# Machine Learning to Predict Protein Crystallizability

Marta Korpacz

## 1 Introduction

Protein crystallization is one of the laboratory techniques that allow for an extremely accurate analysis of the three-dimensional structure at the atomic level and an understanding of their interactions with other types of molecules such as ligands or RNA. Therefore, the obtained results can be used in drug design, biochemistry, and bioinformatics, as a template for predicting the structure of homologs using computer tools, etc. While many tools, such as AlphaFold[1] and ESMfold[2], have appeared in recent years for predicting structure without conducting a wet experiment, their results are not taken for granted and the result of the crystallization is still considered the most reliable and irreplaceable[3]. Despite the gradual decrease in costs along with the development of techniques, it is still an expensive and long-lasting process. Moreover, a crystal is obtained on average in 30% of trials, and only in some of them, a high quality of diffraction is achieved[4].

Success depends on many factors: the selected technique, the experience of the laboratory team, conditions in the laboratory such as humidity and temperature, the reagents used, but also on the protein itself and its biochemical properties. For this reason, many attempts have been made to create tools to predict whether the crystallization of a given protein is possible under standard laboratory conditions[4, 5, 6, 7].

The idea is to start from the sequence and, thanks to machine learning, predict the crystallographic potential of the tested proteins and, as a result, select the most optimal sequence or sequences from a given set. In creating a model, the challenge is to choose the appropriate architecture, obtain a training set, but also, perhaps above all, select the features that most influence the possibility of crystallization and will be useful in prediction without constituting unnecessary noise. This allows not only to create a tool that may be useful to scientists but also to draw conclusions about biochemical properties that influence the success of crystallization.

XtalPred[6] and XtalPred-RF[8] achieved particular success here, characterized by high prediction efficiency and a wide set of features taken into account.

XtalPred, published in 2007, was based on the Expert Pool method and took into account length, gravy index, isoelectric point (pI), instability index, predicted disorder, and insertion score. Seven years later, the method was improved by expanding the set of features to include surface-related attributes and changing the model architecture to Random Forest[9]. This enhanced version was published as XtalPred-RF and in this work it constitutes a reference point as to the effectiveness of the proposed solutions.

## 2 Methods and Materials

### 2.1 Dataset

In the referenced literature, i.e. in the publications regarding XtalPred[6, 8], XANNpred[5] and CrystalP2[4], records from the PepcDB, TargetDB databases or the database integrating PepcDB and TargetDB, i.e. TargetTrack, were used to create the data set. The database stopped being updated in 2015, but over the fifteen years of its operation, it managed to collect over 350,000 protein sequences as well as the results of their crystallization attempts undertaken by researchers from 35 different research centers[10]. The uniqueness of the database is the mention of the last successful step in the experiment, which, for example, allowed the creators of XANNpred to create a tool that also predicts the chances of success at each of the standard stages of protein processing for crystallization. Moreover sharing the results of unsuccessful attempts is not common in current science. Including sequences that have failed to crystallize is crucial to creating a balanced dataset, which may contribute to effective training and less biased predictions[11].

This work uses a dataset published by the authors of XtalPred, which can be found under this URL: `https://XtalPred.godziklab.org/XtalPred/data.tar`. This is a relatively small set (only 9505 sequences, 4 of which were rejected due to containing non-standard amino acids). The exact steps for obtaining the set are described in Table 1 in the publication *Improving the chances of successful protein structure determination with a random forest classifier*[8], but it is worth emphasizing the key assumptions. A negative label was assigned to proteins whose last stage was ”purified”, which means that the set did not include proteins that were not expressed or whose proper purification failed. Excluded were also targets with predicted signal peptides and transmembrane helices as, according to the researchers, crystallization in this case is extremely difficult. The sequences were also clustered at an identity level of 66%, and the division between the test and training sets was based on the results of the PSI-BLAST program[12] to avoid leakage in the form of similar proteins being found in different sets.

## 2.2 Features

Three sets of features were used in the work:

(A) Features not related to the surface:
length, gravy (grand average of hydropathy), instability index, isoelectric point, count of cysteines, count of methionines, count of tryptophans, count of tyrosines, count of phenylalanines, molecular weight, aromaticity, helix fraction, turn fraction, sheet fraction, extinction coefficient (reduced), extinction coefficient (oxidized), longest disorder region, total percentage of disorder.

(B) Dataset (A) and features related to the surface, which means ASA, RSA, surface ruggedness, weighted gravy, count of amino acids weighted.

(C) Dataset (B) with additional features as average ASA, average RSA, number of amino acids on the surface, weighted disorder, and number of amino acids predicted as disorder, that are not at the surface.

Data set (A), similar to the basic set of features proposed by XtalPred-RF, consists of features calculated directly based on amino acid composition and predicted disorder regions (fragments of the protein that have no fixed tertiary structure). The main difference is the exclusion of the insertion score, a parameter dependent on the similarity to homologs calculated using BLAST[13]. The reason for abandoning this feature is its dependence on available sequence databases, which limits the tool's ability to accurately predict the crystallization potential of *de novo* designed proteins. Additionally, mutations that could potentially enhance crystallographic properties may be assessed negatively based on the insertion score, indicating that this parameter is not ideal from a biological point of view. Moreover, the growth rate of the Uniprot database is rapid[14], which means that the homologs obtained at the time of publication of XtalPred will probably not be the same as in 2024.

Dataset (B) refers to the set extended by the authors of XtalPred in 2014 with features related to the surface.

"Weighted" means that features were calculated with the following formula proposed in the literature[8]:

$$\hat{f} = \frac{\sum_{i=1}^{N} f_i \cdot RSA_i}{\sum_{i=1}^{N} \cdot RSA_i}$$

where RSA corresponds to relative solvent accessibility, $\hat{f}$ is the weighted feature and $f_i$ is the value of the feature for i-th amino acid in the sequence.

Surface ruggedness, in turn, is a feature intended to reflect the number of cavities and protrusions on the protein surface concerning the size of the protein:

$$SR = \frac{\sum_{i=1}^{N} ASA_i}{6.3M^{0.73}}$$

ASA corresponds to Accessible Surface Area, $M$ to molecular weight, and $6.3M^{0.73}$ is a statistic proposed by Miller[15] to approximate the accessible surface predicted based on molecular mass.

The dataset (C) was additionally expanded to include features whose inclusion seemed logical from the point of view of biological knowledge and crystallographic experience by taking into account differences in solvent availability and recognizing that the bigger problem is the disorder inside the protein, which needs to be well folded. $RSA > 0.25$ was assumed as the threshold to determine whether the amino acid was present on the surface[16].

There are also differences in the tools used for calculations, as presented in the table 1. The change was motivated by technical aspects related to available resources and the desire to use the most effective tools possible. A particularly notable difference is the version change in NetSurfP[17], with NetSurfP-3.0 being over 600 times faster than its predecessors[18]. Despite this significant speedup, these calculations still represent the largest computational cost. Additionally, the new versions demonstrate greater accuracy in benchmarks.

| **Feature** | **XtalPred** | **This work** |
|---|---|---|
| Surface Analisys | NetSurfP[17] | NetSurfP-3.0[18] |
| Disorder Prediction | Disopred-2.0[19] | Iupred2a[20] |
| Transmembrane Helices | TMHMM[21] | DeepTMHMM[22] |
| Isoelectric Point | ? | BioPython[23] |
| Secondary Structure | PSIPRED[24] | BioPython |

Table 1: Differences in methods used for feature calculation between this work and XtalPred[8]. The authors did not declare the tool used for Isoelectric Point calculation.

It is also worth emphasizing that the transmembrane region prediction tool is not used, as in the set provided by XtalPred there is not a single protein with a transmembrane helix due to the assumption of its authors that its presence predetermines failure in crystallization. The tool is therefore proposed and chosen to enable the model to be used on proteins from outside the proposed set, however, it is worth noting that transmembrane proteins are also deposited in the Protein Data Bank (PDB)[25], which proves the possibility of their crystallization[26, 27].

The data was preprocessed using *StandardScaler* implemented in Scikit-Learn[28] which use the formula: $z = \frac{x-\mu}{\sigma}$, where $x$ is the training sample, $\mu$ is the mean of the $x$ and $\sigma$ is the standard deviation of $x$.

### 2.3 Hyperparameters tuning and model selection

Many different classifier architectures were tested, including sequential neural networks (MLPs) with Adam and RMSprop optimizers at learning rates of 0.01, 0.001 and 0.0001 using layers implemented in TensorFlow[29] and Random Forest, Gaussian Naive Bayes, Logistic Regression, SVM (Support Vector Machine), Gradient Boosting and KNN (k-nearest neighbors) implemented in the Scikit-Learn[28] package. Additionally, two configurations of stacking classifiers were tested, where the final prediction is obtained using linear regression based on the results of the stacked models. In the case of classifiers implemented by Scikit-Learn, a grid search of hyperparameters was also performed using 5-fold cross-validation. This means that for each of the selected hyperparameter configurations, five training sessions were carried out, each time treating disjoint subsets of the training set (always 20% of the initial set) as the validation set. 5-fold was chosen as a compromise between the reduction in the size of the training set and the computational cost resulting from repeating the training $n$ times. In the case of sequential models, the validation set consisted of a randomly selected 20% of the training set samples.

The main metric in the selection of hyperparameters and models was accuracy, but sensitivity, specificity and Matthews Correlation Coefficient (MCC)[30] were also taken into account in the comparison of the final results.

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

$$\text{Specificity} = \frac{TN}{TN+FP}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

$TP$ - true positives, $TN$ - true negatives, $FP$ - false positives, $FN$ - false negatives. Apart from MCC, the metrics results oscillate between 0 and 1 (the best possible result), and in MCC between -1 and 1, where -1 means assigning reverse labels and 1 means perfect prediction.

In order not to deviate from the approach presented in the reference work[8], the architectures were finally compared based on the evaluation of the test set.

# 3 Results

## 3.1 Dataset A

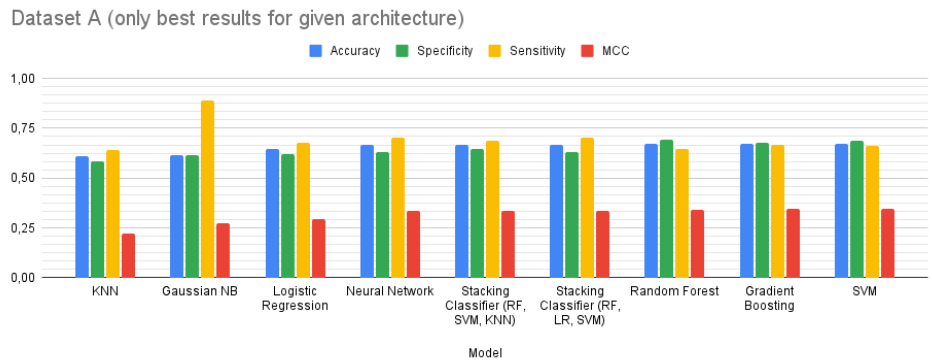


Figure 1: Results sorted ascending by accuracy. For each architecture, the values obtained for the best set of hyperparameters are shown.

As mentioned, dataset A is inspired by the XtalPreda base set, and therefore the results are compared to the accuracy declared in the mentioned work[8], which was 68%. The plots in Fig. 1 present the results obtained.

Achieved accuracy (67% for SVM, Gradient Boosting, and Random Forest) does not differ much from those mentioned in the publication, as a 1% change is not statistically significant and may depend on the seed used and random factors. What is particularly important, the obtained results are more balanced in terms of sensitivity and specificity than in the revised work (see Fig. 2), which means that incorrect label assignment during prediction is more balanced. It remains to be discussed whether false positive or false negative tendencies are more optimal in the context of crystallization. Specifically, it is important to determine whether it is better to attempt crystallizing a protein with low crystallization potential or to abandon efforts to crystallize a protein that is easy to crystallize. The

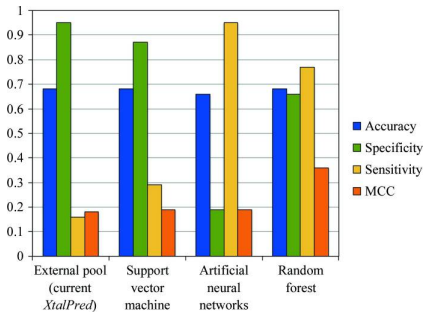


Figure 2: The results obtained at basic feature set published as Figure 1 in work *Improving the chances of successful protein structure determination with a random forest classifier*[8].

optimal approach may depend on the research purpose. For instance, in some protein families, there are only a few proteins with high crystallographic potential, making false negatives particularly harmful. Conversely, in sets of proteins with high crystallization potential, false positives could lead to unnecessary costs and efforts.

### 3.2 Dataset B

Efficiency in terms of each of the metrics used was achieved by expanding the set to include features related to the area. However, they did not exceed the final accuracy of XtalPred-RF (74%), as can be seen in Fig. 3. Here, too, the difference remains small, as the Stacking Classifier (RF, LR, SVM) has achieved an accuracy of 72.6%.

The parameters of the used Stacking Classifier are 200 estimators and 10 as maximal depth for Random Forest, 'L2' as a penalty, C = 100, and a maximal number of iterations set to 300 for logistic regression and in case of SVM C = 100, gamma = 0.001 and Radial Basis Function as the kernel.

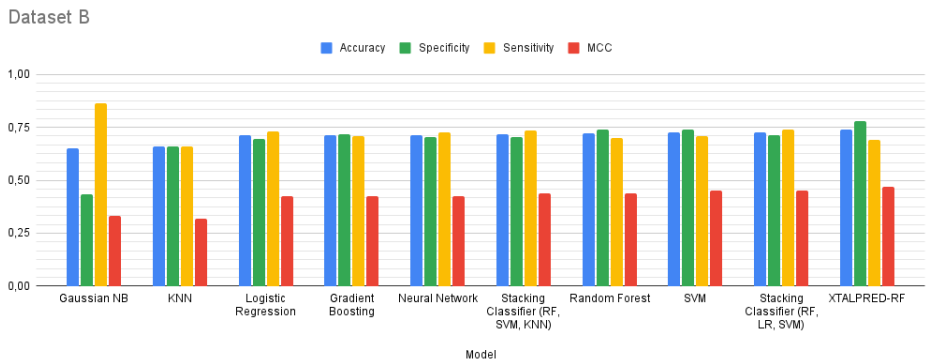


Figure 3: Results obtained for the test set at the model trained on feature set B presented as in Fig. 1

### 3.3 Dataset C

The last set of features allowed us to achieve an accuracy as in the reference work and a higher MCC value, which proves high effectiveness (Fig. 4). The improvement compared to previous data sets applies to all models, which means that expanding the data set with the proposed features is a beneficial action. Hyperparameters remain as in the previous section.

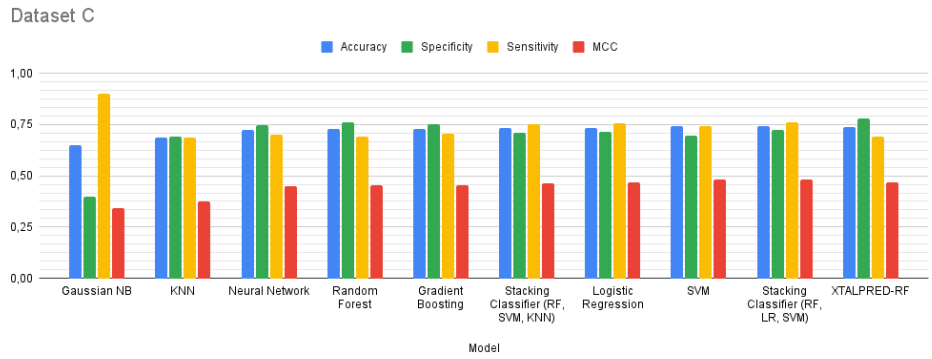


Figure 4: Results obtained for dataset C.

### 3.4 Conclusion

In all cases, the worst results were achieved by the KNN and Naive Bayes classifier. The effectiveness of the remaining methods turned out to be quite similar. However, unlike XtalPred-RF the best results were achieved by using SVM or stacking it with other classifiers. It is worth noting, however, that Random Forest's results were not significantly worse.

The results prove the possibility of predicting the success or failure of crystallization with high efficiency based on the sequence and the importance of selecting features, in particular taking into account predictions related to solvent availability and protein surface. Achieving effectiveness comparable to XtalPred using faster feature prediction tools and the lack of reference to homologs proves the possibility of using tools of this type in the selection of de novo designed or mutated sequences and increasing their availability for laboratories without large computer resources. Importantly, many families deposited in the InterPro database[31] contain several thousand sequences, which means that the time needed to calculate features will be as long as the time needed to calculate them to create a training and test set.

It is worth noting that some of the features (dependent on NetSurfP and disorder results) are predicted, so the effectiveness of the model based on them also depends on the accuracy of the mentioned tools.

## References

[1] Jumper, J. *et al.* Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589 (2021). URL http://dx.doi.org/10.1038/s41586-021-03819-2.

[2] Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023). URL http://dx.doi.org/10.1126/science.ade2574.

[3] Terwilliger, T. C. *et al.* Alphafold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. *Nature Methods* **21**, 110–116 (2023). URL http://dx.doi.org/10.1038/s41592-023-02087-4.

[4] Kurgan, L. *et al.* Crystalp2: sequence-based protein crystallization propensity prediction. *BMC Structural Biology* **9**, 50 (2009). URL http://dx.doi.org/10.1186/1472-6807-9-50.

[5] Overton, I. M., van Niekerk, C. A. J. & Barton, G. J. Xannpred: Neural nets that predict the propensity of a protein to yield diffraction-quality crystals. *Proteins: Structure, Function, and Bioinformatics* **79**, 1027–1033 (2011). URL http://dx.doi.org/10.1002/prot.22914.

[6] Slabinski, L. *et al.* Xtalpred: a web server for prediction of protein crystallizability. *Bioinformatics* **23**, 3403–3405 (2007). URL http://dx.doi.org/10.1093/bioinformatics/btm477.

[7] Ghadermarzi, S., Krawczyk, B., Song, J. & Kurgan, L. Xrrpred: accurate predictor of crystal structure quality from protein sequence. *Bioinformatics* **37**, 4366–4374 (2021). URL http://dx.doi.org/10.1093/bioinformatics/btab509.

[8] Jahandideh, S., Jaroszewski, L. & Godzik, A. Improving the chances of successful protein structure determination with a random forest classifier. *Acta Crystallographica Section D Biological Crystallography* **70**, 627–635 (2014). URL http://dx.doi.org/10.1107/s1399004713032070.

[9] Ho, T. K. *Random decision forests*, Vol. 1, 278–282 (IEEE, 1995).

[10] Helen M. Berman, M. J. G. & Protein Structure Initiative Network Of Investigators. Protein structure initiative - targettrack 2000-2017 - all data files (2017). URL https://zenodo.org/record/821654.

[11] Olson, D. L. *Data Set Balancing*, 71–80 (Springer Berlin Heidelberg, 2004). URL http://dx.doi.org/10.1007/978-3-540-30537-8_8.

[12] Altschul, S. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389–3402 (1997). URL http://dx.doi.org/10.1093/nar/25.17.3389.

[13] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403–410 (1990).

[14] Uniprot: a hub for protein information. *Nucleic Acids Research* **43**, D204–D212 (2014). URL http://dx.doi.org/10.1093/nar/gku989.

[15] Miller, S., Janin, J., Lesk, A. M. & Chothia, C. Interior and surface of monomeric proteins. *Journal of Molecular Biology* **196**, 641–656 (1987). URL http://dx.doi.org/10.1016/0022-2836(87)90038-6.

[16] Gong, H. *et al.* Improving prediction of burial state of residues by exploiting correlation among residues. *BMC Bioinformatics* **18** (2017). URL http://dx.doi.org/10.1186/s12859-017-1475-5.

[17] Petersen, B., Petersen, T. N., Andersen, P., Nielsen, M. & Lundegaard, C. A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Structural Biology* **9** (2009). URL http://dx.doi.org/10.1186/1472-6807-9-51.

[18] Høie, M. H. *et al.* Netsurfp-3.0: accurate and fast prediction of protein structural features by protein language models and deep learning. *Nucleic Acids Research* **50**, W510–W515 (2022). URL http://dx.doi.org/10.1093/nar/gkac439.

[19] Ward, J. J., McGuffin, L. J., Bryson, K., Buxton, B. F. & Jones, D. T. The disopred server for the prediction of protein disorder. *Bioinformatics* **20**, 2138–2139 (2004). URL http://dx.doi.org/10.1093/bioinformatics/bth195.

[20] Mészáros, B., Erdős, G. & Dosztányi, Z. Iupred2a: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Research* **46**, W329–W337 (2018). URL http://dx.doi.org/10.1093/nar/gky384.

[21] Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes11edited by f. cohen. *Journal of Molecular Biology* **305**, 567–580 (2001). URL http://dx.doi.org/10.1006/jmbi.2000.4315.

[22] Hallgren, J. *et al.* Deeptmhmm predicts alpha and beta transmembrane proteins using deep neural networks (2022). URL http://dx.doi.org/10.1101/2022.04.08.487609.

[23] Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009). URL https://doi.org/10.1093/bioinformatics/btp163.

[24] McGuffin, L. J., Bryson, K. & Jones, D. T. The psipred protein structure prediction server. *Bioinformatics* **16**, 404–405 (2000). URL http://dx.doi.org/10.1093/bioinformatics/16.4.404.

[25] Berman, H. M. The protein data bank. *Nucleic Acids Research* **28**, 235–242 (2000). URL http://dx.doi.org/10.1093/nar/28.1.235.

[26] Lu, P. *et al.* crystal structure of transmembrane protein tmhc2_e (2018). URL http://dx.doi.org/10.2210/pdb6B87/pdb.

[27] Rasmussen, S. *et al.* Crystal structure of the beta2 adrenergic receptor-gs protein complex (2011). URL http://dx.doi.org/10.2210/pdb3SN6/pdb.

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

[29] Abadi, M. *et al.* TensorFlow: Large-scale machine learning on heterogeneous systems (2015). URL https://www.tensorflow.org/. Software available from tensorflow.org.

[30] Matthews, B. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure* **405**, 442–451 (1975). URL http://dx.doi.org/10.1016/0005-2795(75)90109-9.

[31] Paysan-Lafosse, T. *et al.* Interpro in 2022. *Nucleic Acids Research* **51**, D418–D427 (2022). URL http://dx.doi.org/10.1093/nar/gkac993.