Naval Research Laboratory
Washington, DC 20375-5320

NRL/6910/MR—2022/2

Applying Machine Learning with Model Interpretability for Sequence-Based Protein Solubility Prediction

JEROME ANTHONY E. ALVAREZ
PATRICIA M. LEGLER
SCOTT N. DEAN

Laboratory for Bio/Nano Science and Technology Branch
Center for Bio/Molecular Science & Engineering Division

ANTHONY P. MALANOSKI

Laboratory for Biomaterials and Systems Branch
Center for Bio/Molecular Science & Engineering Division

September 16, 2022

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
16-09-2022	NRL Memorandum Report	04/01/2022 – 08/23/2022

4. TITLE AND SUBTITLE

Applying Machine Learning with Model Interpretability for Sequence-Based Protein Solubility Prediction

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Jerome Anthony E. Alvarez, Patricia M. Legler, Anthony P. Malanoski, and Dean N. Scott

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
1S67

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Research Laboratory
4555 Overlook Avenue, SW
Washington, DC 20375-5320

8. PERFORMING ORGANIZATION REPORT NUMBER

NRL/6910/MR--2022/2

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Defense Threat Reduction Agency
8725 John J Kingman Rd Ste 6201
Fort Belvoir, VA 22060

10. SPONSOR / MONITOR'S ACRONYM(S)

11. SPONSOR / MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A: Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The prediction of protein solubility is essential for basic research on natural proteins but increasingly for production and investigation of engineered or designed proteins, where experimental confirmation of the engineered properties hinges on the ability to produce it. Thus, accurate predictions of protein solubility are widely sought after by protein engineers. Here we present a new approach which uses an extreme gradient boosting (XGBoost) algorithm fed by a variety of data sources including predicted solvent\ accessibility, secondary structure, among others, to predict solubility of proteins. Our model achieves a high level of performance using a standard hold-out test set, with an overall accuracy of 72%, among the highest for sequence-based machine learning models. Critically, our system also yields information on the features important for the predictions, making use of explainable artificial intelligence to provide both local and global explainers. Using this information, we found that the certain mono-, di-, and tri-peptides are strongly associated with solubility, as are metrics for protein disorder, relative solvent accessibility, and frequency of certain secondary structures, each of which are derived from other prediction models. Critically, in our graphical user interface for the model, we make use of local explanations to help inform the reasoning behind the predictions and suggest modifications. Our model's accuracy paired with its interpretability should allow for rapid prediction of protein solubility, in particular for proteins and protein families without reliable structural information. This should greatly enhance our ability to experimentally produce and investigate proteins designed by machine learning-guided approaches and other protein engineering strategies.

15. SUBJECT TERMS

Machine learning Proteins

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					Dean N. Scott
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**	U	21	**19b. TELEPHONE NUMBER** *(include area code)*
U	U	U			(703) 587-4188

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

This page intentionally left blank.

CONTENTS

This page intentionally left blank.

EXECUTIVE SUMMARY

The prediction of protein solubility is essential for basic research on natural proteins but increasingly for production and investigation of engineered or designed proteins, where experimental confirmation of the engineered properties hinges on the ability to produce it. Thus, accurate predictions of protein solubility are widely sought after by protein engineers. Here we present a new approach which uses an extreme gradient boosting (XGBoost) algorithm fed by a variety of data sources including predicted solvent accessibility, secondary structure, among others, to predict solubility of proteins. Our model achieves a high level of performance using a standard hold-out test set, with an overall accuracy of 72%, among the highest for sequence-based machine learning models. Critically, our system also yields information on the features important for the predictions, making use of explainable artificial intelligence to provide both local and global explainers. Using this information, we found that the certain mono-, di-, and tri-peptides are strongly associated with solubility, as are metrics for protein disorder, relative solvent accessibility, and frequency of certain secondary structures, each of which are derived from other prediction models. Critically, in our graphical user interface for the model, we make use of local explanations to help inform the reasoning behind the predictions and suggest modifications. Our model's accuracy paired with its interpretability should allow for rapid prediction of protein solubility, in particular for proteins and protein families without reliable structural information. This should greatly enhance our ability to experimentally produce and investigate proteins designed by machine learning-guided approaches and other protein engineering strategies.

This report presents research conducted by:

Alvarez, Jerome Anthony E. [1]

Legler, Patricia M. [2]

Malanoski, Anthony P. [2]

Dean, Scott N. [2]

[1] STEM Student Employment Program, Center for Bio/Molecular Science and Engineering, US Naval Research Laboratory, Washington, DC, United States.

[2] Center for Bio/Molecular Science and Engineering, US Naval Research Laboratory, Washington, DC, United States

This page intentionally left blank.

APPLYING MACHINE LEARNING WITH MODEL INTERPRETABILITY FOR SEQUENCE-BASED PROTEIN SOLUBILITY PREDICTION

INTRODUCTION

The rise in protein engineering and design via machine learning (ML) and deep learning models has highlighted a major hurdle for experimental validation of computer-generated protein sequences: functional sequences produced through these methods can be rare and are commonly overwhelmed by nonfunctional sequences and those that are not superior to the existing examples. In a recent review by Frances Arnold, awarded a Nobel Prize in Chemistry for pioneering directed evolution to engineer enzymes, she acknowledged the difficulty in filtering out functional sequences from machine learning-direct design strategies (Yang, Wu et al. 2019). A large proportion of these sequences are insoluble or otherwise difficult to produce. Additionally, the new sequences typically must be synthesized which can be costly (~$300 per open reading frame). The prediction of protein solubility is essential for basic research on natural proteins but increasingly for production and investigation of engineered or design proteins, where experimental confirmation of engineered properties hinges on the ability to produce it. Thus, accurate predictions of protein solubility are widely sought after by protein engineers, and essential for experimentation regardless of final functionality determination.

Many proteins – not exclusively engineered or designed ones but including those naturally occurring – when heterologously expressed using standard approaches in either *Escherichia coli* or other cells, have low solubility, which greatly reduces their ability to be produced in quantities sufficient for experimentation. Prior to production and evaluation of protein solubility in aqueous buffers, many steps can be taken to enhance the chances of recovering a soluble product, including lowering culture temperatures, growth media modifications, fusion tags or proteins, co-expression of T7 lysozyme to more tightly control expression of toxic proteins among other optimizations (Francis and Page 2010). However, this process can be tedious; particularly for machine learning-guided protein design where many novel proteins are to be evaluated to establish trends and no single protein is permitted sufficient time or expense for solubility troubleshooting. In directed evolution strategies, only soluble and functional proteins are evolved; however, in ML approaches the solubility of a predicted sequence is typically unknown. This issue sparked the development of protein solubility predictors based on protein sequence alone, which were aimed to both replace wet-lab experiments and troubleshooting by prefiltering the mostly likely successful proteins *in silico* while avoiding more computationally expensive solubility prediction model reliant on protein structure prediction.

Although recent advancements in protein three dimensional structure modeling via deep learning, most prominently in the competition-winning predictors I-TASSER (Yang, Yan et al. 2015) and more recently AlphaFold (Jumper, Evans et al. 2021), has led to advancements in solubility predictions, these models are extremely computationally expensive and require some knowledge of near-neighbor three dimensional structures, making them difficult to apply to thousands of newly-generated sequences. Thus sequence-based methods for solubility prediction have significant value, particularly for fields of study without a large number of structures or machine learning-based protein engineering or design. In addition, many studies have demonstrated that protein's solubility is predominantly a function of its sequence. In particular, protein solubility has been shown to be correlated with a range of sequence properties including length, frequency of different amino acid types, and net charge (Herbert 1999, Hebditch, Carballo-Amador et al. 2017, Hou, Bourgeas et al. 2018). These well-studied associations have helped create the attempts at accurate sequence-based solubility prediction models we have available today.

Protein solubility models are generally classifiers outputting a binary soluble or insoluble, as quantitative solubility information paired with sequences is insufficient for training machine learning models. SOLpro and PROSO both used support vector machines (SVMs) (Magnan, Randall et al. 2009, Smialowski, Doose et al. 2012). PaRSnIP, or PRotein SolubIlity Predictor, used a gradient boosting

Manuscript approved September 14, 2022.

machine model trained on a variety of data sources, including components of SCRATCH (Cheng, Randall et al. 2005), finding certain tripeptides and fraction of exposed residues as the most important features (Rawi, Mall et al. 2018). DeepSol, using a deep learning model, achieved accuracy of ~76%, while EPSOL, which used PaRSnIP output as a data source for its novel deep learning model, too achieved a high level of accuracy (Khurana, Rawi et al. 2018, Wu and Yu 2021). More recently, DSResSol, using a deep learning architecture consisting of convolutional neural networks and fully connected layers achieved ~80% accuracy (Madani, Lin et al. 2021). Also recently, training on a different dataset with quantitative solubility values, GraphSol used an attentive graph convolutional network deep learning method to achieve R^2 of 0.48 where most other methods only achieve an R^2 of ~0.1, highlighting the high level of difficulty of protein solubility prediction problem (Chen, Zheng et al. 2021).

Very few of the protein solubility models above provide explanations for their predictions. The PaRSnIP model reported Rawi *et al.* was one example that attempted to provide explanations for their model available via gradient boosting model feature importance. However, these were restricted to global explainers of the whole of the training set, and did not report interpretability of predictions on individual samples, or local explanations. In this study we demonstrate the use of an XGBoost-based model for protein solubility prediction with applied explainable AI. Variable importance and SHapley Additive exPlanations (SHAP) are used to identify the importance or contribution of features used in the predictions on both a global and local level, functions that are available to XGBoost and similar tree-based methods but not for black-box methods, e.g., certain recently developed deep learning techniques. Using these methods, we found that certain tripeptides, disorder metrics, and solvent accessibility were the most influential features for solubility prediction. Using this information, we significantly pruned our model such that it obtained near state-of-the-art performance for a machine learning sequence-based model while using a smaller number of features and parameters.

METHODS

Dataset

The dataset used in this study was obtained from Rawi et al., 2018, containing only sequences and binary soluble or insoluble labeling (Rawi, Mall et al. 2018). Originally n = ~70000, we reduced the set to n = 15000 for our training set as this sample size was found to be sufficient for model evaluation. For our selection of the 15000 sequences, we kept the number of soluble and insoluble proteins at similar levels (6999 insoluble and 8001 soluble). Importantly we retained the same test set as used in Chang et al. (n = 2000) (Chang, Song et al. 2014). Since this test set has been used by a range of other reports evaluating other solubility predictor performance (Khurana, Rawi et al. 2018, Rawi, Mall et al. 2018, Madani, Lin et al. 2021) we kept this as a standard for comparison.

All features in our study, other than the solubility label, were continuous and numeric. The overall distributions and mean values for the training and test sets and for each feature were very similar (**Figure 1**). Since certain feature distributions were highly skewed and were made more normal by log10 transformation, this was applied to sequence length and disorder features. The alignment investigated in our study was generated by aligning our training set and calculated the associated similarity matrix via the parSeqSimDisk function from the protr R package (Xiao, Cao et al. 2015), with parameters set to batches set to 50, submat set to BLOSUM62, gap.opening set to 10, and gap.extension set to 4.

Feature Selection

Features used were outputs from a variety of sources. Structural features were obtained from using NetSurfP2 software which predicts the relative solvent accessibility (RSA), and accessible surface area (ASA), frequency of each three- and eight-state secondary structure prediction, maximum, median, and minimum disorder, and median phi/psi dihedral angles of amino acids in an amino acid sequence (Klausen,

Jespersen et al. 2019). Sequence-based features were obtained through R packages. Finally we added features output from Peptides R package functions (Osorio, Rondón-Villarreal et al. 2015): length, aliphatic index, hydrophobicity, molecular weight, net charge, alpha turn, and isoelectric point (pI). Also from the protr R package (Xiao, Cao et al. 2015) we used extractCTDC to obtain solvent accessibility features, positive, neutral, and negative charge, and polarizability. Flexibility index was obtained from the extractMoreauBroto function of protr. These features totaled to 158 (including sequence and solubility label).

Machine Learning Models

The Random Forest (RF) classifier which build multiple trees in randomly selected subspaces of the feature space (Ho, 1995). The optimal parameters used were set as ntree (number of trees used in aggregation) =2000, proximity (independent contribution of each feature to the decision-making process) = TRUE, mtry (number of randomly sampled predictors in each bootstrapping split) = 9. The mtry parameter was also tuned by using the RF tuning parameters including ntreeTry (number of trees used at tuning) = 2000, stepFactor (iteration step) = 1.5, and improve (relative improvement of out-of-bag error) = 0.01. While the samples of the dataset in the RF classification are bootstrapped, RF also mitigates the need for tree-pruning and cross-validation since the bootstrap replicates new learning sets in each randomized sampling (Breiman 2001).

We made use of an XGBoost model (Chen et al., 2016). Final parameters for our model with the following parameters: tree depth = 11, minimal node size = 7, minimum loss reduction = 0.00642, proportion of observations sampled = 0.492, number of randomly selected predictors = 3, and learn rate = 0.0348, each with trees set to 150. Parameters for our XGBoost model were arrived at via grid-based tuning. One thousand different combinations of parameters were tested using five-fold cross validation (k-fold) using the grid_latin_hypercube function in the Dials R package (Kuhn et al., 2022).

Final model evaluation of performance and analysis of its features were performed using a variety of methods. Final fitting of the model was carried out using function last_fit from the Tune R package (Kuhn, 2022), where a final fit on the entire training set is applied and then evaluated on the Chang et al. test set (Chang, Song et al. 2014).

Explainable AI

Global variable feature importance for the XGBoost model was obtained using the VIP R package (Greenwell et al, 2018). Features were ordered by importance and the top 30 were selected for plotting. Local feature importance was obtained using the DALEX and DALEXtra packages (Maksymiuk et al., 2020) for functions providing local feature importance (i.e., per sample basis). Specifically, the predict_parts function was used for computing the most important features over many possible orderings with SHAP, with the number of random orderings set to 20. Representative soluble and insoluble examples provided were selected at random. For partial dependence profiles showing how the prediction for an individual sequence changes as a function of a selected feature we plotted 500 individual profiles aggregated them using model_profile from DALEX.

Graphical User Interface

The GUI for our model was created using Shiny. This GUI uses a saved binary of the model used for inference. We restricted input to be a protein/peptide greater than 10 amino acids in length. A box- and column plot from SHAP via DALEX output were used for showing the local explainer output for the input protein sequence. If the protein was predicted to be insoluble, the GUI provides a suggested sequence-based modification (e.g., a short tag) which is be applied and then the prediction is rerun to determine whether

the evaluation has changed to be soluble. Suggestions are decided logically using both the sequence and DALEX explainer outputs.

Experimental Solubility Assessment

Our in-house set of single domain antibodies were expressed in the periplasm of E. coli BL-21(DE3) or BL-21(DE3) STAR cells using a low temperature overnight induction (25 ºC) period and 0.3 mM IPTG (isopropyl beta-D-1-thiogalactopyranoside). Culture (1.5 L) were grown at 37 ºC to an OD600 of ~1.0. Cells were washed in 30 mM Tris pH 8.0, 20% w/v sucrose, and 0.5 mM EDTA, and then pelleted at 6,000 x g for 10 min. Cells were osmotically shocked for 1 hour in 5 mM Tris pH 8.0, 1 mM EDTA containing ~30 mg of DNAse and lysozyme. The lysed cells were briefly sonicated for 60 seconds in an ice bath and Tris pH 7.6 buffer and NaCl were added to make the solution 50 mM Tris pH 7.6, 150 mM NaCl. The lysate was clarified by centrifugation (20,500 x g for 30 min at 4 ºC) and then loaded on to a nickel-charged Chelating Sepharose column equilibrated with 1x PBS pH 7.4. The column was washed with the same buffer containing 45 mM imidazole. Soluble protein was eluted using 1x PBS pH 7.4 containing 300 mM imidazole. If no protein was recovered, the protein was labeled as insoluble. To enhance solubility and lower the pI of the protein an enterokinase cleavage site sequence (DDDDK) was added to the C-terminus of the protein prior to the his-tag.

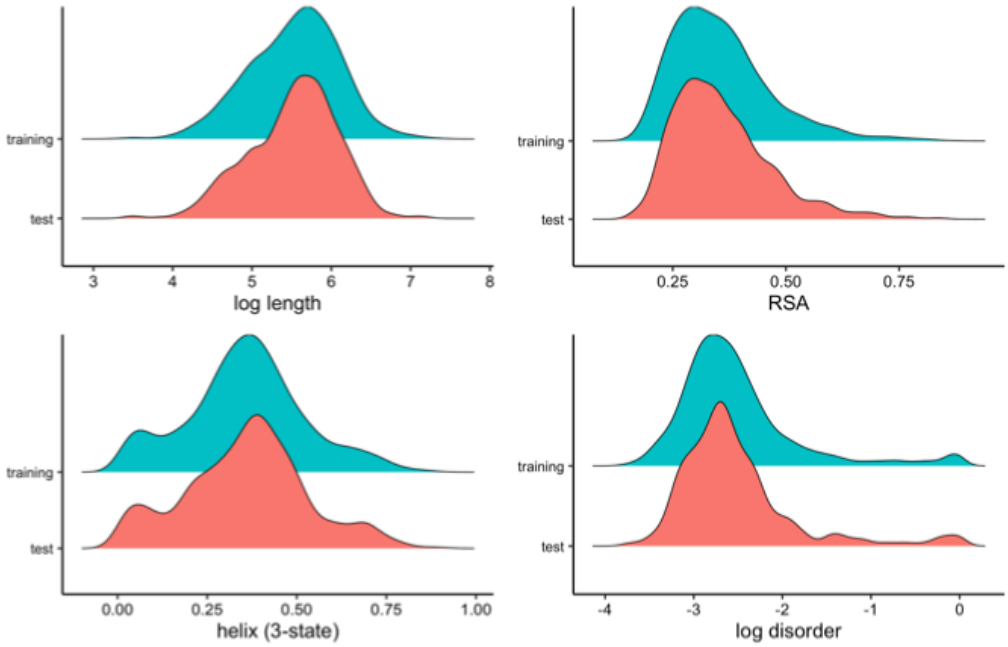

Figure 1. Distributions of features in the training and test sets.

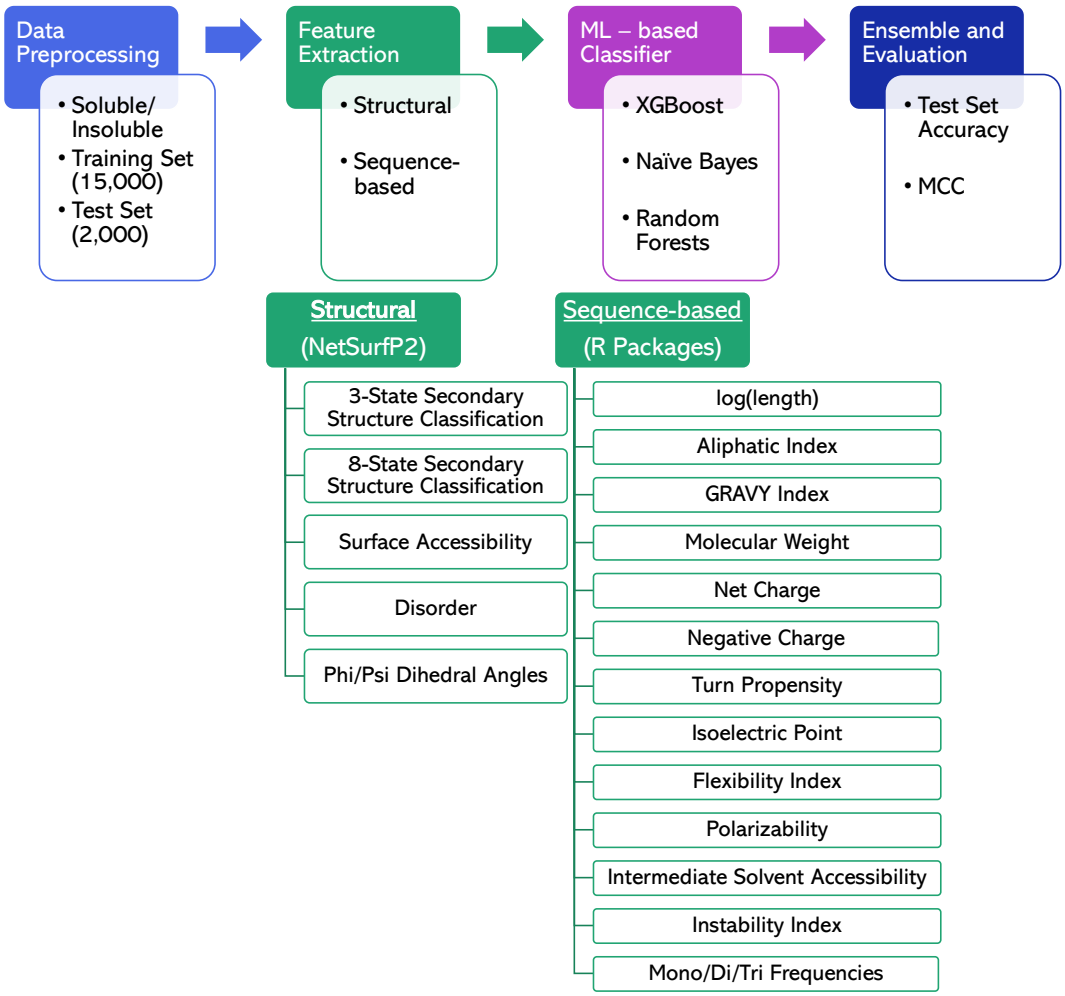

Figure 2. Dataset and model development flowchart. Sequence dataset from Rawi et al. were reduced into 15000 training set and retained the 2000 sequences in the test set. Structural and sequence-based features were obtained, including output from NetSurfP-2.0 and various R packages. XGBoost, RF, and Naïve Bayes models were trained on the compiled dataset and accuracy was determined throughout training with cross validation and finally assessed on the holdout (n = 2000) Chang et al. test set.

RESULTS

Dataset Generation and Data Exploration

Our starting dataset was obtained from *Rawi et al.,* initially >70000 sequences labeled with binary labels for soluble and insoluble (Rawi, Mall et al. 2018). A schematic for our dataset development is shown in **Figure 2**. The dataset was reduced to 15000 sequences for the training set and 2000 sequences (from the Chang *et al*. test set) held out for final model evaluation. Both structural and sequence-based features were extracted. Aside from the sequence-based features, polarizability, flexibility index, negative charge, isoelectric point, intermediate solvent accessibility, and instability index were added as additional predictors. Flexibility index was added since it was demonstrated that global structural flexibility accurately predicts the solubility of over 10000 recombinant proteins expressed in *E. coli* (Bhandari, Gardner et al. 2020). Negatively charged amino acid contributions in protein solubility were demonstrated to be the most beneficial for protein solubility (Qiao, Jiménez-Ángeles et al. 2019). Isoelectric point of proteins is strongly

influenced by the composition of amino acids, their local distribution in protein structure, as well as the structural conformation of proteins. Residues of helical folds have an intermediate solvent accessibility of ~20% and it was observed that most α-helices of protein three dimensional structures are amphipathic (Chou, Zhang et al. 1997), in which amphipathic helices have most of the hydrophobic residues oriented toward the protein core (Lins, Thomas et al. 2003). Finally, instability index was added as an extra measure of stability in vivo since it was correlated that weight values of instability for a protein of known sequence could be used as an index for predicting protein stability characteristics (Guruprasad, Reddy et al. 1990).

To investigate the dataset and possible models that could be used for solubility prediction, we looked at PCA applied to a similarity matrix on a subset of the training set had decent separation between insoluble and soluble groups (**Figure 3**). Although this separation appeared that it could be valuable in combination with a support vector machine, and possible as ensemble with our other machine learning models of interest – random forest, XGBoost, and naïve bayes – we found PCA-SVM in isolation and in combination with the other models was subpar (data not shown) so it was ruled out for the remainder of the study. In our evaluation of model performance naïve bayes appeared yielded similar accuracy to PCA-SVM at 66%, while random forest resulted in significantly higher accuracy at ~72%. Because of this performance, we further investigated the RF model including analyzing its feature importance (**Figure 4**), however we found its classification imbalance to be slightly worse than with XGBoost (see *Model performance* section), so we decided to utilize XGBoost for the remainder of the study.

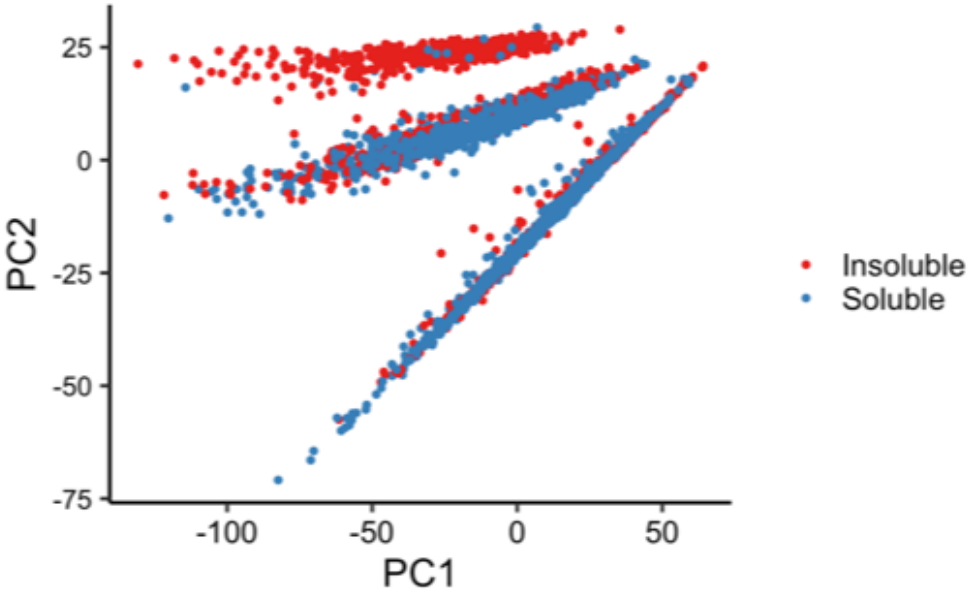

Figure 3. PCA plot of similarity matrix. Points are colored by class, insoluble or soluble.

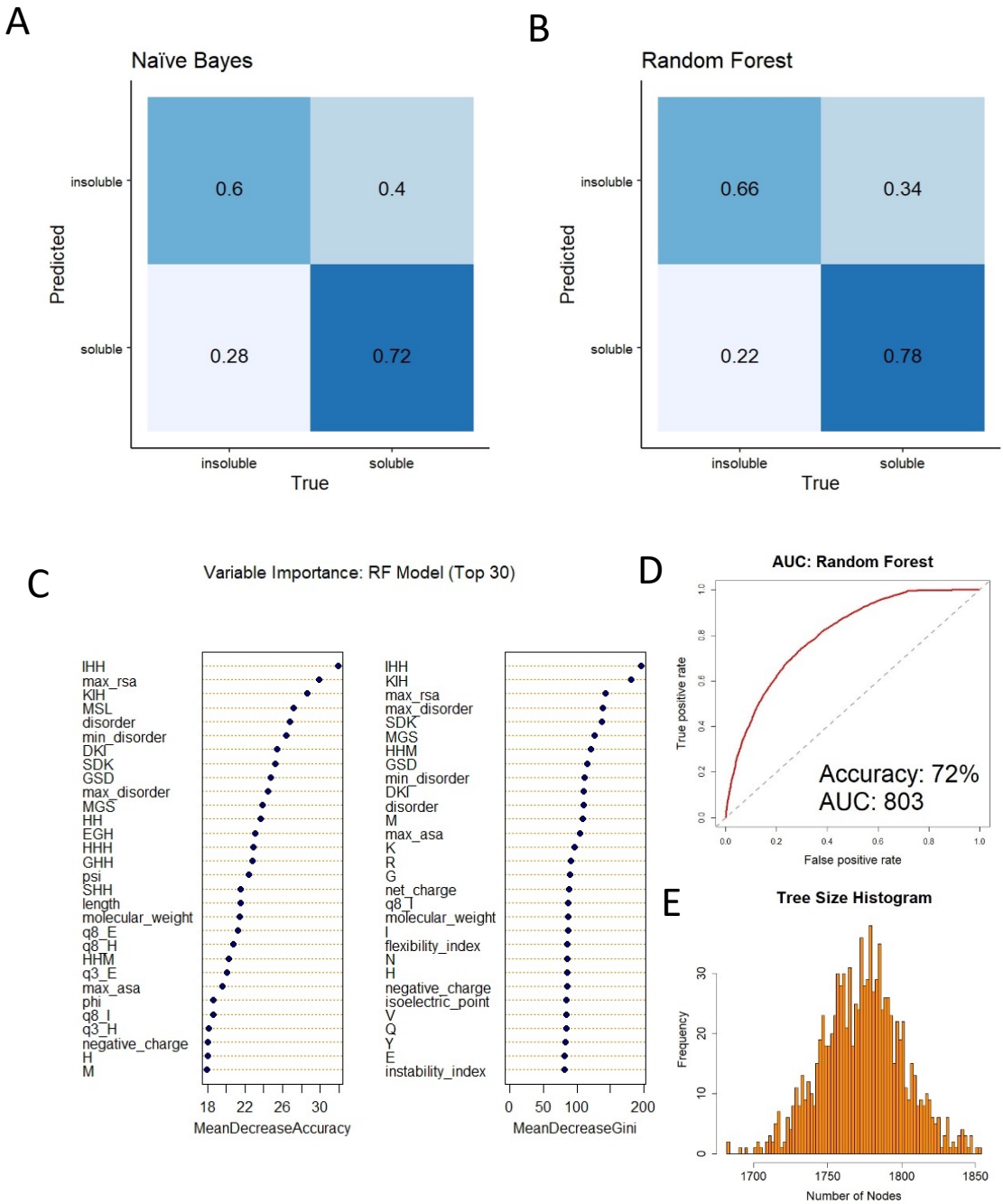

Figure 4. Evaluating other models as predictor. Both A) Naïve Bayes and B) RF models were assessed for accuracy and balance by confusion matrix. RF was further analyzed by C) variable importance and D) ROC curve. E) Shows the distribution of the number of nodes used in the model.

Training History

To train our XGBoost model using our reduced dataset (n = 15000), we split the data 9:1 into training and test sets. The model underwent optimization by tuning tree depth, min n, loss reduction, sample size,

mtry, and learn rate, each with trees set to 150 (parameter definitions found in Methods section), over 1000 different random combinations.

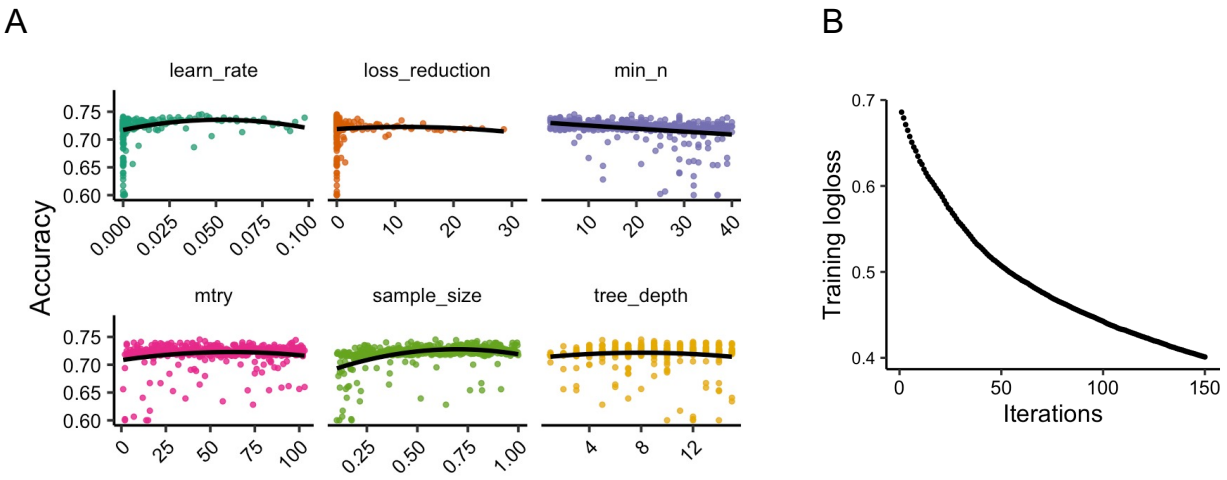

Figure 5. Training history from grid search parameter tuning. A) One thousand different random selections of six different XGBoost parameters were evaluated for accuracy, taking the mean of 5-fold cross validation steps. Each point represents an accuracy outcome. Second degree polynomial fit applied to each parameter to help visualize average trend. B) Shows the logloss during training over 150 iterations.

Results showed that our optimal parameters were tree depth = 11, min_n = 7, loss reduction = 0.00642, sample size = 0.492, mtry = 3, and learn rate = 0.0348, with trees set to 150. **Figure 5A** shows the results from all 1000 iterations of the grid search process. Although the training logloss continued to decrease beyond the point achieved by 150 iterations (**Figure 5B**), this did not lead to further improvement in final accuracy. Finally, we chose to reduce the training set to only 15000 sequences rather than the tens of thousands used by other groups in an effort to reduce the chances of overfitting the model and the computational load in creating the model.

Model Performance

Although our model showed ~80% accuracy during k-fold cross validation in training and parameter tuning above, we analyzed our model performance by using a confusion matrix and receiver operating characteristic curve applied to the Chang et al. test set (n = 2000) used as a benchmark by a variety of different protein solubility predictors (Chang, Song et al. 2014) (**Figure 6**). Here we can see the relative strengths and weaknesses of our model, where the number of true insoluble sequences in the confusion matrix has better performance than the true soluble sequences (**Figure 6A**). For the sequences predicted to be insoluble, 74% are true insoluble while 26% are incorrectly classified as soluble. Meanwhile, for the sequences predicted to be soluble, 69% are correctly classified as soluble with the remaining 31% misclassified as insoluble. This imbalance in the model performance is further highlighted by the receiver operating characteristic curve (**Figure 6B**) where the model has both a relatively high true negative rate and high false negative rate. By adjusting the two-class classification threshold from 0.5 to 0.4, we found that the sequences falsely identified as soluble are significantly reduced. The accuracy within the true soluble group falls and the overall model accuracy under this condition is also reduced from ~72% to ~70%. Similarly, the imbalanced receiver operating characteristic curve is favorably balanced, but the area under the curve is reduced from 0.81 to 0.74. While using a significantly reduced number of features and not using a novel deep learning model, but a widely used machine learning algorithm – XGBoost – we obtained similar performance to Rawi *et al.* on the Chang test set (using accuracy values from Madani *et al.*) (Rawi, Mall et al. 2018, Madani, Lin et al. 2021). The model outperforms other machine learning models, in particular SoluProt by ~4% and PROSO II by ~9%.

In addition, we compared our predictions against two webserver solubility predictors using our in-house ML-design single domain antibodies (sdAbs; **Table 1**). As the series of sdAb variants are very similar in sequence identity in certain cases, including varying by only several amino acids or by addition of a DDDDK FLAG-tag-like sequence inserted before a C-terminal His-tag, this is a potentially challenging task. Here we obtained the solubility status of 18 sdAbs and applied our model and the predictors Protein-Sol and SOLpro (Magnan, Randall et al. 2009, Hebditch, Carballo-Amador et al. 2017). For both webserver predictors, their output should be interpreted as predicted solubility upon overexpression in *E. coli*. We found that our model obtains 89% accuracy, or only two sequences classified incorrectly, while Protein-Sol and SOLpro yield 39% and 61% accuracy, respectively.

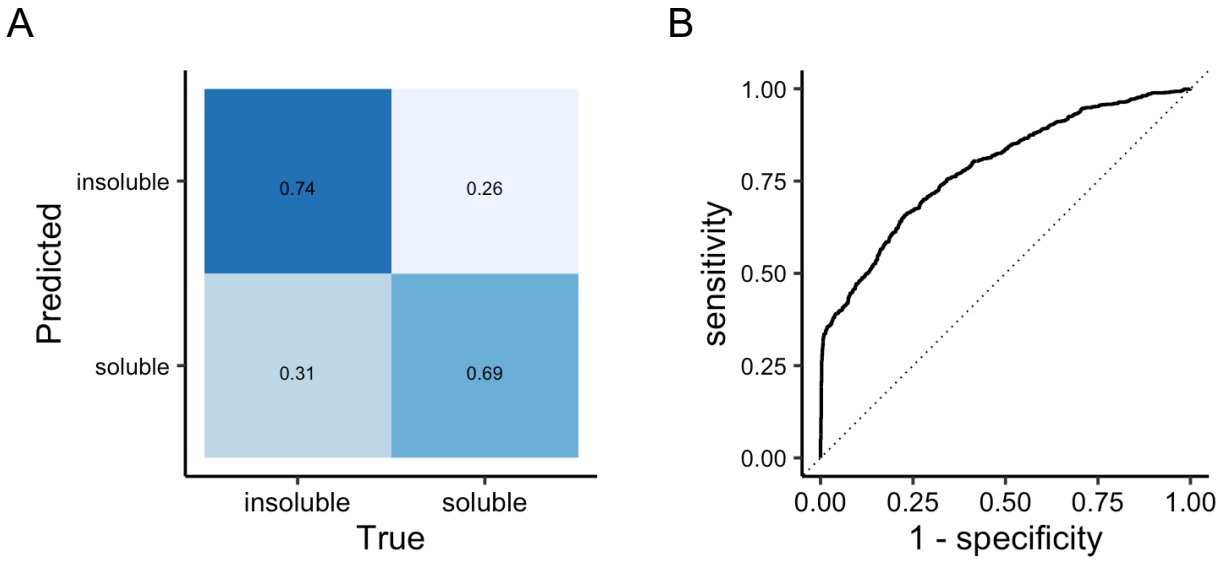

Figure 6. Final classifier performance by the model. A) Confusion matrix showing prediction performance on the test set (n = 500). B) Receiver operating characteristic curve.

Table 1. Evaluation of solubility prediction on in-house single domain antibody dataset

Protein name	Truth	This study	Protein-Sol*^	SOLpro**^
LIME0_1	1	1	0 (0.335)	1 (0.999672)
LIME0_1_DDDDK	1	1	0 (0.417)	1 (0.999705)
LIME0_2	1	1	0 (0.353)	0 (0.545842)
LIME0_2_DDDDK	1	1	0 (0.398)	1 (0.552942)
LIME0_3	0	0	0 (0.447)	0 (0.679541)
LIME1_1	1	1	0 (0.352)	1 (0.513691)
LIME1_1_DDDDK	1	1	0 (0.417)	1 (0.594717)
LIME1_2	0	1	0 (0.368)	0 (0.841232)
LIME1_2_DDDDK	1	1	0 (0.445)	0 (0.749929)
LIME1_3	0	1	0 (0.402)	0 (0.834084)
LIME17_1	1	1	0 (0.400)	0 (0.810661)
LIME17_2	1	1	0 (0.427)	0 (0.845693)
LIME17_3	0	0	1 (0.499)	0 (0.857016)
LIME18_1	1	1	0 (0.348)	1 (0.973666)
LIME18_2	1	1	0 (0.361)	0 (0.530553)
LIME18_3	0	0	1 (0.551)	0 (0.857016)
Accuracy (%)		**88**	**19**	**69**

1 if soluble; 0 if insoluble
^ values in parens are output probabilies
* 0 if lower than 0.45 and 1 if higher 0.45
** INSOLUBLE with probability of

Feature Importance

One of the main advantages of using a tree-based machine learning model over a black box deep learning method is the ability to more easily interpret the feature importance of the data used to train the model. **Figure 7A** shows the top 30 features for our model, ranked in order from most to least important. We found that top features were largely tripeptide frequencies, including those with a high number of histidines, with the most important feature being frequency of IHH (Ile-His-His), followed by KIH (Lys-Ile-His), SDK (Ser-Asp-Lys), and HHM (His-His-Met), although the patterns for the other tripeptides are difficult to discern. Other than tripeptides, attributes relating to RSA, ASA, and disorder were identified in important. Interestingly, each of the minimum, median, and maximum disorder features there derived from NetSurfP-2.0 outputs were found in the top 15 most important features. Maximum RSA (max_rsa) was highlighted as the ninth most important attribute, while none of the secondary structure predictions output by NetSurfP-2.0 were deemed important by the model, except for frequency of π-helix (q8_I). Throughout the rest of the top 30 features, mono- and dipeptides predominate, with ASA and physical or physicochemical descriptors (e.g., molecular weight [mol.wt] and negative charge) making up the remainder. Importantly only top seven attributes were given scores of over 5%, all tripeptides, with the scores tailing off quickly following this point, suggesting that the lesser importance features can be removed without greatly impacting predictive performance of the model. Other scores not shown in **Figure 7A**, and descriptions of each feature, are provided in Supplemental Information.

To further this analysis, we categorized each of the features into five different categories: mono-, di-, and tripeptides, secondary structure, or physicochemical (**Figure 7B**). Interestingly, secondary structure – those frequencies of either three- or eight-state predictions output by NetSurfP-2.0 – were deemed collectively to be unimportant for the prediction. Similarly, by either cumulative or relative importance, were the mono- and dipeptide frequencies, with dipeptides least important after accounting for the large number of features in that category. Finally, both physicochemical and tripeptide attributes were the most important categories, with tripeptides being the single most informative data category in the training set for solubility prediction – over three times more important than secondary structure features, even after correcting for the larger number of features (relative importance).

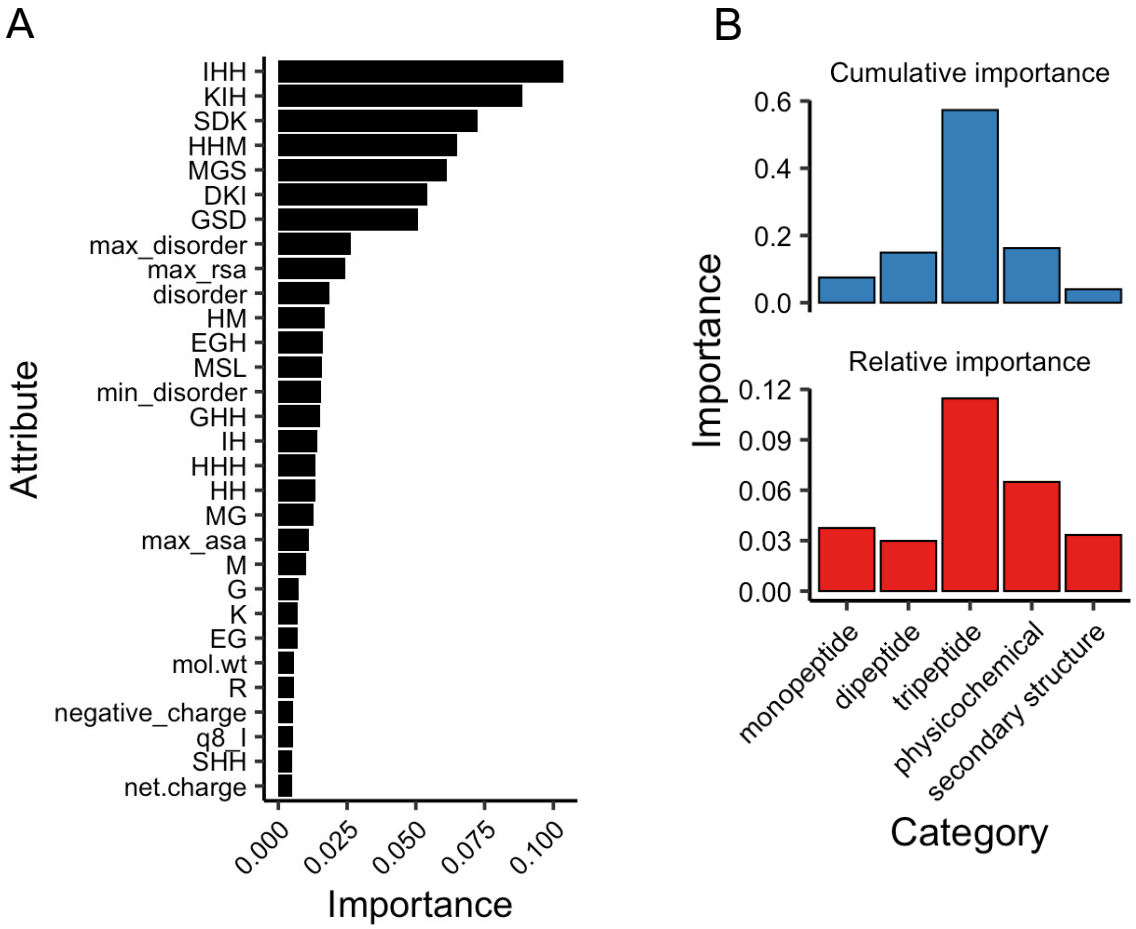

Figure 7. Feature importance by attribute and category. A) Feature importance ranked by attribute (top 30 shown). B) Feature importance after categorization. Cumulative importance is the sum of importance values; relative importance is cumulative divided by the number of features in the category.

Next we analyzed our model by the SHAP method, which unlike the feature importance above, can interrogate the importance of features to an individual prediction (i.e., why did a single sequence receive the prediction of soluble or insoluble?). We chose two representative samples to look at: **Figure 8A** shows an insoluble protein; **Figure 8B** shows a soluble protein. Both of these are correctly evaluated by the model to be insoluble and soluble, respectively. For the representative insoluble protein, we see that the majority of contributing factors with negative contributions (i.e., those that push the prediction toward insoluble) are a series of tripeptide frequencies, including IHH which was deemed by feature importance analysis above to be the most important single tripeptide. Since this protein contains an IHH, it's predicted to be insoluble, when paired with the other tripeptides listed in **Figure 8A**. Few other features positively contribute to this proteins prediction: since it has few lysines, it receives a small increase in likelihood to be soluble, but this and other positive contributions do not outweigh the negatives. In contrast, the soluble representative sample has largely positive contributors, and higher in magnitude ones relative to the previous example (see x-axis; **Figure 8B**). Here, since the soluble protein does not contain the tripeptide IHH this contributes positively toward it being predicted as soluble. Interestingly, in other examples with His-tags and the resulting HHH tripeptide, this feature contributes positively toward a soluble prediction, unlike the negative presence of IHH in the insoluble example. Similarly because the maximum RSA is 0.96 this protein receives a positive contribution, in contrast to the insoluble sample which had a maximum RSA of 0.90, which is deemed to be too low by the model and thus provides a negative contribution.

This fine-line difference in the value of certain features is highlighted by the partial dependence profiles of maximum disorder and maximum RSA aggregating local model explanations shown in **Figure 8C**. Here the small difference in disorder and RSA can have an effect on the probability of classification as soluble, where in particular a maximum disorder approaching 0.0 (note disorder here is $\log_{10}$ transformed) will greatly decrease the likelihood that the protein is predicted to be soluble.

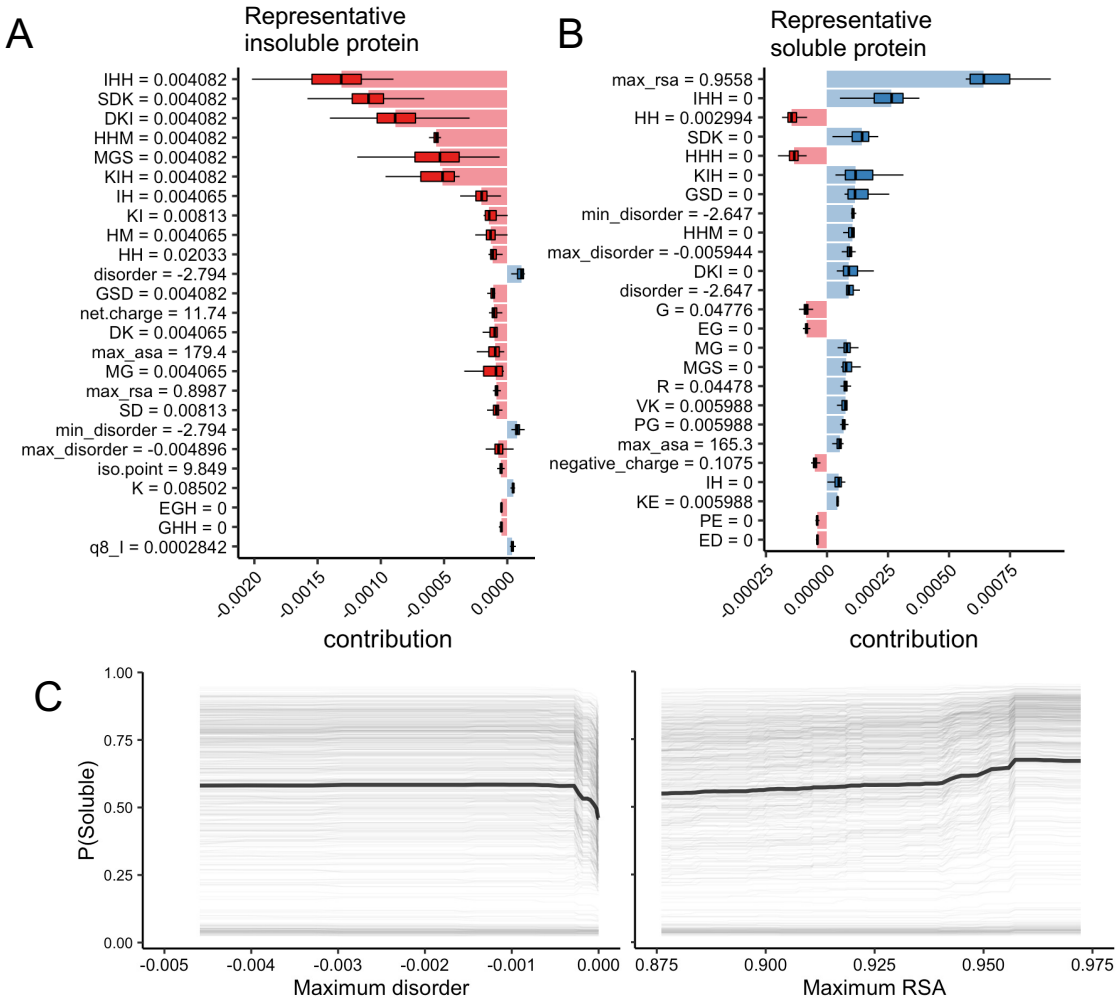

Figure 8. Local explanations of the predictor. SHAP feature ranking according to contribution of individual features for representative insoluble proteins (A) and representative soluble protein (B). Red colored attributes that negatively contribute toward the prediction (toward insoluble); attributes colored blue positively contribute. Boxplots indicate permutations of feature orderings, while the shaded bar plots highlight the mean contribution found for those permutations. C) Partial dependence profiles of maximum disorder and maximum RSA aggregating local model explanations, indicating probability of solubility on the y-axis.

Graphical User Interface

Finally, to increase the number of potential users of our solubility prediction model and explainer output, we produced a simple graphical user interface (see **Figure 9** for a representative screenshot). Made using Shiny, the basic workflow is typing or pasting a protein sequence into the upper-right text box and pressing Run. Following prediction and application of the SHAP explainer, the prediction of either soluble or insoluble is output in the panel on the right side. Along with the prediction is a plot displaying the SHAP output via the DALEX R package. Finally, we supply functionality that uses the local explainer output to suggest a modification to the protein, if the initial prediction is insoluble. Modifications include addition of various protein tags (if not already present in the input protein sequence) and other simple additions or

mutations that could be carried out via standard cloning or protein engineering techniques. For example, as described in *Feature importance* section His-tags and the associated HHH tripeptide can positively contribute to a soluble prediction, although related tripeptides such as IHH have a strong negative contribution requiring modifications to be evaluated on a case-by-case basis.

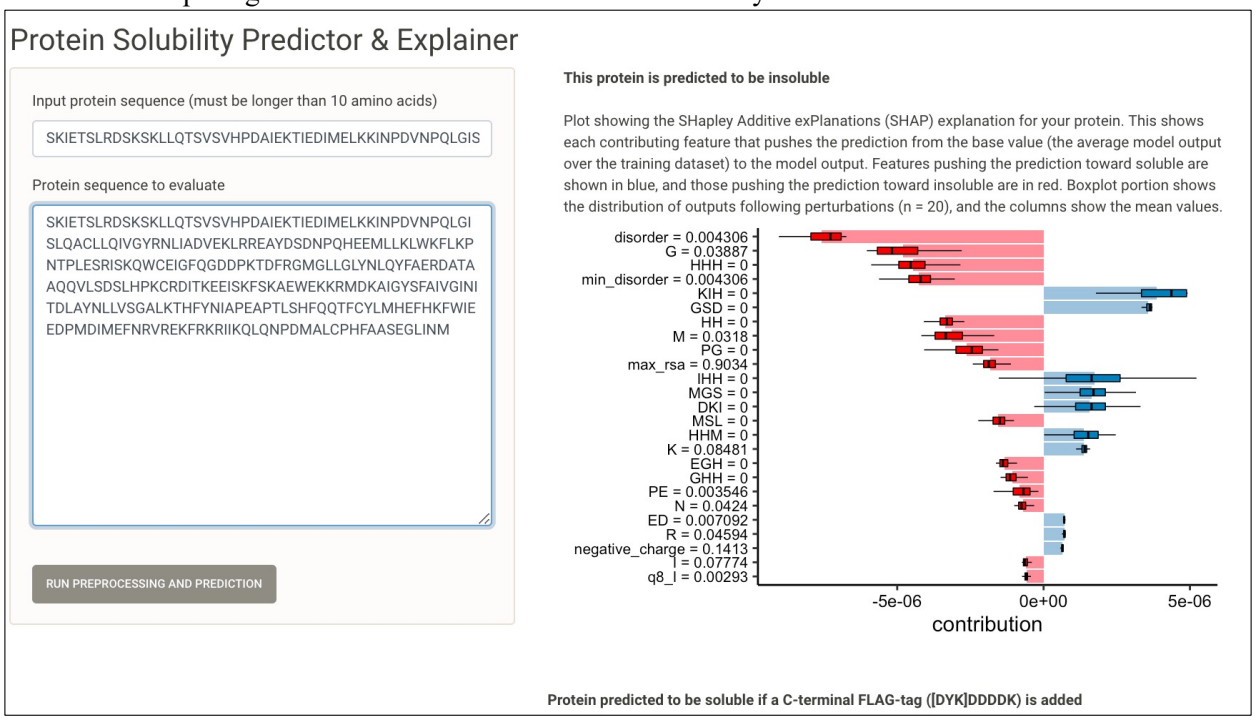

Figure 9. Graphical user interface screenshot of protein solubility predictor and explainer. The protein sequence is input into the text box on the upper left. Sequences must be longer than 10 amino acids. After the protein sequence is shown in the lower text box, the Run Preprocessing and Prediction box can be pressed. After the protein sequence is analyzed, the information on the right is shown. The prediction is given at the upper right, with the SHAP explainer plot below that. (Feature definitions are provided in a different tab of the GUI; e.g., q8_I indicates frequency of π-helix.) If the protein is predicted to be insoluble, the GUI outputs a chosen modification and second prediction based on the information provided by the SHAP explainer (lower right).

DISCUSSION

Despite at least two decades of efforts by various research groups taking different approaches to the problem, the creation of a highly accurate sequence-based protein solubility predictor is still a barrier to successful protein design and engineering experiments. In this report we describe a XGBoost-based model for prediction of solubility based on information derived directly from the protein sequence. We found our model accuracy of 72% to be competitive with the state-of-the-art machine learning models although less accurate than far more complex and computationally expensive deep learning models. Importantly, the model run time (inference) takes under one minute to complete a prediction and provide analysis via explainable AI techniques, including variable importance and the recently developed SHAP, meaning the solubility predictor is not a black box and its results are easily interpretable by humans, a task our tree-based model is well-positioned to perform.

The level of performance of our model is related to several different elements, including the usage of XGBoost, which is among the top performing widely used machine learning models for a wide variety of problems. We also make use of a variety of features focused on attributes that differ between the insoluble and soluble groups. Following training this, we underwent intensive grid search model tuning, using cross validation which increased performance (accuracy) from ~70% to ~80% on the test split, taking the final

performance from mediocre although still valuable, to similar to PaRSnIP with final evaluation on a common test set (Chang, Song et al. 2014, Rawi, Mall et al. 2018), which achieved 72% accuracy (Madani, Lin et al. 2021). Similarly, using Matthews correlation coefficient (MCC), our model achieves a high level of performance when compared to other machine learning model-based predictors. Notably, many machine learning-based predictors perform poorly on the Chang *et al.* test set benchmark (Madani, Lin et al. 2021). And, related to the *Chang et al.* benchmark, the lack of labeled data and the expense in obtaining more data and its labeling forces nearly all published studies to train and evaluate on a handful of datasets, e.g., Chang et al. and the NESG test set (Hon, Marusiak et al. 2021).

Another element differentiating our model is the inclusion of NetSurfP model output. NetSurfP-2.0 (Klausen, Jespersen et al. 2019) provides structural predictions based on sequence, including 8-state secondary structure and solvent accessibility which may be superior to other models used for similar feature accumulation in other solubility predictors. Additionally, from NetSurfP, we included other sequence-derived features like protein disorder metrics not used in some other systems, as well as simple physicochemical descriptors, some of which were ranked highly globally in feature importance, most notably disorder, RSA, and charge. Negative charge, specifically, was ranked in the model's top 30 features, and has previously been identified as correlated with solubility when located on the protein surface (Kramer, Shende et al. 2012).

We can be confident in the importance of these features since we made extensive use of explainable AI techniques: variable importance and SHAP. Each ranking, according to different systems, the importance or contribution of each feature to the final solubility prediction. Critically, we were able to significantly prune our training set using both global and local variable importance rankings such that our starting dataset of >150 features was reduced to 75 features. This starting level of features was already reduced from the thousands (>8000) of features originally evaluated by Rawi *et al*., Wu *et al*., and others further refining which attributes of the sequence descriptors was necessary for solubility prediction. This should lessen the computational load of using our predictor relative to others as few features need to be computed for new inputs.

We also introduce a GUI that both predicts the input protein solubility, but provides feature importance on a local level using SHAP via DALEX and delivers a suggestion based on the local feature importance information to potentially assist in improving the solubility of the input. Specifically, certain protein tags like His- and FLAG-tags are associated with soluble proteins, as the HHH and DDK tripeptides are more common in that group of sequences than in the insoluble group, sufficient enough of a frequency difference that the model identified that as important features. The addition of certain tags, including FLAG-tags, are listed in protocols for enhancing solubility and aiding purification (Walls and Loughran 2011); however, these need to be treated on a case-by-case basis as certain tripeptides such as IHH are found by the model to strong negative contributors to insoluble prediction and could be introduced by adding a His-tag, for example, whereas the highly charged/polar residues of the enterokinase cleavage site DDDDK and other tags may enhance solubility more generally with fewer negatively contributing di- or tripeptides. Constructs with pI values < 9 appeared to be more favorably expressed in the periplasmic space which is oxidizing.

Although our model has some disadvantages, including misclassifying approximately a quarter of sequences in the widely used benchmark Chang *et al*. test set, similar to others in the protein solubility prediction space, our model accuracy when paired with its prediction explanations should enhance the ability to experimentally produce and investigate proteins designed by machine learning-directed approaches and other protein production purposes by alleviating wet-lab work related to solubility issues.

ACKNOWLEDGEMENTS

We acknowledge funding support through base funds of the Naval Research Laboratory (WU# 1V33) and funds from the Defense Threat Reduction Agency (HDTRA1033536).

REFERENCES

Bhandari, B. K., P. P. Gardner and C. S. Lim (2020). "Solubility-Weighted Index: fast and accurate prediction of protein solubility." Bioinformatics **36**(18): 4691-4698.
Chang, C. C., J. Song, B. T. Tey and R. N. Ramanan (2014). "Bioinformatics approaches for improved recombinant protein production in Escherichia coli: protein solubility prediction." Brief Bioinform **15**(6): 953-962.
Chen, J., S. Zheng, H. Zhao and Y. Yang (2021). "Structure-aware protein solubility prediction from sequence through graph convolutional network and predicted contact map." J Cheminform **13**(1): 7.
Cheng, J., A. Z. Randall, M. J. Sweredoski and P. Baldi (2005). "SCRATCH: a protein structure and structural feature prediction server." Nucleic Acids Res **33**(Web Server issue): W72-76.
Chou, K. C., C. T. Zhang and G. M. Maggiora (1997). "Disposition of amphiphilic helices in heteropolar environments." Proteins: Structure, Function, and Bioinformatics **28**(1): 99-108.
Francis, D. M. and R. Page (2010). "Strategies to optimize protein expression in E. coli." Curr Protoc Protein Sci **Chapter 5**: Unit 5 24 21-29.
Guruprasad, K., B. B. Reddy and M. W. Pandit (1990). "Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence." Protein Engineering, Design and Selection **4**(2): 155-161.
Hebditch, M., M. A. Carballo-Amador, S. Charonis, R. Curtis and J. Warwicker (2017). "Protein–Sol: a web tool for predicting protein solubility from sequence." Bioinformatics **33**(19): 3098-3100.
Herbert, B. (1999). "Advances in protein solubilisation for two-dimensional electrophoresis." ELECTROPHORESIS: An International Journal **20**(4-5): 660-663.
Hon, J., M. Marusiak, T. Martinek, A. Kunka, J. Zendulka, D. Bednar and J. Damborsky (2021). "SoluProt: Prediction of Soluble Protein Expression in Escherichia coli." Bioinformatics.
Hou, Q., R. Bourgeas, F. Pucci and M. Rooman (2018). "Computational analysis of the amino acid interactions that promote or decrease protein solubility." Scientific reports **8**(1): 1-13.
Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek and A. Potapenko (2021). "Highly accurate protein structure prediction with AlphaFold." Nature **596**(7873): 583-589.
Khurana, S., R. Rawi, K. Kunji, G. Y. Chuang, H. Bensmail and R. Mall (2018). "DeepSol: a deep learning framework for sequence-based protein solubility prediction." Bioinformatics **34**(15): 2605-2613.
Klausen, M. S., M. C. Jespersen, H. Nielsen, K. K. Jensen, V. I. Jurtz, C. K. Sonderby, M. O. A. Sommer, O. Winther, M. Nielsen, B. Petersen and P. Marcatili (2019). "NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning." Proteins **87**(6): 520-527.
Kramer, R. M., V. R. Shende, N. Motl, C. N. Pace and J. M. Scholtz (2012). "Toward a molecular understanding of protein solubility: increased negative surface charge correlates with increased solubility." Biophysical journal **102**(8): 1907-1915.
Lins, L., A. Thomas and R. Brasseur (2003). "Analysis of accessible surface of residues in proteins." Protein science **12**(7): 1406-1417.
Madani, M., K. Lin and A. Tarakanova (2021). "DSResSol: A Sequence-Based Solubility Predictor Created with Dilated Squeeze Excitation Residual Networks." Int J Mol Sci **22**(24).
Magnan, C. N., A. Randall and P. Baldi (2009). "SOLpro: accurate sequence-based prediction of protein solubility." Bioinformatics **25**(17): 2200-2207.
Qiao, B., F. Jiménez-Ángeles, T. D. Nguyen and M. O. De La Cruz (2019). "Water follows polar and nonpolar protein surface domains." Proceedings of the National Academy of Sciences **116**(39): 19274-19281.
Rawi, R., R. Mall, K. Kunji, C. H. Shen, P. D. Kwong and G. Y. Chuang (2018). "PaRSnIP: sequence-based protein solubility prediction using gradient boosting machine." Bioinformatics **34**(7): 1092-1098.
Smialowski, P., G. Doose, P. Torkler, S. Kaufmann and D. Frishman (2012). "PROSO II--a new method for protein solubility prediction." FEBS J **279**(12): 2192-2200.

Walls, D. and S. T. Loughran (2011). "Tagging recombinant proteins to enhance solubility and aid purification." Protein Chromatography: 151-175.
Wu, X. and L. Yu (2021). "EPSOL: sequence-based protein solubility prediction using multidimensional embedding." Bioinformatics.
Yang, J., R. Yan, A. Roy, D. Xu, J. Poisson and Y. Zhang (2015). "The I-TASSER Suite: protein structure and function prediction." Nature methods **12**(1): 7-8.