XRRpred: Accurate Predictor of Crystal Structure Quality from Protein Sequence

Sina Ghadermarzi[1], Bartosz Krawczyk[1], Jiangning Song[2,3], Lukasz Kurgan[1*]

[1]Department of Computer Science, Virginia Commonwealth University, Richmond, VA 23284, USA; [2]Monash Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, VIC 3800, Australia; [3]Monash Centre for Data Science, Faculty of Information Technology, Monash University, Melbourne, VIC 3800, Australia

*To whom correspondence should be addressed.

Abstract

Motivation: X-ray crystallography was used to produce nearly 90% of protein structures. These efforts were supported by numerous sequence-based tools that accurately predict crystallizable proteins. However, protein structures vary widely in their quality, typically measured with resolution and R-free. This impacts the ability to use these structures for some applications including rational drug design and molecular docking and motivates development of methods that accurately predict structure quality.

Results: We introduce XRRpred, the first predictor of the resolution and R-free values from protein sequences. XRRpred relies on original sequence profiles, hand-crafted features, empirically selected and parametrized regressors, and modern resampling techniques. Using an independent test dataset, we show that XRRpred provides accurate predictions of resolution and R-free. We demonstrate that XRRpred's predictions correctly model relationship between the resolution and R-free and reproduce structure quality relations between structural classes of proteins. We also show that XRRpred significantly outperforms indirect alternative ways to predict the structure quality that include predictors of crystallization propensity and an alignment-based approach. XRRpred is available as a convenient webserver that allows batch predictions and offers informative visualization of the results.

Availability: http://biomine.cs.vcu.edu/servers/XRRPred/.

Contact: lkurgan@vcu.edu

1 Introduction

Knowledge of protein structures is invaluable to decipher protein functions (Kim, et al., 2003; Zhang and Kim, 2003) and to address practical applications, such as the rational drug design (Grey and Thompson, 2010; Jazayeri, et al., 2015; Maveyraud and Mourey, 2020). X-ray crystallography is the main approach to determine protein structures (Ilari and Savino, 2008; Lieberman, et al., 2013). As of Nov 2020, it accounts for over 88% of structures in the Protein Data Bank (PDB) and 80% of structures that were added to PDB in 2020 (wwPDB consortium, 2019). One of the key challenges of the X-ray crystallography is its low success rate. Studies have shown that only between 2% and 10% of crystallization trials lead to protein structures i.e., diffraction-quality crystals (Jahandideh, et al., 2014; Kurgan and Mizianty, 2009; Terwilliger, et al., 2009; Zimmerman, et al., 2014). Moreover, the resulting structures offer varying levels of quality, which impacts the ability to use them for some of the applications. According to a recent survey, average X-ray crystallization costs are at about $150,000 per protein based on the reported $2 billion funding to produce 13,500 structures (Grabowski, et al., 2016). Inability to successfully crystallize protein targets accounts for over 60% of these relatively high protein structure production costs (Jahandideh, et al., 2014; Slabinski, et al., 2007). Consequently, target selection approaches that aim to identify proteins which are both relevant and likely to be solvable by X-ray crystallography were developed (Chandonia, et al., 2006; Marsden and Orengo, 2008; Robin, et al., 2008). Effective target selection requires computational tools that accurately predict whether diffraction-quality crystal structure can be produced from a given protein sequence (Gao, et al., 2018; Grabowski, et al., 2016; Rupp and Wang, 2004; Wang, et al., 2018). Recent surveys reveal that crystallization predictors have improved over time to the point where they now provide very accurate results (Gao, et al., 2018; Wang, et al., 2018). For instance, recently released DeepCrystal, BCrystal and DCFCrystal methods provide predictions with AUC > 0.91 (Elbasir, et al., 2020; Elbasir, et al., 2019; Zhu, et al., 2020).

The crystal structures vary widely in terms of their quality, ranging from very low quality that delineates only an overall protein shape to high quality that provides accurate and precise positions of all atoms. The structure quality is typically described by resolution and R-values. Resolution can be understood as the smallest distance between atoms in the crystal which are seen as separate in the inferred structure (Dubach and Guskov, 2020). The resolutions of PDB structures range between ~0.5Å and several Å, with the lower values corresponding to better quality. Atomic resolution structures, which allow distinguishing individual atoms with little error in their placements, have resolutions <1.2Å (Morris and Bricogne, 2003). Near-atomic structures for which backbone atoms can be located with high confidence but details of side-chains and their orientation can be inaccurate are defined by resolutions ranging between 1.2Å and 2Å. The medium, low and very low resolution structures are defined by the 2Å to 3Å, 3Å to 5 Å, and >5Å intervals, respectively (Dubach and Guskov, 2020). The R-values measure the degree of match between simulated models and experimentally observed diffraction patterns. The typically used R-free values are computed from experimental data that is excluded from the calculation of the structure, reducing intrinsic bias associated with the R-values that are modelled and computed on the same data (Brunger, 1992; Kleywegt and Jones, 1997). These out-of-sample values are considered as the most useful measures of the model-to-measured data agreement (Read, et al., 2011). The R-free values of PDB structures vary between 0.045 and 0.512, where lower values denote better matching. Although resolution and R-free are

correlated, they convey complementary information about the structure. For instance, PDB structure of endothiapepsin (PDB ID: 5RDH) has resolution = 0.85 coupled with R-free = 0.334, suggesting that this atomic resolution structure is in relatively poor agreement with the corresponding structural model. Importantly, certain applications require specific levels of structure quality. For instance, rational drug design (Fernández-Ballester, et al., 2011; Grey and Thompson, 2010; Jazayeri, et al., 2015) and molecular docking applied to predict protein-protein interactions (Movshovitz-Attias, et al., 2010; Park, et al., 2015) rely on the atomic resolution structures.

While some applications require specific levels of structure quality, currently there are no tools that are able to directly predict whether a given protein sequence would produce structure with the desired quality. A couple indirect attempts were recently made to adapt results produced by the crystallization predictors to quantify the structure quality; however, they show low levels of predictive performance (Gao, et al., 2018; Wang, et al., 2018). In other words, while the available tools accurately find whether a given sequence will produce diffraction-quality crystals, the resulting structure quality of the crystallizable proteins cannot be accurately predicted. The quality of the X-ray crystallography-solved structures depends on many factors including protocols used at different stages of structure determination pipelines and the intrinsic properties of the protein itself. While there are many crystallization protocols, they are quite similar and typically utilize a common set of suggested strategies (Graslund, et al., 2008). We argue that intrinsic properties of the protein sequence provide useful information that at least partially determines the resulting structure quality. We demonstrate this based on a relation between one of the most basic sequence characteristics, its length, and quality of the corresponding structures using a large dataset of 128,017 PDB structures defined in Section 2.1 (**Suppl. Fig. S1**). We observe a gradual decrease in structure quality (higher resolution and R-free values) as the chain length grows (denoted by darker shades). This relation is true for R-free (horizontal bar along x-axis in **Suppl. Fig. S1**), resolution (vertical bar along the y-axis) and when considering both quality values (area inside **Suppl. Fig. S1**), in spite of the fact that these structures were solved by different groups/centers using different crystallization protocols. The feasibility of using intrinsic sequence characteristics to predict the structure quality is further supported by the success of the current crystallization predictors. These methods also rely solely on the sequence-derived characteristics to accurately determine whether these sequences could produce diffraction-quality crystals (Canaves, et al., 2004; Elbasir, et al., 2020; Elbasir, et al., 2019; Kurgan, et al., 2009; Mizianty and Kurgan, 2011; Wang, et al., 2016; Zhu, et al., 2020).

Given the availability of the very accurate methods that identify crystallizable sequences, we focus on solving the subsequent, equally challenging and important problem of predicting quality of the resulting structures. We introduce a new predictive tool, XRRpred (**X**-ray crystallography **R**esolution and **R**-free **pred**ictor), that:

- predicts R-free and resolution values directly from protein sequence;
- facilitates prediction for multi-chain proteins that are common in PDB;
- relies on modern resampling and regression algorithms to offer accurate predictions that significantly outperform alternative approaches that can be used to indirectly predict structure quality;
- is freely available as a webserver that conveniently performs computations on the server side.

2 Methods

2.1 Dataset

We collect a large set of proteins with known structure quality to train, optimize and compare the predictive model (**Table 1**). Using PDB, we extract 128,017 X-ray structures of proteins that have experimental values of resolution and R-free and which exclude structures where proteins are in complex with RNA, DNA and RNA/DNA hybrids. The exclusion of these complexes is driven by the fact that we rely on the protein sequence as the sole input and since the presence of these large ligands would inevitably affect the structure quality. Preliminary analysis of this datasets reveals that the size of the protein structures, which we quantify with the total length of their chains, has grown over the years (**Suppl. Fig. S2A**). This suggests that progressively larger protein structures are being solved. Interestingly, the ratio of resolution to the structure size (**Suppl. Fig. S2B**) and R-free to the structure size (**Suppl. Fig. S2C**) are relatively similar for a significant majority of these structures, i.e., for structures deposited after 2000. We hypothesize that this could be explained by the increasing size of the solved structures. Thus, we do not limit the data to a specific timeframe to maximize the dataset size. Next, we utilize the PDB facilities to cluster these structures at the 30% sequence similarity and select one structure per cluster to evenly sample the sequence space. The corresponding query is: "*Resolution is between 0.0 and 100.0 and XrayRefinementQuery: refine.ls_R_factor_obs.comparator=between refine.ls_R_factor_all.comparator=between refine.ls_R_factor_R_work.comparator=between refine.ls_R_factor_R_free.comparator=between refine.ls_R_factor_R_free.min=0 refine.ls_R_factor_R_free.max=100 and Chain Type: there is a Protein chain but not any DNA or RNA or Hybrid and Representative Structures at 30% Sequence Identity*". Finally, we remove peptides (chains with length <20 residues) and sequences with non-standard amino acids among the returned results. The latter was motivated by the fact that tools that we use to produce predictive features from the input sequences and to implement alternative ways to generate the predictions (which we compare to) could not produce results for such sequences.

We divide the remaining proteins into the training and test datasets. We assign 2,037 proteins deposited to PDB after January 1, 2018 into the test dataset; the 18,305 older depositions comprise the training dataset. This simulates a scenario where the information about older structures is used to build a model that predicts the quality of newer structures. These datasets are summarized in **Table 1**. The key relevant characteristics, such as the R-free and resolution values, the sequence length and the number of chains per protein are similar between the training and test datasets, suggesting that our model should be similarly applicable for future structures. The datasets are available at http://biomine.cs.vcu.edu/servers/XRRPred/.

We use cross-validation of the training set to conceptualize, design, and optimize the predictive model. Subsequently, we comparatively test the optimized model on the set-aside (excluded from training) test dataset. Given the above clustering, proteins in the test dataset share low (<30%) sequence similarity to proteins in the training dataset. This ensures that they could not be accurately predicted using alignment/sequence similarity and is also in line with existing studies that develop crystallization predictors where the similarity is limited to a range between 25% and 40% (Elbasir, et al., 2020; Elbasir, et al., 2019; Meng, et al., 2017; Wang, et al., 2018; Zhu, et al., 2020).

Table 1. Summary of the training and test datasets. We report mean ±stdev values.

Dataset characteristics	Training dataset	Test dataset
Resolution	1.96 ±0.52	1.95 ±0.57
R-free	0.226 ±0.039	0.220 ±0.039
Average chain length per protein	273 ±183	312 ±208
Number of chains per protein	2 ±2	2 ±2
Number of proteins	18,305	2,037

2.2 Evaluation metrics

We assess the quality of predictions of the real-valued resolution and R-free generated by various models, including XRRpred, by comparing their outputs with the experimentally measured resolution and R-free values. We use common metrics for this evaluation including Mean Absolute Error (MAE), Mean Squared Error (MSE), Pearson Correlation Coefficient (PCC), and Spearman Correlation Coefficient (SCC). We define these metrics in the Supplementary file.

2.3 Resampling

Our preliminary attempts to design and train an accurate predictor have revealed a significant problem. While our predictions were correlated with the native values (e.g., we secured SCC = 0.38 for resolution), they were impractical since they covered a much narrower range than the native values. For instance, the standard deviation of the predictions was at around 0.2 while the native resolution has the standard deviation at about 0.5. The underlying reason is the imbalance in the distribution of resolution and R-free values across their range, with a larger number of objects (proteins) having close to average resolution and R-free values and much fewer proteins having large and small values (i.e. these are long tailed distributions). Correspondingly, predictive models focus on the center of the distributions while disregarding the tails, leading to small standard deviations (Krawczyk, 2016). We mitigate this problem by balancing the training distributions using resampling. Note that we only resample the training data, and use the original/non-resampled test data.

While there are many methods to resample imbalanced data for classification (prediction of a categorical output variable), the literature concerning resampling for regression problems (prediction of real-valued output variables, which in our case are resolution and R-free) is relatively scarce (Krawczyk, 2016). We consider a comprehensive set of six resampling methods for regression. Our selection covers two main types of resampling: under-sampling (random undersampling RU (Branco, et al., 2019)) and over-sampling (SMOTE (Torgo, et al., 2013) and RBOR (Krawczyk, et al., 2020)). Under-sampling balances the distribution by removing a subset of objects that have over-represented output values. Over-sampling methods introduce new synthetic objects with output values that are under-represented in the dataset (which in our case are located in the distribution tails). We also include variations of these three methods that are augmented with a noise reduction step: RU-ENN for the under-sampling (Fernández, et al., 2018), and SMOTE-ENN (Branco and Torgo, 2019) and RBOR-C (Koziarski, et al., 2020) for the over-sampling. The noise reduction aims to minimize situations where similar objects that have very different output values are close to each other. These objects can be disruptive to the resampling procedure and may negatively affect the subsequent model training process. We briefly describe the six methods that we utilize (i.e., RU, RU-ENN, SMOTE, SMOTE-ENN, RBOR and RBOR-C) in the Supplementary file. We perform resampling separately for the prediction of resolution and R-free.

2.4 Predictive model

XRRpred predicts resolution and R-free directly from the protein sequence in three steps (**Fig. 1**): 1) extraction of residue-level profiles (done for each chain from the input protein); 2) extraction of protein-level features from the profiles (done over the chain-level profiles); and 3) prediction of resolution and R-free from the protein-level features using two dedicated regression models. The result (output) are the predicted real-valued resolution and R-free for the whole protein structure. We detail the three steps in the subsequent subsections.

2.4.1 Extraction of the residue-level profiles

We extract the profiles directly from the input protein sequence(s). They cover residue-level information that is relevant to the prediction of structure quality. We calculate profiles for each chain separately. The profiles include intrinsic disorder predicted with IUPred (Dosztányi, et al., 2005), solvent accessibility predicted with ASAquick (Faraggi, et al., 2014), and a selection of pertinent physicochemical and structural properties of amino acids, such as side chain characteristics, polarity, charge, size, hydrophobicity, flexibility, propensity for structured/disordered conformations and propensity for folding. We selected IUPred and ASAquick motivated by their speed and reasonably strong levels of predictive quality (Faraggi, et al., 2014; Walsh, et al., 2015). We note that the use of slower and more accurate predictors could lead to improvements in the predictions of structure quality as a trade-off for longer runtime. We consider intrinsic disorder because the presence of disorder is shown to negatively impact production of structures via X-ray crystallography (Hu, et al., 2018; Oldfield, et al., 2013). We particularly focus on surface residues extracted using the putative solvent accessibility since they drive packing of proteins into crystals that affects structure quality (Seeliger and de Groot, 2007). Inclusion of this sequence-derived information improves the predictive quality of the protein-level features that we extract in the next step when compared to using solely the sequence. We provide further details in the Supplementary file.

2.4.2 Extraction of the protein-level features

We cannot directly use the sequences or the residue-level profiles as an input to the predictive models. This is because they have variable sizes that depend on the chain length and number of chains per protein while predictors require a fixed-length input. Accordingly, we extract a set of handcrafted protein-level features from the profiles. The design of these features aims to capture key characteristics of the profiles that are related to the prediction of the structure quality. We also extract features directly from the chain sequences. The extraction of features involves two steps. First, we compute a given feature at the chain-level. This step quantifies physicochemical and structural properties of a given chain, including disorder, focusing on putative surface residues. Second, we use the chain-level values to calculate the protein-level features. More specifically, we use the minimum, maximum, and average operators to aggregate the chain-level features for the same protein. Suppl. Table S2 details the calculation of these features. We extract a total of 324 features.

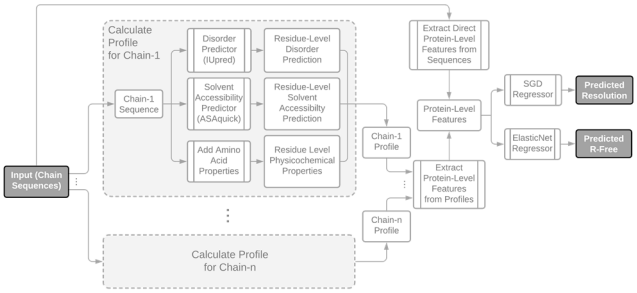

Fig. 1. Block diagram of XRRpred's prediction process. The dark gray boxes denote inputs and outputs. The white boxes with vertical bars mark calculations and white boxes without vertical bars show intermediate data. The light grey boxes delineate the profiles computed for the protein chains.

2.4.3 Prediction of resolution and R-free

We use machine learning models to predict the resolution and R-free from the feature-based representation of the input sequences. We train and optimize the models exclusively on the training dataset. The optimization considers three main aspects including selection and parametrization of machine learning algorithms, selection of relevant features, and selection of resampling methods. Therefore, we perform 3-dimensional grid search to select models that produce the highest SCC based on the 3-fold cross-validation on the training set. Next, we detail each of the three dimensions.

We consider a comprehensive selection of seven regressors (algorithms that produce real-valued predictions) listed in Suppl. Table S3. We exclude deep network-based models since they require a much larger amount of the training data and would likely overfit our training dataset. This table lists hyperparameters and their values that we consider in the grid search. The parametrization involves two passes. In the first pass, we use the hyperparameter values shown in Suppl. Table S3 that uniformly sample the parameter space (exploration step). In the second pass, we fine-tune values around the values selected in the first pass (refinement step).

Some of the considered 324 features could be mutually correlated and/or provide low-quality input for the prediction of resolution or R-free. While some of the considered regressors have an intrinsic ability to identify and utilize a suitable subset of the input features during training, other methods may suffer reduced predictive performance when correlated and poor-quality features are used. Therefore, we consider the following two scenarios: 1) the entire feature set is used; and 2) a subset of empirically selected features is used. The selection relies on a combination of the filter and wrapper feature selections. In the filter-based step, the features are sorted based on

their SCC with the outcome (resolution or R-free) on the training dataset and we remove features with the low SCC values; this eliminates low-quality features. In the subsequent wrapper-based step, we identify a subset of remaining features based on forward selection using the sorted list of features and linear regression in the cross-validation on the training set. Only the features that improve predictive performance are selected. This step eliminates mutually correlated features. This selection protocol has been used in several related studies (Hu, et al., 2019; Meng and Kurgan, 2018; Yan and Kurgan, 2017; Zhang and Kurgan, 2019).

As discussed in Section 2.3, we perform resampling to ensure that the predictions cover the full range of the experimental resolution and R-free values. Thus, the grid search covers the use of the original unsampled dataset as well as the use of each of the considered six resampling methods.

2.4.4 Optimization of the predictive models

We perform the grid search-based optimization using the cross-validation on the training dataset separately for the prediction of resolution and R-free. We examine 4,172 setups (298 combinations of machine learning algorithms and hyper-parameters, 2 feature sets: complete and selected, and 7 resampling options). We search for a setup that results in the highest SCC value while providing predictions with a correct range of resolution and R-free values (detailed in Section 2.3). The results are summarized in Suppl. Table S4. We compare the best results for each resampling method, including also the results where original/non-resampled data is used, since inclusion of resampling substantially impacts the predictive performance.

Using the training data, the best configuration for the prediction of resolution utilizes the Stochastic Gradient Descent (SGD) regressor (Bach and Moulines, 2011), all features, and the SMOTE resampling. For R-free, we secure the best results with the Elastic Net regression (Zou and Hastie, 2005), using the empirically selected features and the SMOTE-ENN resampling. The selected hyperparameters for these two models are listed in Suppl. Table S4. Although we use these specific configurations to implement XRRpred, both types of regressors produce relatively similar results. While the impact of the selection of the regressors is limited, we find that the use of resampling leads to substantial improvements. For both resolution and R-free, we secure the best results using the SMOTE-based resampling that oversamples the "rare" proteins (i.e. by introducing synthetic samples with either low or high values of resolution/R-free) in the training dataset. The use of the undersampling techniques (RU and RU-ENN) provides lower quality results compared to the SMOTE-based oversampling. This can be explained by the fact that undersampling removes the "over-represented" proteins (i.e., proteins with the common/mid-range values of resolution and R-free), therefore reducing the size of the training dataset. Moreover, the application of the original/non-resampled training dataset leads to low predictive performance, when we set the underlying model to produce the predictions that follow a similar range of values as the native resolution/R-free (see **Table 1**). This was our original motivation to introduce resampling. The results also show that feature selection is unnecessary for some regressors, such as SGD and linear regression, which are capable of selecting features indirectly through the optimization of the coefficients. On the other hand, other regressors, such as the Elastic Net, decision tree regressor (Breiman, 1984) and passive aggressive regressor (Crammer, et al., 2006), benefit from the feature selection (Suppl. Table S4). Finally, Suppl. Table S4 reveals the best-performing configurations on the training set also produce the best result on the test dataset. We emphasize that we evaluated these models on the test dataset only after the we finalized the selection of the configurations on the training dataset. This suggests that the use of the cross-validation has been indeed effective to optimize the predictive models.

3 Results

3.1 Comparative assessment

We compare the results produced by XRRpred on the test dataset to a representative set of alternative solutions. To the best of our knowledge, there are currently no other predictors of structure quality quantified with resolution and R-free. Thus, we define several indirect ways to make these predictions including a random predictor, a sequence alignment-based approach and an application of current predictors of the X-ray crystallization propensity. The latter is motivated by recent studies that attempt to use these predictions as a proxy for structure quality (Gao, et al., 2018; Wang, et al., 2018). We selected two recent and runtime-efficient (to process the entire test dataset) crystallization propensity predictors: fDETECT (Meng, et al., 2017; Mizianty, et al., 2014) and DeepCrystal (Elbasir, et al., 2019). They produce a numeric propensity score, for which higher value denotes higher likelihood to produce diffraction-quality crystals. This score is inversely correlated with the resolution and R-free. Therefore, we convert these predictions using the minmax normalization into the range of resolution and R-free values in the training dataset, such that the minimal (maximal) crystallization propensity is mapped to the highest (lowest) resolution/R-free value. Since resolution and R-free are different measures of structure quality we define two alternative indirect predictors based on these two scores. ALT1 uses the converted DeepCrystal scores as putative resolution and the converted fDETECT scores as putative R-free, while ALT2 utilizes the converted fDETECT scores as predicted resolution and the converted DeepCrystal scores as predicted R-free. The alignment-based predictor relies on the premise that proteins with similar sequences should share similar structure quality. To this end, we utilize BLAST (Altschul, et al., 1997) to compute the similarity between a given test protein and each protein in the training dataset and use the resolution and R-free from the most similar training protein as the prediction. Finally, the random predictor produces a random number within the range of the experimental resolution and R-free values. This provides a baseline that corresponds to the lowest possible predictive performance. We use statistical significance tests to assess whether the predictive performance of XRRpred is significantly different from the results provided by each of these reference predictors. This test essentially evaluates whether the differences would hold across different test datasets. Thus, we compare results across 10 disjoint and equally-sized subsets of the test proteins. We use the *t*-test if the measured metrics are normal (we assess normality with the Anderson-Darling test at *p*-value of 0.05) and the Wilcoxon rank-sum test otherwise. **Table 2** compares the correlations (PCC and SCC), errors (MAE and MSE), and the distributions of the predicted scores (represented by the average and standard deviation of resolution/R-free) on the test dataset for XRRpred, ALT1 and ALT2 predictors, the alignment-based approach, the random baseline and the regressors that we optimized in Section 2.4.4 without sampling.

Table 2. Comparison of the predictive performance on the test dataset. The predictors are sorted in the descending order based on the SCC scores, separately for the prediction of resolution and R-free. The stars denote that the difference for a given metric between XRRpred and a given other predictor is statistically significant (*p*-value<0.05). Bold font highlights the best results. "Mean ±Std" denote the mean and standard deviation of the predicted values.

	Method	SCC	PCC	MAE	MSE	Mean±Std
Resolution	**XRRpred**	**0.43**	**0.46**	**0.44**	**0.33**	**1.97 ±0.54**
	Best regressor without resampling	0.22*	0.25*	0.60*	0.62*	2.00 ±0.70
	ALT1	0.22*	0.25*	1.39*	3.07*	3.05 ±1.39
	ALT2	0.16*	0.16*	2.55*	7.61*	4.47 ±1.07
	Random	0.04*	0.04*	1.88*	5.30*	3.50 ±1.63
	Similarity-based	-0.02*	-0.04*	0.60*	0.62*	1.96 ±0.52
R-free	**XRRpred**	**0.35**	**0.36**	**0.034**	**0.002**	**0.231 ±0.033**
	Best regressor without resampling	0.06*	0.05*	0.055*	0.006*	0.220 ±0.068
	ALT2	0.09*	0.10*	0.066*	0.007*	0.235 ±0.073
	ALT1	0.01*	-0.02*	0.100*	0.013*	0.310 ±0.056
	Similarity-based	-0.01*	-0.02*	0.045*	0.003*	0.226 ±0.040
	Random	-0.03*	-0.02*	0.087*	0.011*	0.259 ±0.087

XRRpred provides accurate predictions of resolution and R-free. Its predictions secure SCC = 0.43 and MAE = 0.44 for the prediction of resolution. This means that on average these putative resolutions are 0.44Å away from the actual values. This is a relatively small error considering that the range of the resolution values is about 7Å. Similarly, XRRpred obtains SCC = 0.35 and MAE = 0.034 for the prediction of R-free, when the range of R-free values is about 0.47. The moderate values of correlations stem from the use of the protein sequence as the sole input, which indirectly limits the highest achievable predictive performance. Importantly, these predictions cover the correct range of the resolution and R-free values. To compare, the average and standard deviation of the XRRpred's predictions of resolution are 1.97±0.54 (**Table 2**), while for the experimental resolutions they are 1.95 ±0.57 (**Table 1**). Similarly, for R-free they are 0.231±0.033 (Table 2) for XRRpred and 0.220 ±0.039 (**Table 1**) for the experimental data.

XRRpred strongly benefits from the SMOTE-based resampling. Compared to the regressors that exclude resampling (Table 2), XRRpred provides statistically significant improvements for all metrics (p-value < 0.05). For the prediction of resolution SCC values drop from 0.43 to 0.22 and MAE increases from 0.44 to 0.60. Similarly, for the R-free SCC drops from 0.35 to 0.06 and MAE worsens from 0.034 to 0.055.

Compared to the crystallization propensity predictors (ALT1 and ALT2 methods), XRRpred also secures statistically significant improvements across all metrics for the prediction of resolution and R-free (p-values < 0.05). For instance, for resolution the SCC of the better ALT1 combination is 0.22 and for R-free the SCC of the better ALT2 combination is 0.09. This agrees with the results from recent studies that shows similarly low correlations between the crystallization propensity and resolution (Gao, et al., 2018; Wang, et al., 2018). Interestingly, our results reveal that while the putative crystallization propensities are modestly correlated with resolution (SCC for DeepCrystal is 0.22 vs. 0.16 for fDETECT) they are not correlated with the R-free values (SCC for DeepCrystal is 0.09 vs. 0.01 for fDETECT). This also underscores the fact that resolution and R-free are two different measures of structure quality.

Both the alignment-based approach and the random baseline offer similarly poor performance with SCC and PCC at around zero and large errors. While this is expected for the random predictor, the poor performance of the alignment stems from the fact that test proteins to share low, <30%, similarity to the training proteins. Such low levels of similarity render alignment-based predictions ineffective. However, XRRpred still secures accurate predictions in spite of this low levels of similarity.

Finally, we study a potential impact of similarity between the proteins that were used to train IUPred and ASAquick methods, which we use to derive predictive inputs to XRRpred (**Fig. 1**), and the proteins from our test dataset. To accomplish that, we extract a subset of the test proteins that share low similarity with the IUPred and ASAquick training proteins. More specifically, we compute pairwise similarity of each chain from the test proteins to every training protein using BLAST and eliminate test proteins for which at least one chain had similarity ≥ 30%. Next, we retest XRRpred on the resulting set of 611 dissimilar test proteins. Our predictor secures MAE = 0.41 (vs. 0.44 using the complete test dataset), MSE: 0.27 (vs. 0.33), PCC = 0.48 (vs. 0.46) and SCC = 0.42 (vs. 0.43) for the prediction of resolution. Similarly, it obtains MAE = 0.031 (vs. 0.034), MSE = 0.001 (vs. 0.002), PCC = 0.32 (vs. 0.36) and SCC = 0.31 (vs. 0.35) for the prediction of R-free. We observe that the results on the complete test dataset and its subset that shares low similarity to the training data of IUPred and ASAquick are similar. The lack of sensitivity to the similarity to these training proteins could be explained by the simple design of the IUPred and ASAquick models which do not use sequence alignment, thus reducing possibility of over-fitting training datasets (Dosztányi, et al., 2005; Faraggi, et al., 2014).

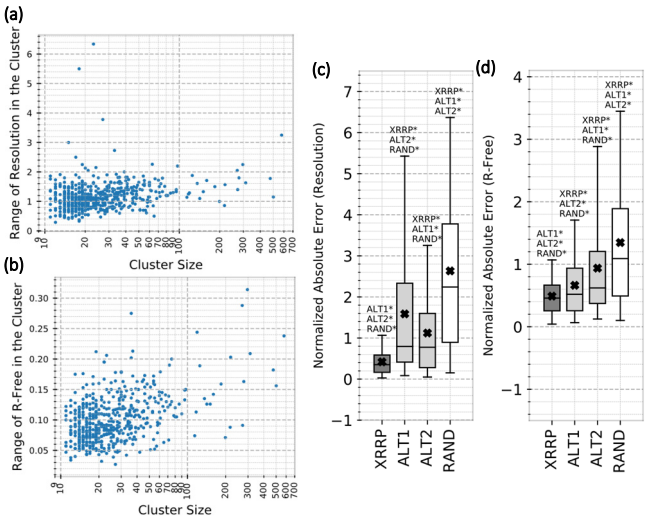

Fig. 2. Analysis of the experimental and putative structure quality values for clusters of structures of the same protein sequences. Panels a and b show the sizes of clusters and the corresponding range = max – min for the experimental values of resolution and R-free within clusters, respectively; each point is a cluster. Panels c and d show the distributions of the normalized absolute errors, which are defined as the absolute difference between predictions and the minimal experimental value in the cluster divided by the range of native values in the cluster. The whiskers are the 5th and 95th percentiles. The middle bar in the box is the median and the cross mark is the mean. The annotations above the whiskers name methods for which the absolute errors are significantly different (p-value<0.05). Statistical test is defined in Section 3.1.

3.2 XRRpred identifies favorable structure quality for proteins with multiple solved structures

PDB offers multiple structures for some of the proteins. This may stem from a variety of reasons including solving structures under different crystallization protocols, across different organisms, and/or in complex with different ligands. For example, there are well over 500 X-ray crystallography structures of the lysozyme. We study the clusters of structures of identical proteins (i.e., proteins in a given cluster have the same number of chains and the same sequences) to investigate whether these structures differ in quality. We focus on the clusters with at least 10 structures to ensure that we can compute reliable statistics. **Figures 2a** and **2b** show the range of resolution and R-free values against the corresponding cluster sizes. We find many clusters that include dozens and even hundreds of structures. Moreover, we observe that the quality of structures in these clusters varies. The majority of the clusters include proteins with structures for which resolutions differ by over 1Å and with R-frees that vary by over 0.08. These are substantial differences considering that the standard deviations of resolution and R-free values in PDB are at about 0.55 and 0.04, respectively.

Given these results, we investigate whether XRRpred and the other methods could produce predictions that fall within the range of the experimental outcomes and that can be used to identify favorable (close to the best) structure quality for these proteins. We measure the absolute error between their predictions and the corresponding cluster minimums. Since different clusters have different ranges of resolution and R-free values, we normalize the error using the range. This results in the following metric:

$$\text{Normalized Absolute Error} = \frac{|\text{predicted value - cluster minimum value}|}{\text{cluster max - cluster min}}$$

Figures 2c and **2d** show that XRRpred secures low normalized absolute errors for the prediction of resolution and R-free. Nearly all of its errors are below 1, which means that XRRpred's predictions do not exceed the range of values inside clusters. XRRpred also produces predictions that are closer to the cluster minimum than the cluster maximum for majority of clusters. This can be observed based on the median values shown in **Figures 2c** and **2d**. Furthermore, statistical analysis reveals that XRRpred's errors are statistically significantly lower (p-value < 0.05) than the errors produced by the other solutions including the usage of the crystallization propensity

predictors (ALT1 and ALT2) and the random baseline. This observation is consistent with the result in **Table 2**. We note that both crystallization propensity predictors are statistically significantly better than the random baseline (p-value < 0.05). Altogether, this analysis suggests that XRRpred outperforms the other approaches since its predictions are within the range of the experimental resolution/R-free and provide useful clues to identify favorable structure quality for these protein clusters.

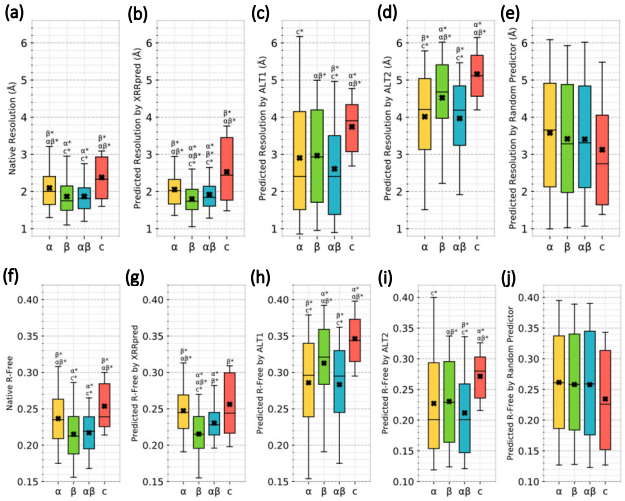

Fig. 3. Distribution of experimental and putative structure quality values for the test proteins grouped into the structural classes defined in the CATH resource. Panels a and f compare the distribution of native resolution and R-free values across the four structural classes. Panels b and g compare the distribution of putative resolution and R-free values predicted by XRRpred. Panels c, d, h and i focus on the putative values produced by the crystallization propensity predictors (ALT1 and ALT2 methods). Panels e and j summarize results for the random baseline prediction. The whiskers give the 5 and 95 percentiles. The middle bar in the box is the median and the cross mark is the mean. The annotations above the whiskers list structural classes for which the differences in resolution or R-free values from the class corresponding to this box plot are significant from (p-value<0.05). Statistical test is defined in Section 3.1.

3.3 XRRpred's predictions reproduce structure quality relations across structural classes of proteins

Protein structures are categorized into several classes that are typically defined based on the overall arrangement and composition of the secondary structure elements. One commonly used classification of protein structures into classes is introduced in the CATH database (Orengo, et al., 2002; Sillitoe, et al., 2021). The top-level of this classification hierarchy relies on a relative composition of secondary structures and covers four major classes: *mainly alpha* (structures composed primarily of alpha-helices; denoted as the α class); *mainly beta* (primarily beta-sheets; denoted as the β class); *alpha and beta* (structures composed of alpha helices and beta-sheets; denoted as the αβ class); and *few secondary structures* (structures mostly devoid of alpha helices and beta-sheets and composed primarily of coils; denoted as the c class). We categorize proteins from the test dataset using the CATH class assignment protocol (Michie, et al., 1996) with the secondary structures that we collect from PDB. Next, we investigate whether proteins from different structural classes differ in their structure quality measured with resolution and R-free (see **Figures 3a** and **3f**). Our analysis reveals that structures of proteins from different structural classes are characterized by statistically significantly different resolution and R-free values. For instance, structures in the β class have on average the best resolution and R-free values, which are significantly better than the corresponding structure quality from the α and c classes (p-value<0.05). On the other end of the spectrum, structures in the c class have on average the worst quality that is significantly inferior to the quality of structures in the β and αβ classes (p-value<0.05). Interestingly, these relations are consistent across both measures of the structure quality.

Considering the above findings, we use the test dataset to study whether XRRpred and the other methods generate predictions of structure quality that reproduce these relationships across the structural classes. **Figures 3b** and **3g** show that XRRpred correctly recapitulates these relationships. It generates similar distributions of the resolution and R-free values for each class when compared to the experimental structure quality distributions (**Figures 3a** and **3f**). The putative structural quality output by XRRpred sorts the structural classes from the best quality β class, through αβ class, α class and finally to the worst quality c class. This is identical with the order of the structural classes based on the experimental data. Moreover, XRRpred reproduces nearly all significant differences that we observe using the experimental data, with the only exception of some of the results for the αβ class. These results are in plain contrast to the predictions from the crystallization propensity predictors (both ALT1 and ALT2 methods) and the random baseline (**Figures 3c, 3d, 3e, 3h, 3i** and **3j**). These approaches do not replicate the correct range of values for any of the classes, mix-up the order of classes (except for ALT1 and ALT2 that correctly identify that the c class obtains the worst structure quality values) and are unable to properly quantify the significance of these differences. To sum up, we demonstrate that XRRpred is the only currently available tool capable of reproducing the structure quality relations between the four structural classes of protein structures.

3.4 XRRpred's predictions reproduce relationship between resolution and R-free values

While resolution and R-free are correlated, they represent complementary information about the structure quality (Read, et al., 2011). This is why XRRpred covers both measures. **Fig. 4a** visualizes and quantifies the relation between these two structure quality measures on the test dataset. The correlation between the experimental resolution and R-free values is 0.75.

Using the test dataset, we examine whether XRRpred and the other indirect predictors of structure quality can accurately model the relationship between the two structure quality measures. **Fig. 4b** shows that XRRpred's predictions of resolution and R-free follow a similar pattern to the experimental data and share a virtually identical 0.75 correlation. The ALT1 and ALT2 methods that rely on the crystallization propensity predictors produces a rather different relation (**Figures 4c** and **4d**) and their outputs features correlations of 0.56. Finally, as expected, the random baseline (**Fig. 4e**) shows no discernable relations and no correlation. This analysis shows that XRRpred not only provides accurate predictions of resolution and R-free but also shows that its predictions preserve the relationship between these two structure quality measures. This also supports the design of XRRpred that relies on two separately trained regressors, which in spite of that are able to provide consistent results across the test proteins.

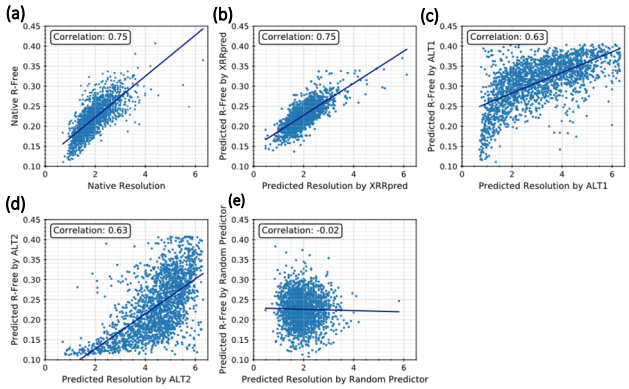

Fig. 4. Relationship between resolution and R-free values for the experimental and predicted data on the test dataset. Panel a shows the experimental data. Panels b, c, d and e show the predictions. The dark blue line is the linear regression line. The

corresponding PCC values between resolution and R-free values are given in the top left corner of each panel.

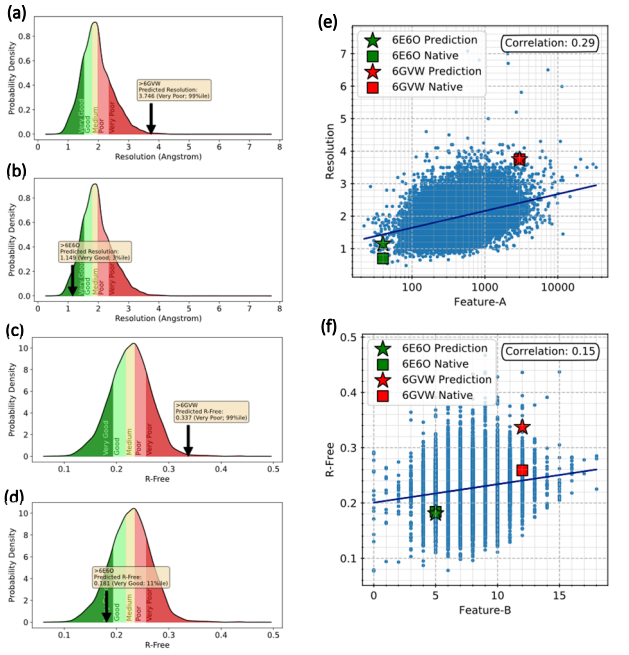

Fig. 5. Case studies that demonstrate predictions from XRRpred for the ER-1 pheromone from *Euplotes raikovi* (PDB ID: 6E6O) and the BRCA1-A complex from *Mus musculus* (PDB ID: 6GVW). Panels a and c illustrate the resolution and R-free predictions for the BRCA1-A complex, respectively. Panels b and d focus on the predictions of resolution and R-free for the ER-1 pheromone, respectively. Panel e shows the relationship between experimental resolution and one of the input features (feature-A that quantifies the sum of input chain lengths) for the test proteins; each point represents one protein. Panel f depicts the relationship between experimental R-free and another input features (feature-B: the maximum number of flexible residues within in a sliding window of size 20 across all chains) for the test proteins. The R-free and resolution values predicted by XRRpred and the corresponding experimental (native) values for the two proteins are identified with the star (for prediction) and square (for native) markers in panels e (for resolution) and f (for R-free). We color-code the two markers in green for the ER-1 pheromone and in red for the BRCA1-A complex. The dark blue line in panels e and f is the linear regression line. The PCC values between a given structure quality measure and the corresponding feature are given in the top right corner in these panels.

3.5 Case studies

We use two test proteins to explain the working of the XRRpred model and to illustrate how its results are presented to the end users. Our objective here is not to compare or evaluate the expected predictive performance but to explain how to interpret XRRpred's predictions and how these predictions are linked to the underlying input features. The examples cover two diverse cases, one that considers a poor-quality structure and the other that presents a good-quality structure.

The relatively poor-quality structure of the BRCA1-A complex from *Mus musculus* (PDB ID:6GVW) has the low resolution of 3.75 Å and a modest-quality R-free of 0.259. **Fig. 5a** and **5c** show that XRRpred predicts resolution of 3.746 and R-free of 0.337, which are relatively close to the experimental values. These figures also illustrate how the results produced by the XRRpred's webserver are visualized. We divide the experimental resolution and the R-free values into five color-coded ranges: *very good* (dark green), *good* (light green), *medium* (yellow), *poor* (light red) and *very poor* (dark red) scores. This helps the end users in placing the predicted values into a proper context. The results produced by XRRpred, which are shown in the yellow callout box, include the predicted resolution/R-free, the corresponding range name, and the percentile with respect to the distribution of the corresponding experimental values. For instance, the XRRpred's prediction of resolution for the BRCA1-A complex (**Fig. 5a**) includes the 3.746 putative resolution which is identified as very poor and placed in the 99th percentile of the resolution values.

We also explain how XRRpred arrived at this prediction. The scatter plots in **Fig. 5e** and **5f** show the relations between two input features (feature-A: sum of input chain lengths; feature-B: max number of flexible residues within in the sliding window of size 20 across all chains) and the experimental values of resolution and R-free for the test proteins. Feature-A has a modest (0.29) correlation with resolution, which means that the resolution worsen as the size of the input protein gets larger. Feature-B has a low (0.15) correlation with R-free, which suggests that R-free worsens for proteins that have sequence regions with many highly flexible residues. Both of these features have relatively large values for the BRCA1-A complex (see red star markers in **Fig. 5e** and **5f**), and this in part leads the XRRpred's regressors to predict relatively large values of resolution and R-free.

The second case study concerns a good-quality structure of the ER-1 pheromone from *Euplotes raikovi* (PDB ID: 6E6O). This structure has the atomic level resolution of 0.70 and a good R-free of 0.184. **Fig. 5b** and **5d** reveal that XRRpred predicts resolution of 1.149 (in the very good range) and R-free of 0.181 (in the medium range) for this protein. These predictions are in a good agreement with the experimental data. The predictions stem in part from the low values of feature-A (i.e., this is a small protein; see green star marker in **Fig. 5e**) and feature-B (i.e., this protein does not include regions with many flexible residues; see green star marker in **Fig. 5f**). The regressors used by XRRpred utilize many such features together to provide accurate predictions of the structure quality.

3.6 XRRpred webserver and standalone code

We provide a webserver that implements the XRRpred predictor at http://biomine.cs.vcu.edu/servers/XRRPred/. Users provide the input protein sequences in the FASTA format where individual chains for a multi-chain protein must share a common prefix in their identifiers. We provide explanation and examples of the input format on the webserver page. Users benefit from a batch processing of predictions. We allow up to 50 sequences in a single run. We offer an option to provide an email address where the notification of the completion of the prediction and the link to the results is sent after the webserver processes the user's query. The processing of the predictions and the results are also available via the browser window. We process the user's requests using a queue that serves multiple webservers from our lab and which ensures a proper load balancing between users. The entire prediction process is automated and done on the server side, freeing the end users from installing software and having access to computational hardware. We visualize the results, which include the predicted values of resolution and R-free, using the graphics shown in Figures 5a, 5b, 5c and 5d. We explain these graphics in Section 3.5. We also provide the results in a parsable comma-separable csv file.

The standalone source code for XRRpred is available at https://github.com/sinaghadermarzi/XRRpred-predictor, with a convenient docker version at https://github.com/sinaghadermarzi/XRRpred-docker.

4 Summary

X-ray crystallography is the main driver to solve protein structures (Grabowski, et al., 2016). However, these structures vary widely in their quality. The last decade has produced accurate tools that identify sequences that produce diffraction quality crystals and thus can be solved via the X-ray crystallography (Elbasir, et al., 2020; Elbasir, et al., 2019; Gao, et al., 2018; Wang, et al., 2018; Zhu, et al., 2020). This calls for the development of tools that predict the quality of structures for the crystallizable proteins.

We introduce the first predictor of the protein structure quality, XRRpred, which targets prediction of the two key structure quality measures: resolution and R-free. XRRpred relies on original sequence profiles, hand-crafted features, and an extensive design process that utilizes modern

resampling. Empirical tests on an independent (low similarity to the training data) test set show that XRRpred provides accurate predictions for resolution and R-free. We find that the inclusion of resampling provides statistically significant improvements while the other design considerations (feature and regressor selection) provide modest benefits. We show that XRRpred's predictions correctly model correlation between resolution and R-free, reproduce structure quality relations between structural classes of proteins, and suggest favorable structure quality for the commonly found clusters of different structures for identical protein sequences. Tests reveal that XRRpred significantly outperforms alternative indirect ways to predict the structure quality, such as predictors of crystallization propensity and alignment. XRRpred server at http://biomine.cs.vcu.edu/servers/XRRPred/ processes user's requests on the server side, allows batch predictions, offers informative visualization of the results, and provides links to the standalone software.

As a potential future direction, one can consider using other inputs, beyond the sequence. These could include details concerning experimental parameters of the crystallization protocol, hardware used, and taxonomy. While inclusion of these data would likely result in an improved predictive performance, it would also constrain applications to the scenarios where information about these factors is available and where these factors are covered in the training dataset. XRRpred uses sequence as the sole input, which means that it can be applied to any protein for which sequence is known, even if it would be solved using hardware or protocols that were not explicitly included in the training dataset. Another option is to consider other types of methods used to solve protein structures, such as nuclear magnetic resonance (NMR) and cryogenic electron microscopy (cryo-EM). While these methods are much less popular than X-ray crystallography, this can change in the future. In particular, recent and rapid developments in the cryo-EM technology (Callaway, 2020; Garcia-Nafria and Tate, 2020) position this technology as a likely target for our future development efforts.

Funding

This work was supported in part by the Robert J. Mattauch Endowment to L.K.

Conflict of Interest: none declared.

References

Altschul, S.F., *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25(17):3389-3402.

Bach, F. and Moulines, E. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In, *Proceedings of the 24th International Conference on Neural Information Processing Systems*. Granada, Spain: Curran Associates Inc.; 2011. p. 451–459.

Branco, P. and Torgo, L. A Study on the Impact of Data Characteristics in Imbalanced Regression Tasks. In, *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE; 2019. p. 193-202.

Branco, P., Torgo, L. and Ribeiro, R.P. Pre-processing approaches for imbalanced distributions in regression. *Neurocomputing* 2019;343:76-99.

Breiman, L. Classification and regression trees. Belmont, Calif.: Wadsworth International Group; 1984.

Brunger, A.T. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* 1992;355(6359):472-475.

Callaway, E. 'It opens up a whole new universe': Revolutionary microscopy technique sees individual atoms for first time. *Nature* 2020;582(7811):156-157.

Canaves, J.M., *et al.* Protein biophysical properties that correlate with crystallization success in Thermotoga maritima: Maximum clustering strategy for structural genomics. *Journal of Molecular Biology* 2004;344(4):977-991.

Chandonia, J.M., Kim, S.H. and Brenner, S.E. Target selection and deselection at the Berkeley Structural Genomics Center. *Proteins* 2006;62(2):356-370.

Crammer, K., *et al.* Online passive-aggressive algorithms. *J Mach Learn Res* 2006;7:551-585.

Dosztányi, Z., *et al.* IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 2005;21(16):3433-3434.

Dubach, V.R. and Guskov, A. The Resolution in X-ray Crystallography and Single-Particle Cryogenic Electron Microscopy. *Crystals* 2020;10(7):580.

Elbasir, A., *et al.* BCrystal: an interpretable sequence-based protein crystallization predictor. *Bioinformatics* 2020;36(5):1429-1438.

Elbasir, A., *et al.* DeepCrystal: a deep learning framework for sequence-based protein crystallization prediction. *Bioinformatics* 2019;35(13):2216-2225.

Faraggi, E., Zhou, Y. and Kloczkowski, A. Accurate single‐sequence prediction of solvent accessible surface area using local and global features. *Proteins: Structure, Function, and Bioinformatics* 2014;82(11):3170-3176.

Fernández-Ballester, G., *et al.* Ionic channels as targets for drug design: a review on computational methods. *Pharmaceutics* 2011;3(4):932-953.

Fernández, A., *et al.* Learning from imbalanced data sets. Springer; 2018.

Gao, J., *et al.* Survey of Predictors of Propensity for Protein Production and Crystallization with Application to Predict Resolution of Crystal Structures. *Curr Protein Pept Sci* 2018;19(2):200-210.

Gao, J., *et al.* Survey of predictors of propensity for protein production and crystallization with application to predict resolution of crystal structures. *Current Protein and Peptide Science* 2018;19(2):200-210.

Garcia-Nafria, J. and Tate, C.G. Cryo-Electron Microscopy: Moving Beyond X-Ray Crystal Structures for Drug Receptors and Drug Development. *Annu Rev Pharmacol Toxicol* 2020;60:51-71.

Grabowski, M., *et al.* The impact of structural genomics: the first quindecennial. *J Struct Funct Genomics* 2016;17(1):1-16.

Graslund, S., *et al.* Protein production and purification. *Nat Methods* 2008;5(2):135-146.

Grey, J.L. and Thompson, D.H. Challenges and opportunities for new protein crystallization strategies in structure-based drug design. *Expert Opin Drug Discov* 2010;5(11):1039-1045.

Grey, J.L. and Thompson, D.H. Challenges and opportunities for new protein crystallization strategies in structure-based drug design. *Expert opinion on drug discovery* 2010;5(11):1039-1045.

Hu, G., *et al.* Taxonomic Landscape of the Dark Proteomes: Whole-Proteome Scale Interplay Between Structural Darkness, Intrinsic Disorder, and Crystallization Propensity. *Proteomics* 2018:e1800243.

Hu, G., *et al.* Quality assessment for the putative intrinsic disorder in proteins. *Bioinformatics* 2019;35(10):1692-1700.

Ilari, A. and Savino, C. Protein structure determination by x-ray crystallography. *Methods Mol Biol* 2008;452:63-87.

Jahandideh, S., Jaroszewski, L. and Godzik, A. Improving the chances of successful protein structure determination with a random forest classifier. *Acta Crystallogr D Biol Crystallogr* 2014;70(Pt 3):627-635.

Jazayeri, A., Dias, J.M. and Marshall, F.H. From G Protein-coupled Receptor Structure Resolution to Rational Drug Design. *J Biol Chem* 2015;290(32):19489-19495.

Jazayeri, A., Dias, J.M. and Marshall, F.H. From G protein-coupled receptor structure resolution to rational drug design. *Journal of Biological Chemistry* 2015;290(32):19489-19495.

Kim, S.H., *et al.* Structure-based functional inference in structural genomics. *J Struct Funct Genomics* 2003;4(2-3):129-135.

Kleywegt, G.J. and Jones, T.A. Model building and refinement practice. *Methods Enzymol* 1997;277:208-230.

Koziarski, M., Wozniak, M. and Krawczyk, B. Combined Cleaning and Resampling algorithm for multi-class imbalanced data with label noise. *Knowl-Based Syst* 2020;204.

Krawczyk, B. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence* 2016;5(4):221-232.

Krawczyk, B., Koziarski, M. and Wozniak, M. Radial-Based Oversampling for Multiclass Imbalanced Data Classification. *Ieee T Neur Net Lear* 2020;31(8):2818-2831.

Kurgan, L. and Mizianty, M.J. Sequence-Based Protein Crystallization Propensity Prediction for Structural Genomics: Review and Comparative Analysis. *Nat. Science* 2009;1(2):93-106.

Kurgan, L., *et al.* CRYSTALP2: sequence-based protein crystallization propensity prediction. *Bmc Structural Biology* 2009;9.

Lieberman, R.L., Peek, M.E. and Watkins, J.D. Determination of soluble and membrane protein structures by X-ray crystallography. *Methods Mol Biol* 2013;955:475-493.

Marsden, R.L. and Orengo, C.A. Target selection for structural genomics: an overview. *Methods Mol Biol* 2008;426:3-25.

Maveyraud, L. and Mourey, L. Protein X-ray Crystallography and Drug Discovery. *Molecules* 2020;25(5).

Meng, F. and Kurgan, L. High-throughput prediction of disordered moonlighting regions in protein sequences. *Proteins* 2018;86(10):1097-1110.

Meng, F., Wang, C. and Kurgan, L. fDETECT webserver: fast predictor of propensity for protein production, purification, and crystallization. *BMC bioinformatics* 2017;18(1):580.

Michie, A.D., Orengo, C.A. and Thornton, J.M. Analysis of domain structural class using an automated class assignment protocol. *J Mol Biol* 1996;262(2):168-185.

Mizianty, M.J., *et al.* Covering complete proteomes with X-ray structures: a current snapshot. *Acta Crystallogr D Biol Crystallogr* 2014;70(Pt 11):2781-2793.

Mizianty, M.J. and Kurgan, L. Sequence-based prediction of protein crystallization, purification and production propensity. *Bioinformatics* 2011;27(13):i24-33.
Morris, R.J. and Bricogne, G. Sheldrick's 1.2 angstrom rule and beyond. *Acta Crystallographica Section D-Structural Biology* 2003;59:615-617.
Movshovitz‐Attias, D., London, N. and Schueler‐Furman, O. On the use of structural templates for high‐resolution docking. *Proteins: Structure, Function, and Bioinformatics* 2010;78(8):1939-1949.
Oldfield, C.J.*, et al.* Utilization of protein intrinsic disorder knowledge in structural proteomics. *Biochim Biophys Acta* 2013;1834(2):487-498.
Orengo, C.A.*, et al.* The CATH protein family database: a resource for structural and functional annotation of genomes. *Proteomics* 2002;2(1):11-21.
Park, H., Lee, H. and Seok, C. High-resolution protein–protein docking by global optimization: recent advances and future challenges. *Current opinion in structural biology* 2015;35:24-31.
Read, R.J.*, et al.* A new generation of crystallographic validation tools for the protein data bank. *Structure* 2011;19(10):1395-1412.
Robin, G.*, et al.* A general target selection method for crystallographic proteomics. *Methods Mol Biol* 2008;426:27-35.
Rupp, B. and Wang, J. Predictive models for protein crystallization. *Methods* 2004;34(3):390-407.
Seeliger, D. and de Groot, B.L. Atomic contacts in protein structures. A detailed analysis of atomic radii, packing, and overlaps. *Proteins* 2007;68(3):595-601.
Sillitoe, I.*, et al.* CATH: increased structural coverage of functional space. *Nucleic Acids Res* 2021;49(D1):D266-D273.
Slabinski, L.*, et al.* The challenge of protein structure determination--lessons from structural genomics. *Protein Sci* 2007;16(11):2472-2482.
Terwilliger, T.C., Stuart, D. and Yokoyama, S. Lessons from structural genomics. *Annu Rev Biophys* 2009;38:371-383.
Torgo, L.*, et al.* Smote for regression. In, *Portuguese conference on artificial intelligence*. Springer; 2013. p. 378-389.
Walsh, I.*, et al.* Comprehensive large-scale assessment of intrinsic protein disorder. *Bioinformatics* 2015;31(2):201-208.
Wang, H.*, et al.* Critical evaluation of bioinformatics tools for the prediction of protein crystallization propensity. *Brief Bioinform* 2018;19(5):838-852.
Wang, H.*, et al.* Crysalis: an integrated server for computational analysis and design of protein crystallization. *Sci Rep* 2016;6:21383.
wwPDB consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res* 2019;47(D1):D520-D528.
Yan, J. and Kurgan, L. DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. *Nucleic Acids Res* 2017;45(10):e84.
Zhang, C. and Kim, S.H. Overview of structural genomics: from structure to function. *Curr Opin Chem Biol* 2003;7(1):28-32.
Zhang, J. and Kurgan, L. SCRIBER: accurate and partner type-specific prediction of protein-binding residues from proteins sequences. *Bioinformatics* 2019;35(14):i343-i353.
Zhu, Y.H.*, et al.* Accurate multistage prediction of protein crystallization propensity using deep-cascade forest with sequence-based features. *Brief Bioinform* 2020.
Zimmerman, M.D.*, et al.* Data management in the modern structural biology and biomedical research environment. *Methods Mol Biol* 2014;1140:1-25.
Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B* 2005;67(2):301-320.