unlikely that the limited EFR dataset that is currently available will grow extensively.

Therefore, to enable interpretation of the EFR determinants we propose a semi-supervised classification approach, where we leverage unlabeled and non-homologous protein sequence data for which protein structure data are available [37]. By labelling these data with EFR residues as identified by the 'black box' approach (using F for early folding and N for not early folding), we enlarge the interpretable training data, assuming it helps in elucidating the separation of the classes by interpretable classifiers. The goal is to obtain an interpretable model with better performance compared to only using experimentally labelled data, as well as obtaining a large dataset of (predicted) early folding data that can be analyzed statistically. Our self-labeling 'grey-box' (SlGb) approach [19] therefore aims to find a balance between accuracy and interpretability in a semi-supervised classification setting, so leveraging both labeled and unlabeled data, and providing a more flexible approach to interpretability [25]. In the learning process, the enlarged interpretable dataset is amended to avoid propagating misclassifications in the self-labeling. We experiment with rule-based classifiers as a proxy for interpretability, since these approaches are capable of providing both global holistic views of the model and local interpretations that explain a particular prediction. We show that the self-labeling grey-box approach achieves competitive results against the EFoldMine 'black box' in terms of sensitivity and specificity, through a leave-one-group-out cross-validation. Yet, it is able to represent the classification model with an average of 43 rules. Further analysis of these rules, combined with more classical analyses of the enlarged predicted dataset, enables us to gain mechanistic residue-level insights into the early folding process as well as a better definition of what constitutes an early folding fragment, which can provide useful information for protein design strategies. An overview of the datasets and essential elements of the strategy we developed is given in Fig. 1.

The prediction rules are fully interpreted for the SlGb approach, and analysed in relation to sequence patterns and secondary structure adopted in the folded protein, with all information provided via http://xefoldmine.bio2byte.be/, a resource for the community to help understand and steer early protein folding. Our interpretation confirms the importance of backbone rigidity for early folding (Pancsa et al., 2016a), and reveals the importance of inherent sheet propensity for the early folding residue itself, and strong helix propensity for the residue at position −2. This indicates that very particular specific restrictions on local conformations could be driving the formation of more stable local structures that then initiate the folding process.

2. Methodology

2.1. Datasets

The ground truth **start2fold** dataset is derived from the Start2Fold database [28] (http://start2fold.eu/), consists of 30 non-overlapping sequences and was described for the development of EFoldMine [30]. The **default** set is based on a set of proteins with less than 20% shared sequence identity, for which high-resolution (<1.6 Å) x-ray diffraction determined structures are available in the Protein Data Bank (PDB) [1,5,37]. These sequences were converted into their non-gapped equivalents, where residues are included that are missing in the PDB structures, by querying the PDBe API [36] and comparing the sequences. This resulted in a final set of 3020 proteins. The **scrambled** set is based on the default set and equally contains 3020 proteins, but the amino acids in each individual protein sequence were randomly scrambled, so retaining amino acid content but not their original order. The **denovo** set contains 98 proteins selected from the PDB that were designed *de novo*, so that have not evolved, and do not exist in nature. The selection was based on the search terms "*Primary citation author = Baker, D.*" and "*organism = Synthetic construct*". Proteins with a sequence length of less than 20 residues were subsequently removed, as well as entries with more than 80% sequence identity with any other protein in the dataset.

The secondary structure information for the **default** and **denovo** sets was acquired via the PDBe API and is based on the DSSP analysis [21] of the protein structure information, where only the H (Helix), E (sheet) and C (coil) classes are retained. For all sequences in all three sets, the backbone and sidechain dynamics, as well as the helix, sheet and coil secondary structure propensities, were predicted using DynaMine [6] and derived tools [30], with the original EFoldMine predictions (**ef**) also calculated. The newly developed black box predictor (**bb**) and the grey box predictor (**SlGb**) were also applied on all datasets.

2.2. Data analysis on predictions

For each dataset and each prediction, every amino acid was further labelled whether it was at a single (**S**) early folding or non-early folding residue, or whether it appeared at the beginning (**B**), middle (**M**) or end (**E**) of a given early folding or non-early folding fragment (Fig. 2). For each residue in each protein, a comparison was then performed on the overlap between the **ef**, **bb** and **SlGb** predictions, in terms of true/false positives and true/false negatives.

In the next step, all proteins were, per dataset, divided into 5-residue fragments (the window size used for the predictors), and each such amino acid fragment was classified based on its secondary structure string (*e.g.* CCHHH), with pre-N-terminal and post-C-terminal positions designated with a '+' character (*e.g.* an N-terminal residue could appear as ++CCC). Fragments were clustered together if there was only one change between H or E to C states, or vice versa (*e.g.* HHHHC and HHHHH), on the assumption that the DSSP software can mis-assign the secondary structure designation for one residue, as well as increasing the size of the fragment pools to obtain better statistical comparisons. For each prediction type, the overall ratio of predicted early folding residues to the total number of residues in that dataset was calculated and used as the cutoff for a binomial statistical test per (clustered) secondary structure fragment. Fragments that had a statistically

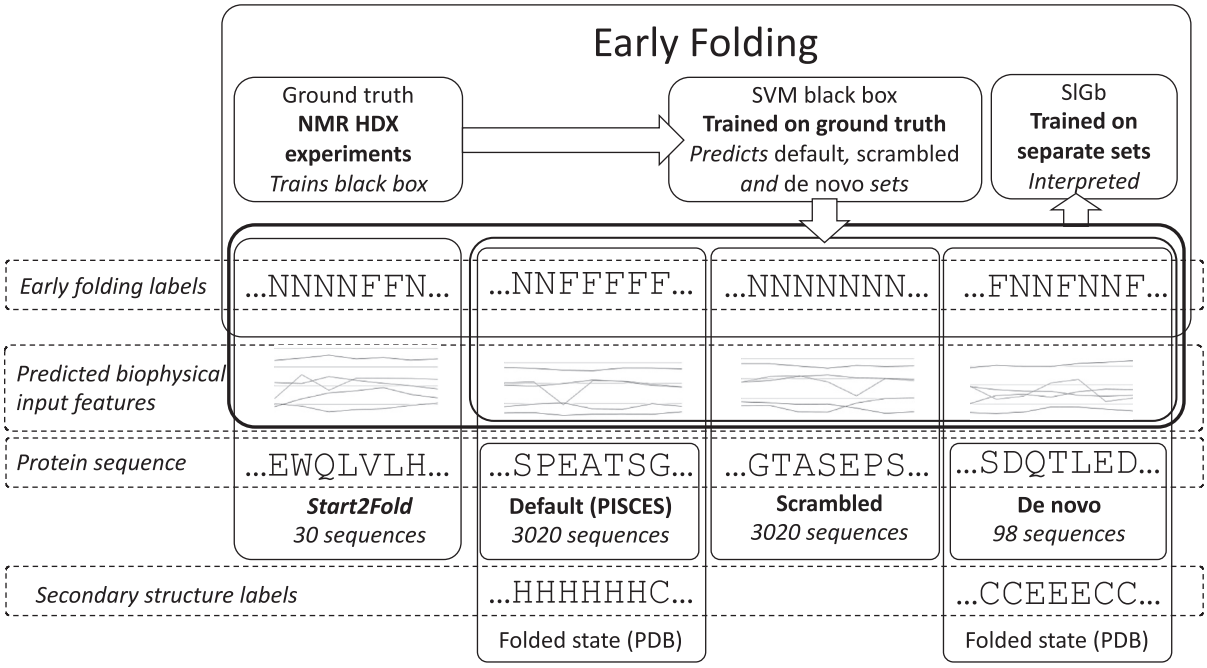

Fig. 1. Overview of the datasets and workflow of XEFoldMine (eXplainable EFoldMine). In a first step, ground truth labelled data is used for training a black box support vector machine classifier. In a second step, the black box is used as a component of the SlGb classifier for predicting the labels (F or N) of default, scrambled and de novo datasets (self-labelling process). In a third step, the SlGb model expressed in rules is further interpreted and finally, a per-fragment analysis using the secondary structure is made.

significant higher number of central early folding residues (using a p value cutoff of 0.05) were classed as **'over-represented'**, conversely fragments with a significantly lower than average number as **'under-represented'**, with all other fragments labelled **'neutral'**.

For all fragments, a Sankey plot was generated where the inter-amino acid connections can be explored (based on the amino acids in each 5-residue fragment), a Chord diagram and heatmap to visualize the rules for the per-fragment trained predictor, as well as an

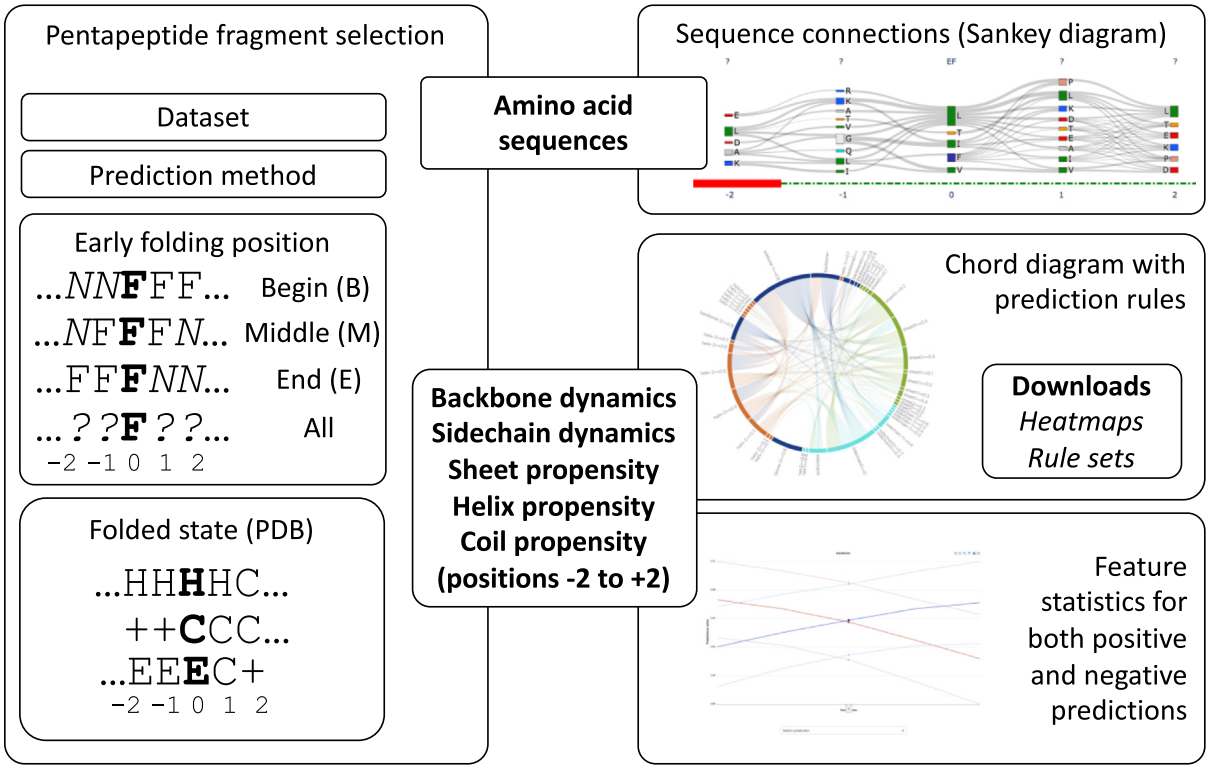

Fig. 2. Overview of the available data interpretation and analysis accessible from http://xefoldmine.bio2byte.be/. Sankey **plot are available for all fragments where the inter-amino acid connections can be explored. Chord diagrams and heatmaps** visualize the rules for the per-fragment trained predictor, as well as an overview of the distributions of the features used for the prediction for these fragments.

overview of the distributions of the features used for the prediction for these fragments (backbone and sidechain dynamics, and secondary structure propensities, for each of the 5 positions of that fragment) (Fig. 2). This information is available via http://xefold-mine.bio2byte.be/.

2.3. Self-labeling grey-box classifier

The prediction problem at hand is a semi-supervised classification setting since unlabeled data is available in a high quantity and labeled data is limited due to a costly process of labeling. There exist several machine learning algorithms for tackling semi-supervised classification problems, including transductive support vector machines which are the extension of vanilla support vector machines for this type of problem. However, transductive support vector machines and the majority of semi-supervised classifiers consist of complex ensemble structures which lack interpretability features.

The SlGb approach [19] is a recently proposed semi-supervised classification strategy for building a model which requires a certain degree of interpretability. In a first step, a black-box classifier is trained with available labeled data and used to predict the decision class of unlabeled instances in a process called self-labeling. Given the successful performance of EFoldMine predictor [30] we choose a support vector machine-based classifier for the black box component of the grey box model. From this step, we obtain an enlarged training set comprising the originally labeled instances and the extra labeled ones. An amending procedure [18] is used for weighting the instances according to the confidence in their self-classification. Afterward, a surrogate white-box classifier is used to build an interpretable predictive model based on the enlarged dataset. The aim is to outperform the base white-box component using only the originally labeled data while maintaining a good balance between performance and interpretability.

Based on EFoldMine and using the same choice of hyperparameters as reported in [30] a black-box component for the classification of unlabeled instances is developed (**bb**). The choice of white box is guided by the type of interpretability we want to obtain, either a global tree structure that allows to inspect the model as a whole, or a set of rules that describe the decision space. We consider tree white box methods in our experiments: C4.5 decision trees (**c45**) [29], partial decision lists (**part**) [16] and propositional rule learning (**rip**) [8]. The hyperparameters reported for each model were determined by hyperparameter optimization, using a grid of possible values on each case, minimizing complexity while retaining prediction performance [18].

Decision trees. We use the C4.5 algorithm for inducing rules in the form of a pruned decision tree. The hyperparameters used are: two as the minimum number of objects per leaf, 0.25 as the confidence factor for pruning, and the use of subtree raising operation when pruning. The flow-like structure of decision trees allows dividing the space with a series of tests on each attribute until reaching a conclusion, providing a global view of the model as well as the possibility of simulation from input to output by a human. This transparency and simulatability are proxies for claiming the interpretability of decision trees.

Partial decision lists. Partial decision lists use a separate-and-conquer strategy for building rules. It generates a partial C4.5 decision tree in each iteration, makes the "best" leaf into a rule, separates the instances covered by this rule, and repeats the process until all instances are covered. The hyperparameters are idem to the decision trees above. Decision lists are a set of rules which should be interpreted in order. It generally starts covering the cases from more general rules to more specific. The algorithm is still transparent, and the outputs can be computed by a human being using the model.

Propositional rule learning. This propositional rule learner is based on association rules and implements reduced error pruning. The training data is split into a growing set and a pruning set. The rules obtained from the growing set are simplified by pruning operators, such as it yields the greatest reduction of the error on the pruning set. The hyperparameters used for learning are: two as the minimum total weight of instances in a rule, three folds where one is used for pruning, and the other two for optimization steps. The decision list that this model produces covers the rarest instances first; therefore, is more suitable for explaining the classification of minority classes. It should also be interpreted in order, similar to the partial decision lists. Like decision trees or partial decision lists, the manual simulation from input to output to explain the outcome of a particular instance is possible.

In general, the IF-THEN structure of the explanations is straightforward interpretable, although the total number of rules or the rules' length could affect this advantage. Decision lists tend to produce more compact rule sets than decision trees while being similarly expressive. The models generated by the chosen white boxes are generally sparse, which is another desideratum for interpretability since they select only the relevant features for the model.

3. Results

3.1. Performance comparison between the different predictors

We first determine which configuration of the SlGb approach performs best compared to the baseline EFoldMine predictor and the white-box baselines. The baseline models are trained on labeled data only (start2fold dataset). We perform leave-one-group-out cross-validation where each fold groups one labeled protein from start2fold and an equal share of unlabeled proteins per fold from the unlabeled datasets (default, scrambled, or denovo). This selection results in a 27-fold cross-validation with residues from one labeled protein and 112 unlabeled proteins (for default and scrambled datasets) or three or four unlabeled proteins (for the denovo dataset). Only the ground truth labelled instances distributed over the different CV folds are used for validation. The different unlabeled datasets were so chosen to investigate whether there are differences among the rules sets (and therefore the predictions) obtained for natural proteins (default), rationally designed sequences based on natural proteins (denovo) and nonsense proteins that still contain the same amino acids as natural proteins (scrambled).

In order to compare the results, we use several measures for evaluating the prediction performance. Since the binary classification problem at hand is highly imbalanced and the minority class (positive or "early folding") is the label of interest, we opt for exploring several measures beyond accuracy. The sensitivity and precision indicate the prediction power of the model regarding the positive class, while specificity focuses on the negative prediction power. To complement the overall accuracy, we include other measures that consider the class imbalance, such as the balanced accuracy, Mathew's correlation coefficient, and Cohen's kappa.

In terms of interpretability, we use the number of rules as an indicator of the complexity of the model. The fewer the number of rules, the more interpretability potential the resulting model has. We also measure the agreement of the SlGb predictions with black box prediction on the data as an indicator of the fidelity of the explanations provided by the SlGb. For measuring the fidelity, we use Cohen's kappa as an agreement measure [7]. Observe that this measure is reported as fidelity in the table, which is different from the kappa

obtained as a prediction performance measure (which indicates the agreement with the ground truth). Table 1 shows the averaged performance across folds of the cross-validation, for the EFoldMine predictor (**ef**), the black box predictor built for this study based on EFoldMine (**bb**), the white box baseline models trained on labeled data (**c45**, **part** and **rip**), and the three configurations of the grey-box model (**SlGb_c45**, **SlGb_part**, **SlGb_rip**), for each dataset combination.

The first section of the table shows that the black box component is able to reproduce the performance of the EFoldMine predictor, which makes it a trusted model for the self-labeling process. The second section of the table contains the results of the baseline white boxes trained on the start2fold dataset, which shows that the propositional rule learner classifier (**rip**) obtains the best results. **Rip** outperforms other white boxes when predicting the positive class and overall considering the imbalance of the dataset, with a considerably simpler model in terms of the number of rules. In general, the white-box baselines are less accurate than the black-box approaches, particularly for detecting true positive instances (see sensitivity and precision). In contrast, each SlGb configuration is able to outperform its white-box baseline, especially for the minority positive class "early-folding", at the cost of an increase in the number of rules. This trade-off between accuracy and interpretability is inherent to grey-box models. Here, the performance of the SlGb is competitive with EFoldMine while it remains an interpretable model. These results support using a semi-supervised grey-box approach since the SlGb is clearly profiting from the unlabeled instances for improving its performance on detecting the minority class while also being an interpretable model.

The last three sections of the table show that decision lists (**part** and **rip**) are preferred compared to decision trees (**c45**). These models achieve competitive or better performance with a significantly lower number of rules, making the grey-box more transparent. In particular, the SlGb configuration using propositional rule learning (**SlGb_rip**) obtains the best results in terms of interpretability measures with the lowest number of rules and high fidelity to the black-box prediction for all unlabeled datasets. Therefore, for the rest of the analysis, we choose the SlGb configuration using a propositional rule learner as the white-box component (**SlGb_rip**).

3.2. White box interpretation

The SlGb approach allows obtaining both global and local interpretability of the machine learning model. Local explanations can be derived for the prediction of a residue by inspecting the rule that leads to that prediction. Fig. 3 shows the structure and interpretation of the rules produced by SlGb.

A global view of the predicting model can also be obtained by inspecting the decision list as a whole. For example, when using the denovo data together with the start2fold dataset for training, the resulting SlGb model contains 24 rules, with an average of 3.66 conditions in the antecedent. We can then analyze the relative frequency of the features appearing in the rule antecedents. Features such as backbone in position −1 and 0 (with 0 being the position that is predicted, see also Fig. 2), helix in position −2 and sheet in position 0 have the highest frequency for this model. Almost all rules require the backbone feature (in combinations of position −2, −1, or 0) to have a high value (>=0.8 indicates a rigid backbone) in order to predict a residue as positive, which suggest that this is a strong condition associated with early folding, as already indicated by a previous analysis [27]. From the rules, it is also evident that the early folding residues combine the high values of the backbone feature with medium values (in the range from 0.3 to 0.7) of sheet0 and helix-2 features.

The SlGb_rip model obtained from the scrambled dataset contains a higher number of rules (146 rules with an average of 4.69 conditions in the antecedent), which is expected as this dataset contains a much higher number of unlabeled proteins than the denovo dataset, resulting in more complex decisions to cover the wider sequence variety. However, the patterns in the frequencies of the features and their cutoffs are relatively similar to the denovo model. We use the Jaccard similarity index, weighted by features frequency [15], to estimate this similarty and obtain a value of 0.74, which could be considered high. Similar to denovo model, the scrambled model relies on high values of backbone feature interacting with helix but focuses more on the values of sheet and sidechain features for detecting early folding residues.

The decision list for the SlGb model for the default dataset contains 137 rules, with an average of four conditions in the antecedent. When analyzing the relative frequency of the features in the antecedents we observe similar patterns compared to the scrambled dataset and, to a lesser extent, the denovo dataset. In this model the backbone, sheet and sidechain features (all in position 0) have a stronger role in the rules, when compared to denovo. The Jaccard similarity index between denovo and default rule models is 0.69, which means both models are relatively similar in the selection of the features for the antecedents of the rules. The default model shares more similarity with the one obtained from the scrambled dataset (jaccard index = 0.83). From the first rule of the default model "*(backbone0 ≥ 0.87) and (sidechain0 ≥ 0.61) and (sheet0 ≥ 0.45) and (sidechain-1 ≥ 0.63) ≥ class = F (12172.24/20.39)*", we can already derive the characteristics that distinguish roughly 10% of the early folding residues of this dataset. A similar analysis can be made for the rest of the rules.

To facilitate the interpretation of bigger models such as the SlGb based on the default dataset, we created heatmaps and chord diagrams for illustrating. The heatmaps show the strength of the pairwise interactions in the rules, based on the number of early folding residues that are correctly predicted. In this way, we illustrate common paired characteristics of the early folding residues. Fig. 4 shows a reduced version of the heatmaps for the default (a) and denovo (b) models. These heatmaps are focused on regions with high interaction between features. For example, for the default dataset, the backbone feature in positions −1 and 0 has a strong interaction with helix-2 and sheet, with approximately 17,000 early folding residues having backbone values greater than 0.8 and sheet values greater than 0.4. In this dataset the sheet in position 0 has a much stronger role in detecting early folding residues than in the denovo dataset. For the denovo dataset, the backbone interaction in position −1 is stronger for medium values of the helix in position −2 and sheet, compared to the backbone in position 0.

The chord diagram is an extended and interactive version of the heatmap, including all interactions without showing sparse regions. Fig. 5 shows a screenshot of the chord diagram corresponding to the rule model for the default dataset. Here, the pairwise interactions between conditions in the antecedents are represented as chords of the circumference. The wider the chord between two features, the more early-folding residues it distinguishes correctly. Besides the pairwise interaction between conditions, these diagrams show the strength of an individual feature by highlighting all its interactions and the total number of early folding residues that share this characteristic. This strength is represented by the length of the arc in the circumference and can be interpreted as a feature attribution measure. The wider the arc, the more support and confidence this condition has from the rules, and therefore the better it helps to distinguish early folding residues. In the screenshot, we highlight the interactions of the backbone with other features. However, it can be observed that

Table 1
Average cross-validated performance comparison between EFoldMine predictor and several configurations of the SlGb using sensitivity (sen), specificity (spe), accuracy (acc), balanced accuracy (bac), precision (pre), Mathew's correlation coefficient (mcc), area under the ROC (auc), Cohen's kappa (kap), the number of rules, and the fidelity (fid) of the SlGb white-box component to the black-box predictor. The best performing models per section are highlighted in bold.

dataset	model	sen	spe	acc	bac	pre	mcc	auc	kap	rules	fid
start2fold	ef	0.73	0.76	0.74	0.74	0.36	0.35	0.81	–	–	–
start2fold	bb	0.74	0.73	0.71	0.73	0.34	0.33	0.80	0.28	–	–
start2fold	c45	0.53	**0.77**	**0.72**	0.65	0.29	0.22	0.64	0.19	165.74	–
start2fold	part	0.59	0.69	0.66	0.64	0.26	0.20	0.68	0.16	37.96	–
start2fold	rip	**0.66**	0.70	0.68	**0.68**	**0.30**	**0.26**	**0.71**	**0.22**	**10.78**	–
start2fold + default	SlGb_c45	0.63	**0.77**	**0.73**	0.70	**0.34**	**0.30**	0.71	**0.26**	207.30	0.75
start2fold + default	SlGb_part	**0.71**	0.71	0.70	**0.71**	0.32	**0.30**	**0.75**	0.25	54.93	0.75
start2fold + default	SlGb_rip	0.69	0.72	0.71	**0.71**	0.33	**0.30**	0.72	0.25	**43.41**	**0.76**
start2fold + scrambled	SlGb_c45	0.62	**0.76**	**0.72**	0.69	0.32	0.28	0.70	0.24	207.85	**0.78**
start2fold + scrambled	SlGb_part	**0.70**	0.71	0.70	**0.71**	**0.33**	**0.31**	**0.75**	**0.26**	51.11	**0.78**
start2fold + scrambled	SlGb_rip	0.69	0.73	0.71	**0.71**	**0.33**	0.30	0.72	**0.26**	**42.74**	0.76
start2fold + denovo	SlGb_c45	0.47	**0.78**	**0.72**	0.62	0.29	0.19	0.62	0.17	172.74	0.65
start2fold + denovo	SlGb_part	0.62	0.74	**0.72**	0.68	**0.31**	0.27	**0.71**	**0.23**	37.37	**0.67**
start2fold + denovo	SlGb_rip	**0.70**	0.70	0.69	**0.70**	**0.31**	**0.28**	**0.71**	**0.23**	**18.33**	**0.67**

residues with sheet values greater than 0.4 or sidechain values in the range from 0.3 to 0.6 are also detected as early folding.

The decision lists for default and denovo SlGb models, as well as the heatmaps and chord diagrams, are available online via http://xefoldmine.bio2byte.be/.

3.3. Differences in predictions and rulesets for the datasets

There are differences between the prediction results of the different predictors in relation to the datasets (Table 2). The **bb** and **SlGb_rip** predictors have roughly similar overall ratios of positive to negative predictions in all three datasets, whereas the **ef** predictor shows larger differences, and compared to default predicts many more positives in the denovo dataset, and fewer in the scrambled dataset. It is therefore difficult to make conclusions about the likely relationships between the number of early folding sites in the natural, scrambled and *de novo* sequences. All predictors, however, show the same trends in relation to the final secondary structure state of the residues, with residues that fold into beta sheets containing the most early folding residues, and residues ending up as coil by far the least. This trend is more pronounced in the default dataset than in the denovo dataset, indicating that helices in *de novo* designed proteins might be, compared to natural proteins, 'overdesigned' with respect to the number of early folding sites they contain.

A comparison of the ratio between positive (early folding) and negative residues on a per-amino acid basis between the 3 datasets for the **SlGb_rip** predictions (Fig. 6) shows that certain amino acids (C, F, I, L, V, W, Y) are more likely to be present in early folding sites, with the ratio in general similar between the datasets, except for C, G and H, which are underrepresented as early folding residues in the denovo dataset. This indicates these amino acids might not be used in their 'natural' context in *de novo* proteins, but are rather positioned only in specific contexts (*e.g.* G is only primarily in turns and loops, whereas it can occur in secondary structure elements in natural proteins).

3.4. Per-fragment analyses

In addition to the overall analysis of the prediction model, we explored the relationship between the early fold predictions and the secondary structure as adopted in the folded protein. Based on 5 residue fragments, with the central residue either predicted as early folding (F) or not (N), we retrained SlGb models on secondary-structure derived subsets of the data, if enough available, and enable full access to this per-fragment information (see also Fig. 2):

- Amino acid sequence information for the fragments, for both F and N predictions, in the form of a **Sankey diagram** that visualizes the connections between the most commonly observed amino acids in this fragment
- A **chord diagram** that gives an overview of the SlGb interpretation of why residues in this fragment are predicted as early folding
- An overview of the **statistics of the 5 input features** for the prediction for 5 fragment positions from −2 to +2, with the positive (F) fragments in red, the negative (N) fragments in blue. This enables, per input feature, to visualize where there are consistent differences between the fragments predicted as F or N.

For example, if you are interested to see where early folding starts in helical fragments of natural proteins, first select the 'default' dataset, then on the left-hand side of the page select the

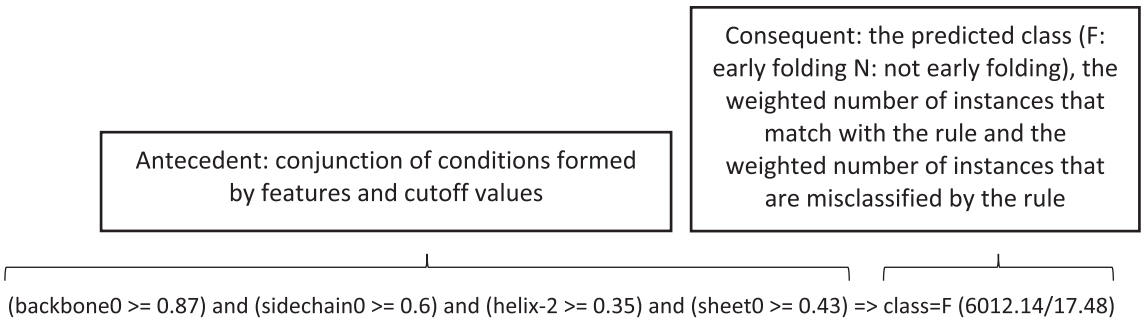

Fig. 3. Structure and interpretation of an SlGb rule. This example rule was extracted from the SlGb model that was built using the start2fold and default datasets.

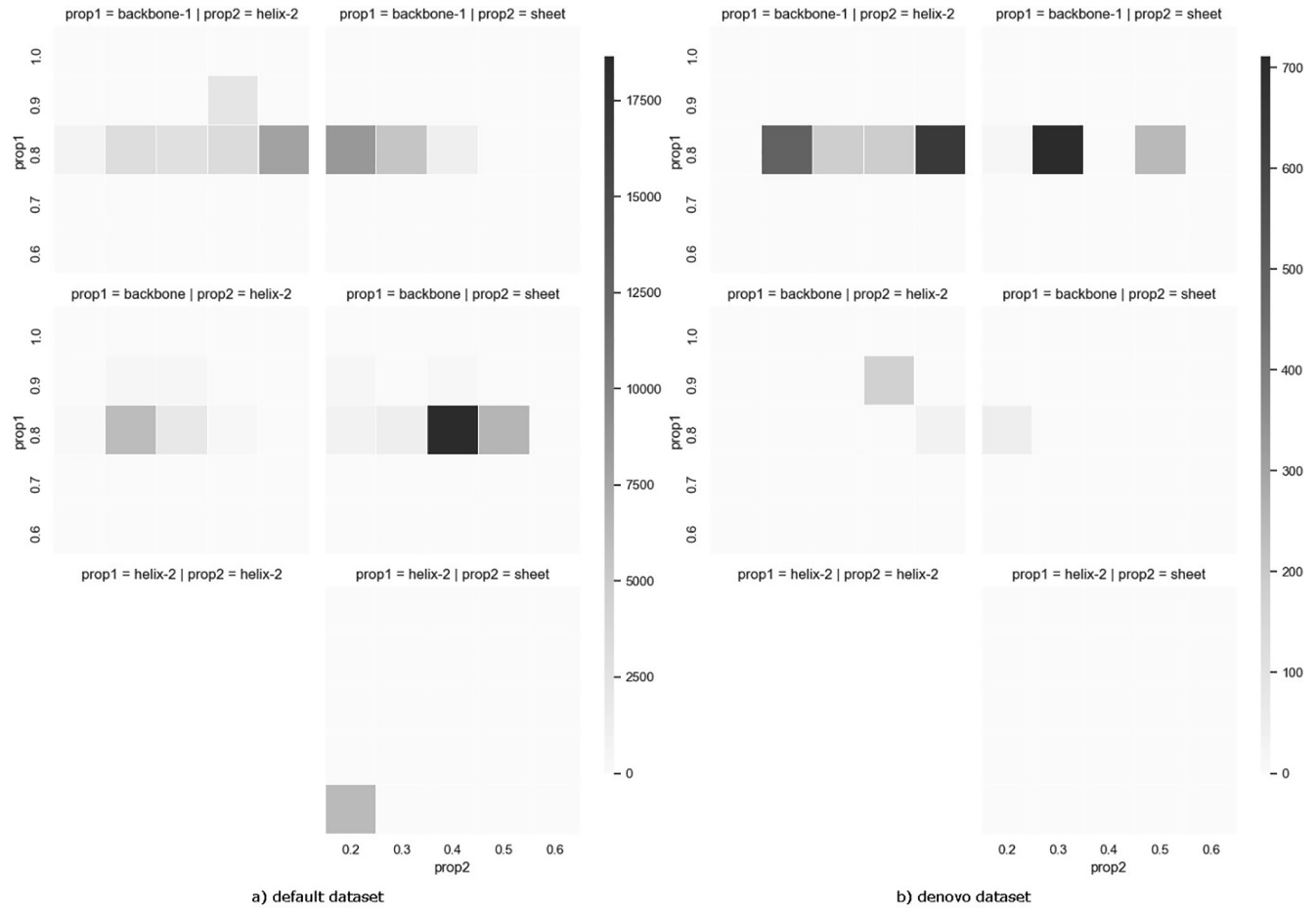

Fig. 4. Heatmaps of pairwise interactions of features (and their cutoffs) in the antecedents of the rule models built from a) the default dataset and b) the denovo dataset. These heatmaps are focused on regions with high interaction between features. For the default dataset, the backbone feature in positions −1 and 0 has a strong interaction with helix-2 and sheet. Meanwhile for the denovo dataset, the backbone interaction in position −1 is stronger for medium values of the helix in position −2 and sheet, compared to the backbone in position 0.

prediction method (*e.g.* EFoldmine), the EF fragment as 'Start', the Folded state as 'HHHHH', and press '*Submit*'. The next page will display the information for these specific fragments, bearing in mind that for this statistical analysis the fragments' CHHHH', 'HHHHC', 'HCHHH', 'HHCHH' and 'HHHCH' were also included, with 'HHHHH' as reference (see Methodology). The top text indicates that the start of early folding in these fragments is significantly over-represented, so is more common than average, and occurs in 25,668 fragments, with the central residue early folding 13,632 times, and not early folding 12,036 times (this represents the cases where a non-early folding fragment starts in a full helical fragment).

The **Sankey diagram** shows that L, A, V, I, F, Y and T are commonly found in the position where early folding starts (position 0), followed by the hydrophobic A, V, I, L amino acids at position + 1, and a mix of charged and hydrophobic amino acids at position + 2 (R, L, K, E, A). In contrast, position −1 is enriched in negatively charged residues (E,D), alanine (A) and to a lesser degree R, K, Q and L, with a similar mix in position −2. Hovering over an amino acid shows which other amino acids are found in relation to that one; for example, V as a central residue is found with A, E and D in position −1, but not with the other amino acids. Hovering over the link also displays the overall occurrence (*e.g.* E in position −1 and V in position i occurs 250 times in the fragment dataset). Changing the EF fragment setting on the left to 'End' and pressing '*Submit*' again will update the page to now show different residue preferences, in this case, for the end of an early folding fragment in helices, hydrophobic residues are strongly preferred at position −1, and negatively and positively charged residues at position + 1. It is also possible to display the negative set (the start of a non-early folding fragment in full helix) by selecting 'Non-early folding' under the Sankey diagram heading.

The per-secondary-structure fragment **chord diagram** is only displayed for the SlGb_rip predictions, as they were directly derived from the rulesets thereof. The content is the same as the chord diagrams for the overall predictions, but in this case relates to a re-training on only elements of a particular secondary structure fragment. This enables specific exploration of the features leading to the prediction of early folding residues for a particular secondary structure fragment, including the ability to distinguish whether this relates to the start, middle or end of an early folding fragment. For 'HHHHH' fragments, for example, the rules for predicting early folding when at the start of an early folding fragment show that the 0 and 1 positions are the most relevant (high backbone, sheet, sidechain and helix predictions), whereas for the end of an early folding fragment, the −2 and −1 positions are dominant (again for the same features).

Finally, the overview of the **statistics of the 5 input features** over positions −2 to +2 are shown as the final plot, with the

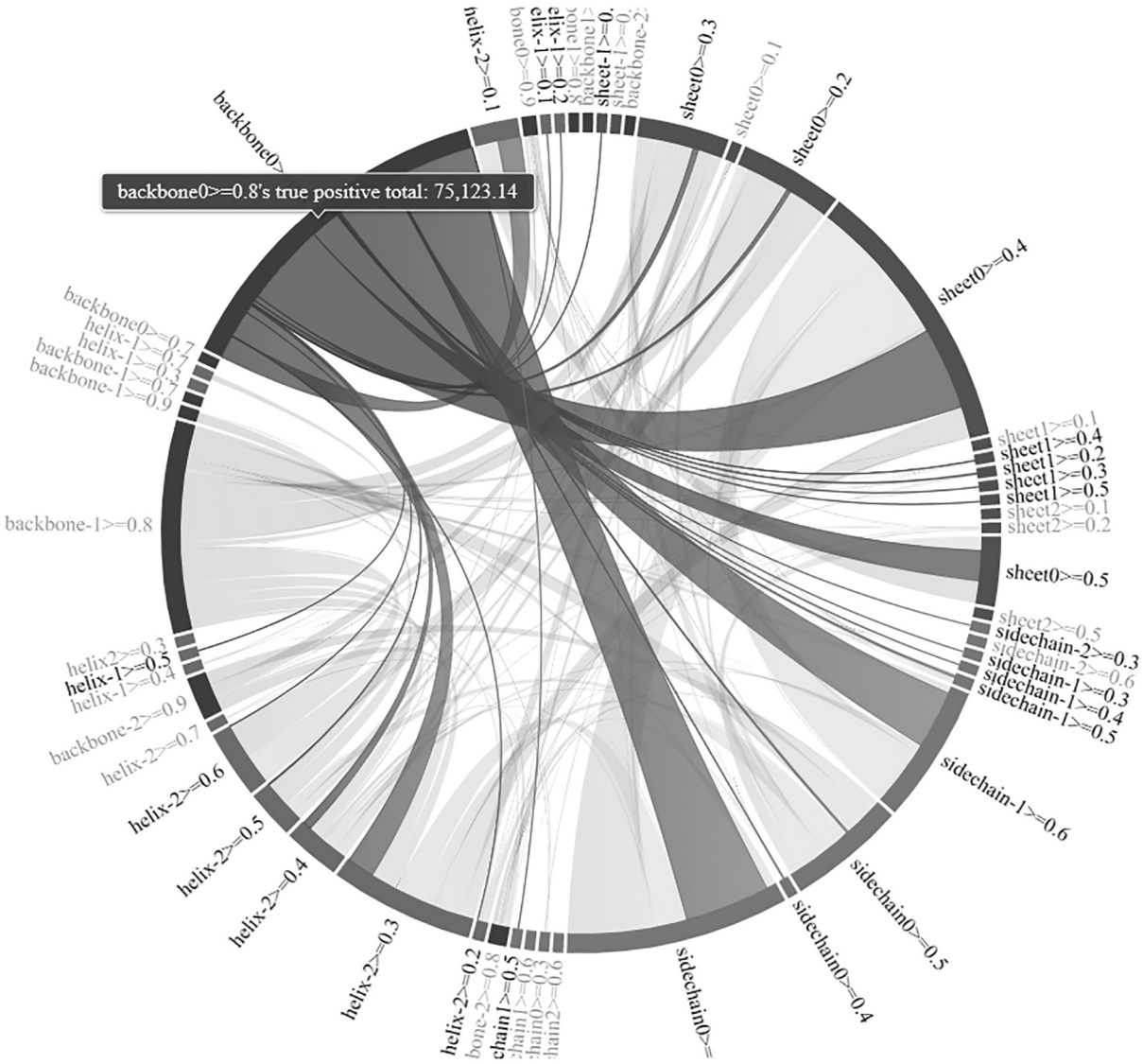

Fig. 5. Screenshot from the chord diagram for the rule model of the default dataset. The interaction of the backbone feature with other features is highlighted, contributing to detect a high number of early folding residues. A coloured and interactive version of this diagram is available online via http://xefoldmine.bio2byte.be/.

median, first and third quartile of the distributions of the predicted biophysical feature input values for positive (F) fragments in red, the negative (N) fragments in blue. By default, the 'backbone' values are shown, with the bottom selection box enabling switching between predictions. For the 'HHHHH' fragment, for example, shows for the start of an early folding fragment that the backbone dynamics predictions are very similar at position −2, but that for position +2 the backbone dynamics predictions are higher (more rigid backbone). A similar trend is present for helix and sheet propensities. These trends are reversed for the end of early folding fragments, and for the middle of early folding fragments for 'HHHHH', consistently higher backbone, helix and sheet values are present. The elevated sheet propensity seems to be consistently present in helices, which indicates that sheet propensity is important to create early folding fragments. This is also captured by the rule sets, as visualized by the chord diagrams. For 'EEEEE' fragments, the helix propensities trends are similar, but only somewhat elevated after the start of the fragment, or before the end, whereas the sheet propensity differences are especially pronounced.

3.5. Case study

To illustrate the rule set in a qualitative example, we compare the backbone dynamics, helix and sheet propensity input features for myoglobin (1myf) and leghemoglobin (1bin) sequences (Fig. 7), as also discussed in the original EFoldMine article [30]. Both are all-helical proteins, but have notably different processes of folding.

The first helix (A dark red box, Fig. 7) folds early in myoglobin, but only folds later in leghemoglobin; this agrees with the original EFoldMine predictions [30]. Although the predicted backbone dynamics and helix predictions are in similar ranges for both proteins, the sheet propensities are much lower in leghemoglobin, which fits with the rule set that indicates sheet propensity is important for early folding in helices. The second helix (B, red box) folds early in both proteins, again correctly predicted by EFoldMine, with the backbone dynamics and helix/sheet propensities adopting very similar values in both proteins, again in adherence with the overall rule sets for high helix propensity at position −2 and high sheet at position 0. Finally, the E helix (grey box) folds early only in leghemoglobin, which has correspondingly

Table 2
Overview of the ratio of positives to negative early fold predictions for the ef, bb and SlGb_rip predictions for all residues, subdivided by residues that form helix, sheet and coil secondary structures in the final fold.

predictor	dataset	overall ratio	helix ratio	sheet ratio	coil ratio
ef	default	0.335	0.438	0.715	0.107
	scrambled	0.239	–	–	–
	denovo	0.597	0.862	0.939	0.089
bb	default	0.289	0.350	0.634	0.099
	scrambled	0.293	–	–	–
	denovo	0.321	0.420	0.534	0.036
SlGb_rip	default	0.280	0.336	0.612	0.100
	scrambled	0.291	–	–	–
	denovo	0.258	0.331	0.428	0.050

higher predictions for the backbone dynamics (more rigid) and a region where helix and sheet propensities are concurrently elevated, whereas in hemoglobin the peak in sheet propensity precedes the peak in helix propensity. This indicates that the understanding derived from the SlGb rules can help us design, for example, computational mutation studies which aim to modify sequences in ways that affect the folding pathway, in this case a possibility would be to change the sequence of the A helix in leghemoglobin to increase its sheet propensity without affecting other biophysical characteristics, and while still agreeing with the overall fold.

A second case study illustrates the meaning of the early folding predictions on superoxide dismutase 1, on which a detailed experimental folding, unfolding and misfolding analysis was performed by [34]. In this protein, combinations of the first four beta strands (β1-β4) form an initial stable core from the unfolded state. Misfolding of the protein then happens after formation of this stable core, with the partially native stable core intermediates mediating misfolding. Especially loops IV (between β4-β5) and VII (between β7-β8) are involved in formation of an aggregation prone interface when flexible. The first four β -strands indeed all have very high early folding and strand propensity (Fig. 8), in line with them forming the stable core. These properties are in general lacking for the region after β4, with only β7 having similar properties, but there are no other early folding regions close to it except for the region just after β6. This indicates that there might not be enough conformationally restricted regions around β7 to create an initial foldon. The loop IV and VII regions lack early folding propensity, indicating that they are dependent on the folding of other regions of the protein to become conformationally restricted, in line with the experimental data.

4. Conclusion

The analysis we present here shows that our interpretation of black box machine learning can reveal which combinations of input features lead to positive predictions, in our case early folding residues. The biophysical meaning of these input features in turn enables us to infer the main determinants for the formation of early folding fragments: especially a more rigid backbone for the central residue (0) and subsequent residue (positions +1), in combination with a strong propensity for sheet conformations for those residues, and a strong helix propensity for the residue at position −2.

It is important to stress that with the term ‘early folding’ we are referring to transient processes, where the protein is initially unfolded, and where the predominant interactions are local,

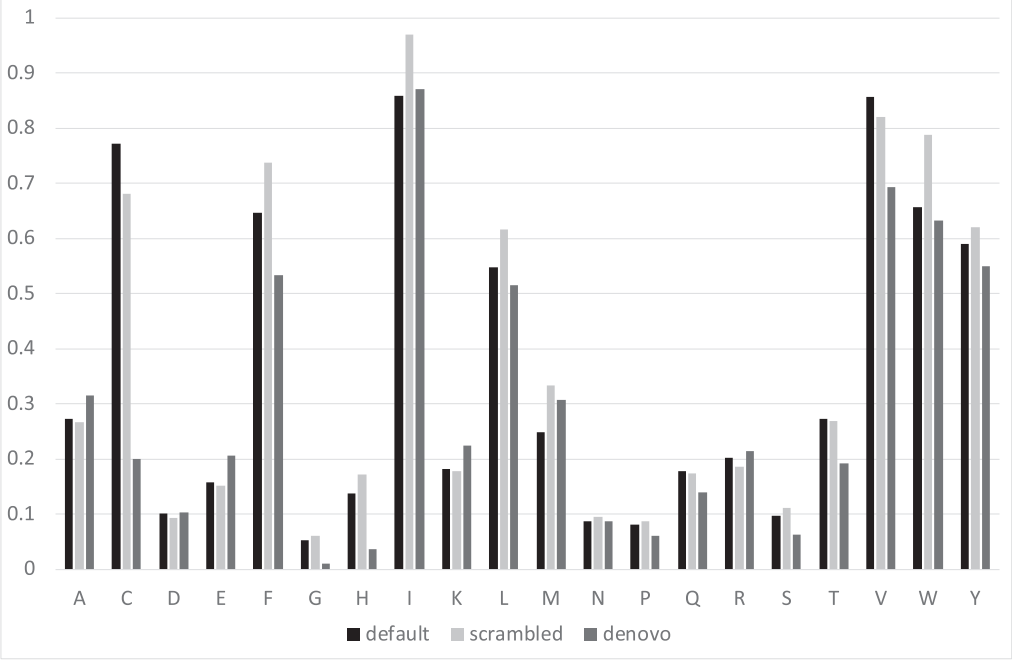

Fig. 6. Ratio between positive and negative predictions for SlGb_rip for the 20 natural amino acids per dataset. Certain amino acids (C, F, I, L, V, W, Y) are more likely to be present in early folding sites, with the ratio in general similar between the datasets, except for C, G and H, which are underrepresented as early folding residues in the denovo dataset.

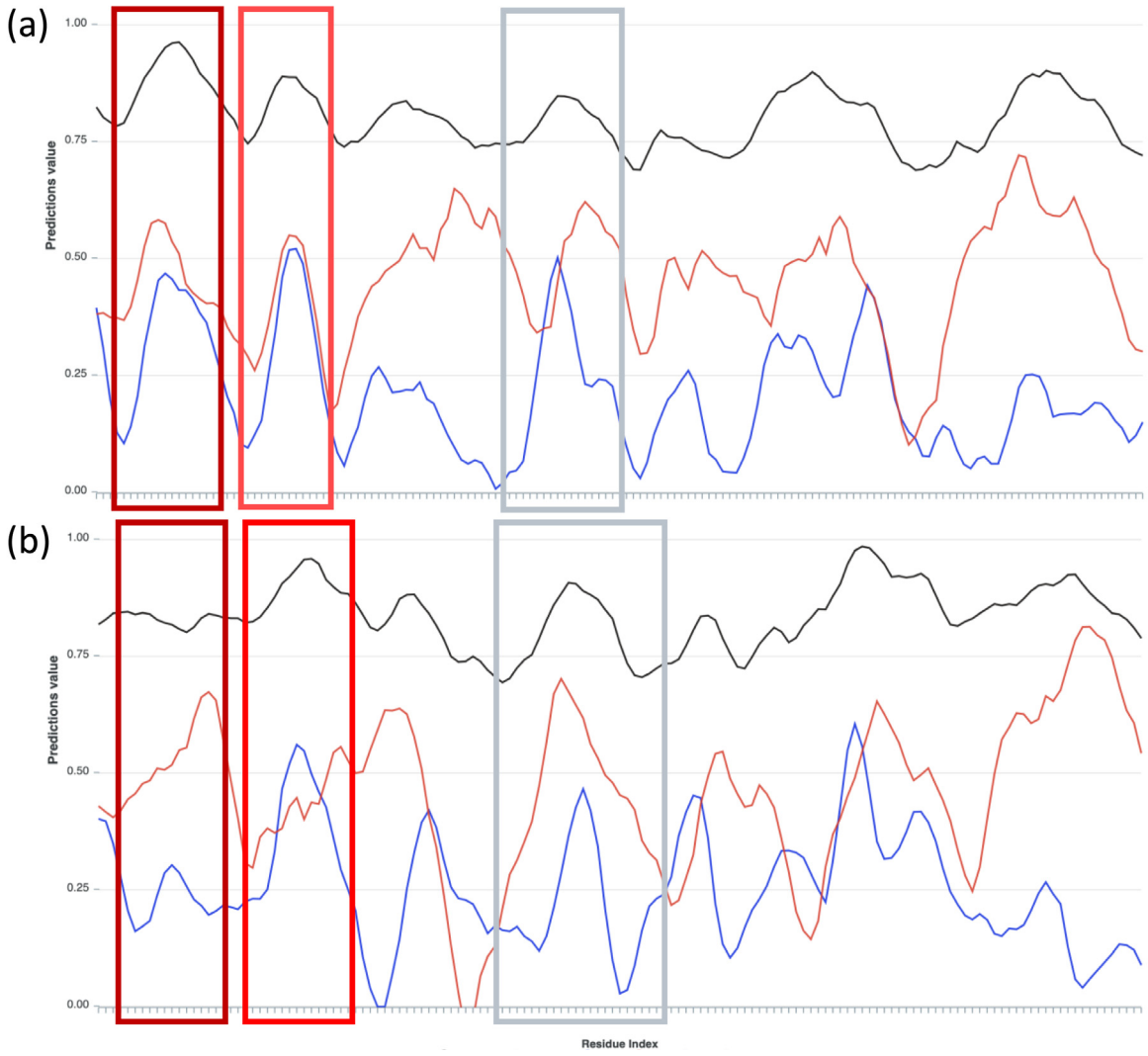

Fig. 7. Comparison between the backbone dynamics (black), helix (red) and sheet (blue) propensities for myoglobin (a) and leghemoglobin (b), which have similar overall folds while folding differently. The regions indicating the matching A helix (dark red box), B helix (red box) and E helix regions (grey box) are indicated. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

between residues close to each other in the amino acid sequence, which are also inevitably close in space because of the permanent covalent peptide bonds. The importance of such local amino acid interactions was already pointed out decades ago based on information from folded protein structures [31,32]. The predictors to generate the input features from the primary sequence are based on information from NMR experiments, so incorporating protein mobility, and cover a wide range of proteins and their behavior; they therefore reflect rather what a protein is capable of in terms of general behavior, not its final fold. On the other hand, the NMR HDX experiments to pinpoint early folding residues indicate only the residues where the backbone NH hydrogen is protected from solvent by hydrogen bonding; this is therefore also what we are predicting. The present study seems to indicate that this hydrogen bond is typically made with the CO carbonyl of the residue at position −2, given the importance of helix propensity at that position, which implies 'turning' the backbone. The importance of sheet propensity for position 0 and 1, on the other hand, seems to indicate an extended conformation where especially the side-chain of the residue at position 1 can temporarily 'protect' the backbone from solvent, so aiding intramolecular hydrogen bond formation. This could still lead to helix formation; at this stage the protein is still highly dynamic, and a temporary formation of an extended conformation does not preclude later helix formation. In NMR experiments, for example, the first indication of local structure formation are NOEs between backbone atoms of residues at positions $i \rightarrow i+2$. This discussion is also only relevant for the environmental conditions in which the proteins were studied with HDX NMR, which are in the pH 3.0–8.0 range at temperatures between 273 and 303 K.

In evolutionary terms, the high frequency of predicted early folding residues in the scrambled dataset for the bb and SlGb predictors indicates that early folding fragments are easily formed randomly, although the original EFoldMine underpredicts them in scrambled sequences, which would rather indicate that early folding fragments are under some evolutionary pressure, and need to be maintained in particular positions to enable folding. The denovo dataset does not indicate a large different of predicted early folding residues, again for bb and SlGb, whilst the original EFoldMine indicates that early folding is (artificially) increased in the *de novo* designed sequences. This is the case even though the performances in themselves are similar (see Table 1), and highlights the importance of testing different approaches and the results they generate. The core rule sets, however, will be similar; in this

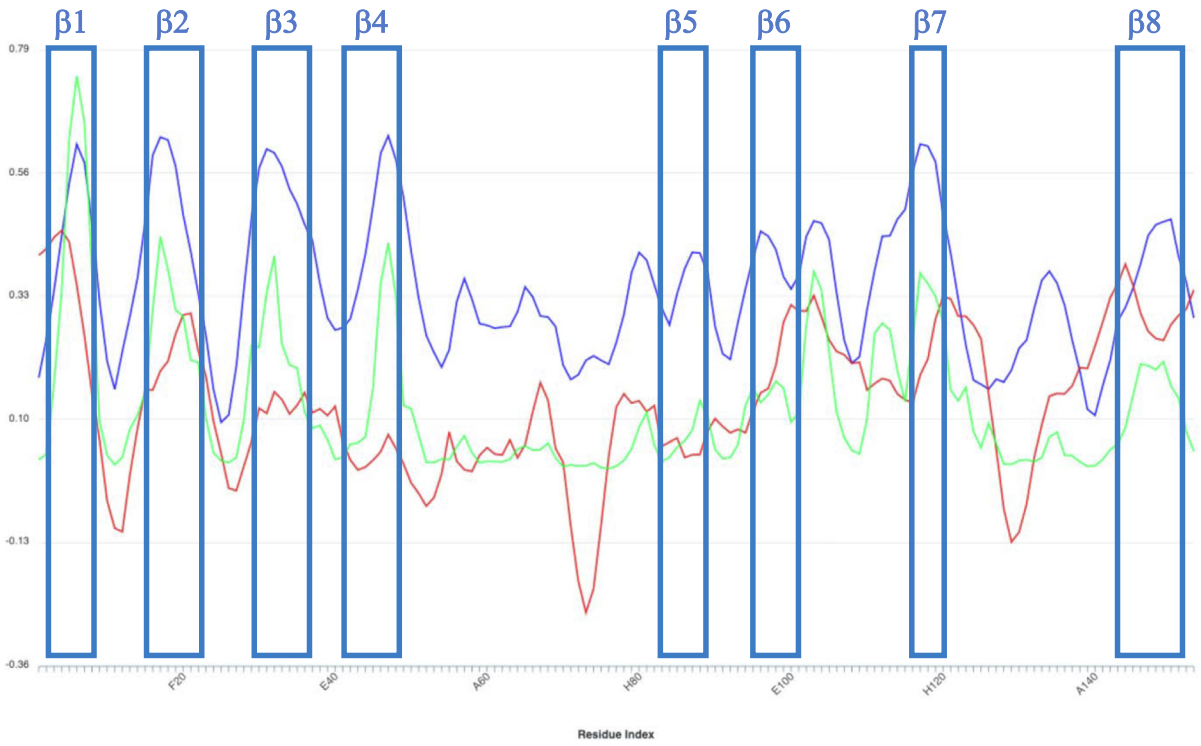

Fig. 8. The helix (red) and sheet (blue) propensities and early folding propensity (green) for superoxide dismutase 1. The positions of the eight beta strands, based on PDB code 3ECU, are indicated by blue boxes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

respect, it is also important to be aware of the fact that we here over-simplify the complex hyperdimensional space employed by the black box into a representation that highlights the essential features.

By providing all information online for easy examination, and by indicating likely core features that underlying the formation of early folding fragments, we hope to stimulate further discussion and especially guided experiments for further illumination. In addition, this resource is likely useful for protein design, to create local sequence fragments that ensure folding in particular sequence positions.

Funding

The Research Foundation Flanders (FWO) - project [grant number G.0328.16 N] to W.V.. I.G. was supported by the Flemish Government (AI Research Program) and the BRIGHTanalysis project, funded by the European Regional Development Fund (ERDF) and the Brussels-Capital Region.

CRediT authorship contribution statement

Isel Grau: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - original draft, Writing - review & editing, Visualization. **Ann Nowé:** Conceptualization, Methodology, Supervision, Funding acquisition, Project administration. **Wim Vranken:** Methodology, Software, Validation, Formal analysis, Data curation, Writing - original draft, Writing - review & editing, Visualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. Nucleic Acids Res 2000. https://doi.org/10.1093/nar/28.1.235.
[2] Bittrich S, Kaden M, Leberecht C, Kaiser F, Villmann T, Labudde D. Application of an interpretable classification model on Early Folding Residues during protein folding. BioData Min 2019;12:1–16. https://doi.org/10.1186/s13040-018-0188-2.
[3] Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. Funnels, pathways, and the energy landscape of protein folding: a synthesis. Proteins Struct Funct Bioinforma 1995;21(3):167–95. https://doi.org/10.1002/prot.340210302.
[4] Bryngelson JD, Wolynes PG. Spin glasses and the statistical mechanics of protein folding. Proc Natl Acad Sci USA 1987;84(21):7524–8. https://doi.org/10.1073/pnas.84.21.7524.
[5] Burley, S.K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G. V., et al., 2021. RCSB Protein Data Bank: Powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. Nucleic Acids Res. 49, D437–D451. doi:10.1093/nar/gkaa1038.
[6] Cilia E, Pancsa R, Tompa P, Lenaerts T, Vranken WF. The DynaMine webserver: predicting protein dynamics from sequence. Nucleic Acids Res 2014;42:W264–70. https://doi.org/10.1093/nar/gku270.
[7] Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20(1):37–46.
[8] Cohen WW. Fast Effective Rule Induction, in: Prieditis, A., Russell, S. (Eds.), Machine Learning Proceedings 1995. Elsevier, San Francisco (CA), pp. 115–123, 1995. doi:10.1016/b978-1-55860-377-6.50023-2
[9] Contessoto VG, Lima DT, Oliveira RJ, Bruni AT, Chahine J, Leite VBP. Analyzing the effect of homogeneous frustration in protein folding. Proteins Struct Funct Bioinforma 2013;81(10):1727–37. https://doi.org/10.1002/prot.24309.
[10] Daggett V, Fersht AR. Is there a unifying mechanism for protein folding? Trends Biochem. Sci. , 2003 doi:10.1016/S0968-0004(02)00012-9.
[11] Dill KA, Chan HS. From Levinthal to pathways to funnels. Nat Struct Biol 1997;4(1):10–9. https://doi.org/10.1038/nsb0197-10.
[12] Dobson CM. Protein folding and misfolding. Nature 2003;426(6968):884–90. https://doi.org/10.1038/nature02261.
[13] Englander SW, Mayne L. The nature of protein folding pathways. Proc. Natl. Acad. Sci. U. S. A. , 2014 doi:10.1073/pnas.1411798111.
[14] Ferreiro DU, Komives EA, Wolynes PG. Frustration, function and folding. Curr Opin Struct Biol 2018. https://doi.org/10.1016/j.sbi.2017.09.006.
[15] Fletcher S, Isla MZ. Comparing sets of patterns with the Jaccard index. Australas J Inf Syst 2018;22. https://doi.org/10.3127/ajis.v22i0.1538.
[16] Frank E, Witten IH. Generating Accurate Rule Sets Without Global Optimization, In Proceedings of the Fifteenth International Conference on

Machine Learning, ICML '98. University of Waikato, Department of Computer Science, San Francisco, CA, USA; 1998, pp. 144–151. 1-55860-556-8.
[17] Frauenfelder H, Sligar S, Wolynes P. The energy landscapes and motions of proteins. Science (80-.) 1991;254(5038):1598–603. https://doi.org/10.1126/science.1749933.
[18] Grau I, Sengupta D, Garcia Lorenzo MM, Nowé A. An Interpretable Semi-supervised Classifier using Rough Sets for Amended Self-labeling. IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). IEEE, 2020.
[19] Grau I, Sengupta D, Garcia Lorenzo MM, Nowé A. Interpretable self-labeling semi-supervised classifier. IJCAI/ECAI 2018 Workshop on Explainable Artificial Intelligence (XAI), 2018.
[20] Hu W, Walters BT, Kan Z-Y, Mayne L, Rosen LE, Marqusee S, et al. Stepwise protein folding at near amino acid resolution by hydrogen exchange and mass spectrometry. Proc Natl Acad Sci USA 2013;110(19):7684–9. https://doi.org/10.1073/pnas.1305887110.
[21] Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22(12):2577–637. https://doi.org/10.1002/bip.360221211.
[22] Kiefhaber T, Bachmann A, Jensen KS. Dynamics and mechanisms of coupled protein folding and binding reactions. Curr Opin Struct Biol, 2012 doi:10.1016/j.sbi.2011.09.010.
[23] Leopold PE, Montal M, Onuchic JN. Protein folding funnels: a kinetic approach to the sequence-structure relationship. Proc Natl Acad Sci USA 1992;89(18):8721–5. https://doi.org/10.1073/pnas.89.18.8721.
[24] Li R, Woodward C. The hydrogen exchange core and protein folding. Protein Sci 1999;8(8):1571–90. https://doi.org/10.1110/ps.8.8.1571.
[25] Molnar C. Interpretable Machine Learning. Leanpub 2019.
[26] Onuchic JN, Luthey-Schulten Z, Wolynes PG. Theory of protein folding: the energy landscape perspective. Annu Rev Phys Chem 1997;48(1):545–600. https://doi.org/10.1146/annurev.physchem.48.1.545.
[27] Pancsa R, Raimondi D, Cilia E, Vranken WF. Early folding events, local interactions, and conservation of protein backbone rigidity. Biophys J 2016;110(3):572–83. https://doi.org/10.1016/j.bpj.2015.12.028.
[28] Pancsa R, Varadi M, Tompa P, Vranken WF. Start2Fold: a database of hydrogen/deuterium exchange data on protein folding and stability. Nucleic Acids Res 2016;44(D1):D429–34. https://doi.org/10.1093/nar/gkv1185.
[29] Quinlan JR. C4.5: Programs for Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1993.
[30] Raimondi D, Orlando G, Pancsa R, Khan T, Vranken WF. Exploring the sequence-based prediction of folding initiation sites in proteins. Sci Rep 2017;7:1–11. https://doi.org/10.1038/s41598-017-08366-3.
[31] Rooman MJ, Kocher JPA, Wodak SJ. Extracting information on folding from the amino acid sequence: accurate predictions for protein regions with preferred conformation in the absence of tertiary interactions. Biochemistry 1992;31(42):10226–38. https://doi.org/10.1021/bi00157a009.
[32] Rooman MJ, Wodak SJ. Extracting information on folding from the amino acid sequence: consensus regions with preferred conformation in homologous proteins. Biochemistry 1992;31(42):10239–49. https://doi.org/10.1021/bi00157a010.
[33] Saibil H. Chaperone machines for protein folding, unfolding and disaggregation. Nat. Rev. Mol. Cell Biol., 2013 doi:10.1038/nrm3658.
[34] Sen Mojumdar S, N. Scholl Z, Dee DR, Rouleau L, Anand U, Garen C, et al. Partially native intermediates mediate misfolding of SOD1 in single-molecule folding trajectories. Nat Commun 2017;8(1). https://doi.org/10.1038/s41467-017-01996-1.
[35] Van Der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, et al.. Classification of intrinsically disordered regions and proteins. Chem. Rev, 2014 doi:10.1021/cr400525m.
[36] Velankar S, van Ginkel G, Alhroub Y, Battle GM, Berrisford JM, Conroy MJ, et al. PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. Nucleic Acids Res 2016;44(D1):D385–95. https://doi.org/10.1093/nar/gkv1047.
[37] Wang G, Dunbrack RL. PISCES: recent improvements to a PDB sequence culling server. Nucleic Acids Res 2005;33(Web Server):W94–8. https://doi.org/10.1093/nar/gki402.