Motivation: Poor protein solubility hinders the production of many therapeutic and industrially useful proteins. Experimental efforts to increase solubility are plagued by low success rates and often reduce biological activity. Computational prediction of protein expressibility and solubility in Escherichia coli using only sequence information could reduce the cost of experimental studies by enabling prioritisation of highly soluble proteins. Results: A new tool for sequence-based prediction of soluble protein expression in Escherichia coli, SoluProt, was created using the gradient boosting machine technique with the TargetTrack database as a training set. When evaluated against a balanced independent test set derived from the NESG database, SoluProt’s accuracy of 58.4% and AUC of 0.60 exceeded those of a suite of alternative solubility prediction tools. There is also evidence that it could significantly increase the success rate of experimental protein studies. SoluProt is freely available as a standalone program and a user-friendly webserver at https://loschmidt.chemi.muni.cz/soluprot/. Availability and Implementation: https://loschmidt.chemi.muni.cz/soluprot/ Contact: jiri@chemi.muni.cz Supplementary Information: Supplementary data are available at Bioinformatics online

File list (2)

| File name                              | Size       | View Options | Download Options |
|----------------------------------------|------------|--------------|------------------|
| Hon_Bioinformatics.docx (301.98 KiB)   | view on ChemRxiv | download file |
| Hon_Bioinformatics_SI.docx (101.40 KiB)| view on ChemRxiv | download file |
SoluProt: Prediction of Soluble Protein Expression in *Escherichia coli*

Jiri Hon\(^1,2,3\), Martin Marusiak\(^3\), Tomas Martinek\(^3\), Antonín Kunka\(^1,2\), Jaroslav Zendulka\(^3\), David Bednar\(^1,2,*\), Jiri Damborsky\(^1,2,*\)

\(^1\)Loschmidt Laboratories, Centre for Toxic Compounds in the Environment RECETOX and Department of Experimental Biology, Faculty of Science, Masaryk University, 625 00 Brno, Czech Republic; \(^2\)International Clinical Research Center, St. Anne’s University Hospital Brno, 656 91 Brno, Czech Republic; \(^3\)IT4Innovations Centre of Excellence, Faculty of Information Technology, Brno University of Technology, 612 66 Brno, Czech Republic; *authors for correspondence

Abstract

**Motivation:** Poor protein solubility hinders the production of many therapeutic and industrially useful proteins. Experimental efforts to increase solubility are plagued by low success rates and often reduce biological activity. Computational prediction of protein expressibility and solubility in *Escherichia coli* using only sequence information could reduce the cost of experimental studies by enabling prioritisation of highly soluble proteins.

**Results:** A new tool for sequence-based prediction of soluble protein expression in *Escherichia coli*, SoluProt, was created using the gradient boosting machine technique with the TargetTrack database as a training set. When evaluated against a balanced independent test set derived from the NESG database, SoluProt's accuracy of 58.4% and AUC of 0.60 exceeded those of a suite of alternative solubility prediction tools. There is also evidence that it could significantly increase the success rate of experimental protein studies. SoluProt is freely available as a standalone program and a user-friendly webserver at [https://loschmidt.chemi.muni.cz/soluprot/](https://loschmidt.chemi.muni.cz/soluprot/).

Availability and Implementation: [https://loschmidt.chemi.muni.cz/soluprot/](https://loschmidt.chemi.muni.cz/soluprot/)

Contact: jiri@chemi.muni.cz

Supplementary Information: Supplementary data are available at Bioinformatics online

**Keywords**

protein solubility, soluble expression, prediction, machine-learning, protein mining
Introduction

Low protein solubility causes severe problems in protein science and industry; insufficient protein solubility is probably the most common cause of failure of protein production pipelines. The importance of solubility is underlined by the findings of the large-scale Protein Structure Initiative (PSI) project (Berman et al., 2017), which sought to produce thousands of protein sequences from different organisms, crystallise them, and resolve their tertiary structure. Unfortunately, in most cases it proved impossible to produce the target proteins in soluble form. The inherent low solubility of natural enzymes also limits the success of emerging high-throughput pipelines that explore protein databases to identify novel enzymes with diverse functions (Vanacek et al., 2018; Hon et al., 2020). Given the rapid growth of protein sequence databases driven by the capabilities of next-generation sequencing technologies, there is an urgent need to focus only on potentially soluble targets to avoid wasting resources on hard-to-produce orthologs. Solubility is thus a key attribute when choosing protein targets for experimental characterisation (Vanacek et al., 2018). Strictly speaking, solubility is a thermodynamic parameter defined as the protein's concentration in a saturated solution in equilibrium with a solid phase under specific conditions. However, it is challenging to quantitatively measure the solubility of large sets of proteins (Kramer et al., 2012), so there is little quantitative experimental data on protein solubility. Moreover, this definition of solubility is too narrow to encompass many of the practical problems that may occur during protein production with common expression systems. Therefore, inspired by existing literature (Magnan et al., 2009; Smialowski et al., 2012; Agostini et al., 2012; Khurana et al., 2018; Raimondi et al., 2020), available data (Berman et al., 2017), and laboratory practice, we use an extended definition of protein solubility in this work. Specifically, by solubility, we mean the probability of soluble protein (over)expression in *Escherichia coli* cells. We thus consider a protein soluble only if it is both (over)expressed and thermodynamically soluble in a given expression system.

Solubility depends on many extrinsic and intrinsic factors. Extrinsic factors are dictated by the choice of expression system and the experimental conditions used in protein production. Expression systems may be either *in vivo* or *in vitro* (Rosano and Ceccarelli, 2014; Carlson et al., 2012). *In vivo* protein expression is induced inside living cells of a host organism, whereas *in vitro* expression relies on the use of cell-free translational systems. Solubility can be increased by adjusting extrinsic solubility factors, especially by using different mutated host strains, codon optimization, coexpression of chaperones and foldases, lowering cultivation temperatures, and adding suitable fusion partners (Costa et al., 2014). However, tuning the expression system or experimental conditions is not always sufficient to confer solubility, and is not feasible in high-throughput protein production pipelines. If extrinsic factors cannot be varied, protein solubility will depend only on the intrinsic properties of the protein sequence. Unfortunately, the relationship between a protein’s sequence and its solubility is poorly understood, mainly due to a lack of reproducible quantitative solubility measurements (Kramer et al., 2012). Recent protein engineering studies suggest that charged amino acids on the protein surface are key intrinsic determinants of solubility (Carballo-Amador et al., 2019; Sankar et al., 2018; Chan et al., 2013). However, this knowledge cannot be directly used for solubility prediction due to a lack of structural data. Despite the continuous growth of structural databases (Burley et al., 2019), the structures of proteins of interest are generally unknown, and the limited availability of template structures prevents their accurate computational prediction.

The simultaneous effects of extrinsic and intrinsic factors make solubility prediction challenging. For example, the prediction of solubility from sequence data using machine learning is hindered by the high level of noise in typical training data sets due to the influence of diverse extrinsic variables. Because the molecular mechanisms governing protein solubility are poorly understood, recent solubility prediction tools rely heavily on statistical analysis and machine learning, using previously reported experimental data to train and validate model parameters. One of the most widely used data sources is the
TargetTrack database (Berman et al., 2017), formerly known as PepcDB or TargetDB, which integrates information from the Protein Structure Initiative projects. This database contains data from over 900,000 protein crystallization trials involving almost 300,000 unique protein sequences, which are referred to as targets. The database does not contain solubility data per se, but target proteins can be considered soluble if they were successfully purified in the experimental trials. A major limitation of this database is the low quality of its annotations. For example, reasons for failure are generally not provided for unsuccessful crystallization attempts. Therefore, it is impossible to distinguish failures due to insolubility from failures due to other problems later in the experimental pipeline. Second, the experimental protocols used for protein production and crystallization are described in free text with no internal structure, making it hard to automatically extract information about experimental conditions and expression systems for a given target. Filtering is therefore needed to reduce noise before using the TargetTrack data for model training. However, the application of stringent filtering rules to the target annotations can dramatically reduce the number of usable records.

eSOL is another well-known and commonly used solubility database (Niwa et al., 2009, 2012) that contains experimentally measured solubilities for over 4000 E. coli proteins produced in the PURE (Shimizu et al., 2001) cell-free expression system. eSOL is an impressive collection of highly homogenous data but has its own limitations. First, it only contains data on proteins originating from E.coli. Second, it has relatively little negative data; adding the three main cytosolic E. coli chaperones (TF, DnaKJE, and GroEL/GroES) to the PURE expression system reduced the number of insoluble proteins from 788 to 24 (Niwa et al., 2012). eSOL is a valuable source of exact solubility data that were generated using a robust pipeline and provide a good quantitative measure of thermodynamic solubility. However, these data cannot be used to assess solubility according to our expanded definition, which also encompasses expressibility.

The relationship between protein sequence and solubility has been studied for over 30 years, leading to the development of several predictive models and software tools. There are 11 such models or tools that use definitions of solubility like that described above and take protein sequences as their sole input. These are the revised Wilkinson-Harrison model (rWH) (Wilkinson and Harrison, 1991; Davis et al., 1999), SOLpro (Magnan et al., 2009), RPSP (Diaz et al., 2010), PROSO II (Smialowski et al., 2012), ccSOL omics (Agostini et al., 2012, 2014), ESPRESSO (Hirose and Noguchi, 2013), CamSol (Sormanni et al., 2015), Protein-Sol (Hebditch et al., 2017), DeepSol (Khurana et al., 2018), SKADE (Raimondi et al., 2020), and the Solubility-weighted index (SWI) (Bhandari et al., 2020). However, the accuracy of these tools is limited, and there is clear room for improvement. Additionally, these tools exhibit poor generality when used to make predictions based on previously unseen data. A comprehensive review of advances in solubility prediction, including predictors that use protein structures as inputs, was published recently (Musil et al., 2019). Here, we present a novel machine learning-based tool, SoluProt, for predicting solubility from protein sequence data. SoluProt benefits from thorough dataset pre-processing and is shown to predict solubility more accurately than previously reported methods.

SoluProt training and test set

We used the TargetTrack database to build the SoluProt training set. Since this database does not directly provide solubility information, we inferred solubility computationally, using an approach similar to those adopted previously (Magnan et al., 2009; Smialowski et al., 2012). A protein was considered soluble if it was recorded as having reached a soluble experimental state or any subsequent state requiring soluble expression (Table S1). If failed expression or purification was mentioned in the experiment record’s stop status, the protein was labelled insoluble. In contrast to a previous approach (Smialowski et al., 2012), we required an explicit stop status relating to insolubility to reduce the
frequency of incorrect classification of insoluble sequences. To improve the quality of the training set, we also performed several additional steps to clean the data.

Most importantly, we performed keyword matching combined with manual checking of TargetTrack annotations to extract only proteins expressed in the most common host organism, *E. coli*. This was necessary because a protein soluble in one organism might be insoluble in another. By focusing solely on the most common expression system, we reduced the noise in the training data. We also used specific keywords to search the unstructured descriptions of experimental protocols provided in the TargetTrack database (Table S2). Generic search phrases like “*E. coli*” or “Escherichia coli” were used to identify potential *E. coli* related protocols. These protocols were then manually checked and confirmed (Table S3). A full list of 1494 TargetTrack protocols signifying expression in *E. coli* is available at the SoluProt website.

We next identified transmembrane proteins in the dataset based on direct annotations from the TargetTrack database and predictions generated using TMHMM (Krogh et al., 2001) with default settings. The transmembrane proteins were then removed, along with sequences shorter than 20 amino acids, and sequences with undefined residues. We also removed sequences that had been classified as insoluble but for which a protein structure was available in the Protein Data Bank (PDB) (Berman et al., 2000). To this end, we compiled an *E. coli* PDB subset containing sequences of proteins whose structures had been solved by NMR or X-ray crystallography and which had been expressed in *E. coli* according to the PDB annotations (64,416 sequences, downloaded 2018-04-17). Because both NMR and X-ray crystallography require soluble proteins, any protein in this PDB subset can be considered soluble in *E. coli*. This step reflects advances in molecular biology: methodological developments have made it possible to produce and crystallize some proteins that were previously considered insoluble.

Finally, we reduced the sequence redundancy in the training set by clustering to 25% identity using MMseqs2 (Steinegger and Söding, 2017) and retaining only representative sequences from each cluster. This was done separately for positive and negative samples to avoid simplifying the prediction problem. We balanced the number of soluble and insoluble samples such that both classes were equally represented. Additionally, we balanced the sequence length distribution so that length alone would not play a dominant role in the predictions. Sequence length correlates with protein solubility – larger proteins are usually less soluble. However, we wanted to suppress its influence in the model because we anticipate that SoluProt would mainly be used to prioritize proteins of similar lengths, usually from a single protein family. A typical expected use case is that of the EnzymeMiner web server (Hon et al., 2020) for automated mining of soluble enzymes. A prediction model relying heavily on sequence length would not perform well in this use case.

The SoluProt test set was built from a dataset generated by the North East Structural Consortium (NESG), which represents 9,644 proteins expressed in *E. coli* using a unified production pipeline (Price et al., 2011). The dataset contains two integer scores ranging from 0 to 5 for each target, indicating the protein’s level of expression and the soluble fraction recovery. The reproducibility of the experimental results in the dataset was validated by performing repeat measurements for selected targets. The NESG dataset targets are actually included in the TargetTrack database because the NESG participated in the PSI project. However, the expression and solubility levels from the NESG dataset were not included in the TargetTrack database; instead, they were provided to us directly by the authors of the original study (W. Nicholson Price II, personal communication). The high consistency and quality of the NESG dataset make it suitable for benchmarking purposes. We processed the NESG dataset using the same procedure as the training set, although the computational solubility derivation and expression system filtration steps were omitted because they were pointless in this case. Instead, we transformed the solubility levels into binary classes: all proteins with a solubility level of 1 or above were considered soluble and all others insoluble.
Finally, we ensured that no pair consisting of a sequence from the test set and a sequence from the training set had a global sequence identity above 25% as calculated using the USEARCH software (Edgar, 2010). This made the test set more independent because it ensured that predictions were not validated against data similar to those used during training. In total, 10,912 protein sequences remained in the SoluProt training set and 2,904 in the independent SoluProt test set. Both datasets had equal numbers of soluble and insoluble samples with balanced sequence length distributions (Figure S1). The datasets are available at the SoluProt website.

Prediction model

The SoluProt predictor is implemented in Python using scikit-learn (Pedregosa et al., 2011), Biopython (Cock et al., 2009), and pandas (McKinney, 2010) libraries. We used a gradient boosting machine (GBM) (Friedman, 1999) to generate the predictive model. Prediction features were selected from a set of 251 sequence characteristics that were divided into eight groups: i) single amino acid content (20 features), ii) amino acid dimer content (210 features), iii) sequence physicochemical features (12 features, Table S4), iv) average flexibility as computed by DynaMine (Cilia et al., 2014) (1 feature), v) secondary structure content as predicted by FESS (Piovesan et al., 2017) (3 features), vi) average disorder as predicted by ESPRITZ (Walsh et al., 2012) (1 feature), vii) content of amino acids in transmembrane helices as predicted by TMHMM (3 features), and viii) maximum identity to the E. coli PDB subset as calculated using USEARCH (1 feature). All sequences having 100% identity to any sequence from the E. coli PDB subset were excluded from the test set because we wanted to assess the model’s predictive accuracy for sequences without a solved NMR or X-ray structure. We standardized all features by subtracting the mean and scaling to unit variance. The means and variances were calculated using the training set.

We removed correlated features in two steps. First, we fitted a GBM with default parameters using the full training set and all features. Second, we calculated Pearson’s correlation coefficient for each pair of features. If the correlation between any two features exceeded 0.75, we removed the feature with the lesser importance in the fitted GBM model. We also removed irrelevant features using LASSO (Tibshirani, 1996). LASSO’s alpha parameter was optimized to maximize the mean AUC of the GBM model with default parameters over 5-fold cross-validation. The alpha parameter was varied between 0.08 to 0 with a step size of 6.25 \(10^{-4}\); its optimal value was 0.005. In total, 96 features were selected for inclusion in the predictive model (Table S5). The DynaMine, FESS, and ESPRITZ features were not included in the final feature set.

We next optimized the hyperparameters of the GBM model, using an iterative 7-stage strategy to maximize the mean AUC over 5-fold cross-validation using the training set (Table S6). In each stage, one or two parameters were optimized using grid search; other parameters were left either at their final values from the previous stages or at the default value if the parameter had not yet been optimized. The best GBM model achieved mean AUC values of 0.84 ± 0.003 for the training part and 0.72 ± 0.02 for the validation part. Overall, the feature selection and hyperparameter optimization had little effect on the mean AUC: without these measures, the mean AUC values for the training and validation sets were 0.83 ± 0.002 and 0.71 ± 0.02, respectively. The main benefit of the feature selection and parameter tuning steps was that they reduced the number of features and thus made the feature calculation step roughly two times faster.

Finally, we used the best GBM hyperparameters to train the final SoluProt model using the full training set. The resulting model had an AUC of 0.84 and an accuracy of 75% for the full training set. The five most important features according to the GBM are: i) maximum identity to the E. coli PDB subset (14.1%), ii) isoelectric point (6.2%), iii) lysine content (3.9%), iv) predicted number of amino acids in...
transmembrane helices in the first 60 amino acids of the protein (3.4%), and v) glutamine content (3.3%) (Table S5).

Performance evaluation and comparison

We used the SoluProt test set to evaluate and compare SoluProt to 11 previously published tools. The evaluation relied on both threshold-independent (area under the ROC curve) and threshold-dependent metrics (accuracy, Matthew's correlation coefficient, and confusion matrices). For the threshold-dependent metrics, we applied a threshold of 0.5 or the thresholds recommended by the authors of the corresponding method (Table 1). SoluProt achieved the highest accuracy (58.4%) and the greatest AUC (0.60) of the tested tools when evaluated against the SoluProt test set (Table 1 and Figure 1), followed by PROSO II and SWI.

While the SoluProt test set is independent of the SoluProt training set, other tools’ training sets might overlap with our test set. We therefore compared the SoluProt test set to the training sets of DeepSol, SKADE, SWI, and SOLpro to quantify their overlaps (Table 2). DeepSol and SKADE have a common training set, which showed the largest overlap (75.1%), followed by the SWI training set (24.9%) and the SOLpro training set (16.5%). SWI benefits from the overlap; it was the third-best tool in our comparison. DeepSol and SKADE ranked 7th and 12th by accuracy with respect to the SoluProt test set despite having the greatest proportion of test sequences in their training set. This comparatively poor performance can be partly explained by differences in solubility annotations between the DeepSol training set and the SoluProt test set (Table 2): 356 (12.3% of the total) sequences annotated as insoluble in the DeepSol training set were annotated as soluble in the SoluProt test set. The total number of disagreements (the sum of false positives and false negatives) ranged from 328 to 525, depending on the binarization threshold applied to the SoluProt test set (Table S7). No training set was published for PROSO II; only an initial set of soluble and insoluble sequences without pre-processing is available. However, the initial set exhibits 97.4% overlap with the SoluProt test set. Therefore, we expect the overlap of the PROSO II training set to also be very high, like the DeepSol training set. Unfortunately, the training sets of other previously developed tools have not been published, preventing a more comprehensive comparison.

The absolute accuracy of the available solubility prediction tools is low (below 60%), so there is clearly room for improvement. Nevertheless, SoluProt and other tools can be useful for protein sequence prioritization (Figure 2), i.e. for selecting a small number of sequences for in-depth experimental characterization from a large database of several hundreds or thousands of sequences. Specifically, predicted solubility values can be used to select a limited number of high-scoring protein sequences. For example, if we use SoluProt predictions to order the SoluProt test set and remove all sequences bar the 10% with the highest scores, we get 199 true positives, i.e. 37.2% more true positives than would be expected with blind selection (145 true positives). This shows that despite their limited accuracy, current solubility predictors are valuable for protein sequence prioritization and can increase the success rate of experimental protein studies.
Table 1. Performance of various solubility predictors using the balanced SoluProt test set of 2,904 sequences.

| Name        | AUC  | T   | ACC  | MCC | TP  | TN  | FP  | FN  |
|-------------|------|-----|------|-----|-----|-----|-----|-----|
| SoluProt    | 0.60 | 0.5 | 58.4% | 0.17 | 829 | 868 | 584 | 623 |
| PROSO II    | 0.60 | 0.6 | 57.6% | 0.16 | 583 | 1089| 363 | 869 |
| SWI         | 0.59 | 0.5 | 56.0% | 0.13 | 1124| 502 | 950 | 328 |
| CamSol      | 0.58 | 1.0 | 55.1% | 0.10 | 662 | 938 | 514 | 790 |
| ESPRESSO    | 0.57 | 0.5 | 54.6% | 0.10 | 960 | 627 | 825 | 492 |
| Protein-Sol | 0.56 | 0.5 | 53.5% | 0.07 | 875 | 678 | 774 | 577 |
| DeepSol     | 0.55 | 0.5 | 52.9% | 0.09 | 224 | 1313| 139 | 1228|
| rWH         | 0.55 | 0.5 | 54.3% | 0.09 | 635 | 941 | 511 | 817 |
| SOLpro      | 0.54 | 0.5 | 52.5% | 0.05 | 621 | 905 | 547 | 831 |
| ccSOL omics | 0.51 | 0.5 | 50.9% | 0.02 | 840 | 638 | 814 | 612 |
| SKADE       | 0.50 | 0.5 | 49.5% | -0.01| 166 | 1272| 180 | 1286|
| RPSP        | 0.50 | 0.5 | 49.7% | -0.01| 467 | 976 | 476 | 985 |

AUC – area under the ROC curve, T – threshold for the soluble class, ACC – accuracy, MCC – Matthew’s correlation coefficient, TP – true positives, TN – true negatives, FP – false positives, FN – false negatives.

Figure 1. Receiver operating curves (ROC) calculated for the balanced SoluProt test set of 2,904 sequences. The predictors are ordered by the area under the receiver operating curve (AUC).

Table 2. Overlaps between the SoluProt test set and available training sets. Two sequences were considered identical if their global sequence identity reported by USEARCH was 100%. Differences in solubility annotations for identical sequences were quantified using confusion matrix terms (TP, TN, FP, FN).
and FN). The solubility annotations of the SoluProt test set are assumed to reflect the true solubilities of the proteins.

Table 2. Overlaps between the SoluProt test set and available training sets. Two sequences were considered identical if their global sequence identity reported by USEARCH was 100%. Differences in solubility annotations for identical sequences were quantified using confusion matrix terms (TP, TN, FP, and FN). The solubility annotations of the SoluProt test set are assumed to reflect the true solubilities of the proteins.

| Dataset             | Size  | Test set overlap | TP  | TN   | FP  | FN  |
|---------------------|-------|------------------|-----|------|-----|-----|
| PROSO II initial    | 129,643 | 2,829 (97.4%)     | 894 | 1378 | 49  | 508 |
| DeepSol/SKADE\(^a\) | 69,420  | 2,181 (75.1%)     | 682 | 1077 | 66  | 356 |
| SWI                 | 12,216 | 723 (24.9%)       | 476 | 188  | 43  | 16  |
| SOLpro              | 17,408 | 480 (16.5%)       | 170 | 115  | 39  | 156 |
| SoluProt            | 10,912 | 0 (0.0%)          | 0   | 0    | 0   | 0   |

TP – true positives, TN – true negatives, FP – false positives, FN – false negatives. \(^a\)DeepSol and SKADE share the same training set.

Figure 2. Increases in the number of true positives resulting from sequence prioritization using the tested solubility prediction tools. The SoluProt test set sequences were ordered by predicted solubility based on each predictor’s output, and a variable percentage of the sequences with the worst predicted solubility was then removed. The increase in the number of true positives was then calculated relative to a baseline random selection. For example, upon randomly removing 90% of the test set sequences (2,614 samples), we would expect half of the remaining 290 sequences to be true positives.
Conclusions

We have developed a novel software tool, SoluProt, for sequence-based prediction of soluble protein expression in *E. coli*. The tool simultaneously predicts the solubility and expressibility of the proteins under consideration. SoluProt achieved a higher accuracy (58.4%) and AUC (0.60) than a suite of alternative solubility prediction tools when evaluated using the balanced independent SoluProt test set of 2,904 sequences. PROSO II, SWI, and CamSol were the next best tools, achieving accuracies of 57.6%, 56.0%, and 55.1%, respectively. SoluProt also performed well in protein prioritization. The main strengths of SoluProt are that it was trained using a dataset generated by thorough pre-processing of the noisy TargetTrack data, and was validated using a high-quality independent test set.

Surprisingly, the recently reported DeepSol (Khurana et al., 2018) and SKADE (Raimondi et al., 2020) tools, which are based on deep learning methods, performed worse than the simpler and mostly older methods PROSO II (Smialowski et al., 2012), SWI (Bhandari et al., 2020), and CamSol (Sormanni et al., 2015) in our comparison. This may be partly due to the overlap of their training set with our test set and disagreements between these sets with respect to the solubility of certain sequences.

The SoluProt predictor is available via a user-friendly web server or as a standalone software package at [https://loschmidt.chemi.muni.cz/soluprot/](https://loschmidt.chemi.muni.cz/soluprot/). The SoluProt web server has already predicted the solubility of over 4,200 unique protein sequences in the six months since its launch in February 2020. It has also been integrated into the web server EnzymeMiner (Hon et al., 2020) for automated mining of soluble enzymes from protein databases ([https://loschmidt.chemi.muni.cz/enzymeminer/](https://loschmidt.chemi.muni.cz/enzymeminer/)).

Funding

This work was supported by Czech Ministry of Education [857560, 02.1.01/0.0/0.0/18_046/0015975, CZ.02.1.01/0.0/0.0/16_026/0008451, CZ.02.1.01/0.0/0.0/16_019/0000868]; Czech Grant Agency (20-15915Y); European Commission [857560, 720776, 814418]; and AI Methods for Cybersecurity and Control Systems project of the Brno University of Technology [FIT-S-20-6293]. Computational resources were supplied by the project ‘e-Infrastruktura CZ’ [e-INFRA LM2018140] and by the ELIXIR-CZ [LM2015047]. Funding for open access charge: Czech Ministry of Education.

References

Agostini, F. et al. (2014) ccSOL omics: a webserver for solubility prediction of endogenous and heterologous expression in Escherichia coli. *Bioinformatics*, 30, 2975–2977.

Agostini, F. et al. (2012) Sequence-Based Prediction of Protein Solubility. *Journal of Molecular Biology*, 421, 237–241.

Berman, H.M. et al. (2017) Protein Structure Initiative - TargetTrack 2000-2017 - all data files. *Zenodo*.

Berman, H.M. et al. (2000) The Protein Data Bank. *Nucleic Acids Res.*, 28, 235–242.

Bhandari, B.K. et al. (2020) Solubility-Weighted Index: fast and accurate prediction of protein solubility. *Bioinformatics*.

Burley, S.K. et al. (2019) RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.*, 47, D464–D474.

Carballo-Amador, M.A. et al. (2019) Surface patches on recombinant erythropoietin predict protein solubility: engineering proteins to minimise aggregation. *BMC Biotechnol.*, 19.

Carlson, E.D. et al. (2012) Cell-free protein synthesis: applications come of age. *Biotechnol. Adv.*, 30, 1185–1194.

Chan, P. et al. (2013) Soluble expression of proteins correlates with a lack of positively-charged surface. *Sci Rep.*, 3.

Cilia, E. et al. (2014) The DynaMine webserver: predicting protein dynamics from sequence. *Nucleic Acids Res.*, 42, W264-270.
Supplementary Information

SoluProt: Prediction of Soluble Protein Expression in *Escherichia coli*

Jiri Hon$^{1,2,3}$, Martin Marusiak$^3$, Tomas Martinek$^3$, Antonin Kunka$^{1,2}$, Jaroslav Zendulka$^3$, David Bednar$^{1,2}$, Jiri Damborsky$^{1,2}$

$^1$Loschmidt Laboratories, Centre for Toxic Compounds in the Environment RECETOX and Department of Experimental Biology, Faculty of Science, Masaryk University, 625 00 Brno, Czech Republic;
$^2$International Clinical Research Center, St. Anne’s University Hospital Brno, 656 91 Brno, Czech Republic;
$^3$IT4Innovations Centre of Excellence, Faculty of Information Technology, Brno University of Technology, 612 66 Brno, Czech Republic
Table S1. TargetTrack experiment states signifying soluble expression. The list was compiled by the authors of PROSO II (Smialowski et al., 2012).

| Experiment states |
|--------------------|
| soluble, purified, crystallized, hsqc, structure, in pdb, native diffraction-data, NMR assigned, phasing diffraction-data, diffraction, in bmrb, nmr structure, crystal structure, diffraction-quality crystals |

Table S2. Specific keywords signifying expression in *E. coli*.

| Specific keywords |
|-------------------|
| BL21, DE3, rosetta, xl10, DH10B, CodonPlus, RIPL, RIL, DB3.1, DB3, arctic, origami |

Table S3. Protocols identified by generic *E. coli* phrases and manually checked to signify expression in *E.coli*.

| Protocol ids |
|--------------|
| NYSGXRCSGX_MOLBIO_TOPO_TRANSFORM |
| JCSG-E_Ecoli_GNF_1 |
| CSGID-NU_SelMet_expression |
| CSGID-NU_native_expression |
| MPP-LP.4341 |
| MCGS-NU_default_expression |
| NYSGXRCSGX_FERM_ECOLI_LB |
| MPP-LP.4813 |
| SSGCID-33 |
| NYSGXRCSGX_FERM_ECOLI_M9 |
| CSGID-NU_default_expression |
| SSGCID-2 |
| SSGCID-31 |
| SSGCID-1 |
| CESG-MAXWELL 16 EXPRESSION TESTING (R D) v.1.0.0 |
| MPP-LP.4814 |
| SSGCID-128 |
| EFI-SeMET expression in HY Media-PSI2 |
| SGX-SGX_FERM_ECOLI_LB_CFTR |
| SGX-SGX_MOLBIO_EXPR_SOL_CFTR |
Figure S1. Sequence length distribution of soluble and insoluble proteins in the SoluProt datasets. The x-axis is limited to the range of 0–1000 amino acids to improve readability. The longest sequences in the test and training sets have 790 and 2842 amino acids, respectively.

Table S4. Sequence physicochemical features. Most of the features were extracted using the Biopython package (Cock et al., 2009).

| Name                              | Description                                                                 |
|-----------------------------------|-----------------------------------------------------------------------------|
| physico_chemical_fracnumcharge    | Fraction of charged amino acids (R, K, D, E).                               |
| physico_chemical_kr_ratio         | Ratio of K and R content.                                                   |
| physico_chemical_aa_helix         | Fraction of helix amino acids (V, I, Y, F, W, L).                           |
| physico_chemical_aa_sheet         | Fraction of sheet amino acids (E, M, A, L).                                |
| physico_chemical_aa_turn          | Fraction of turn amino acids (N, P, G, S).                                 |
| physico_chemical_molecular_weight | Molecular weight.                                                           |
| physico_chemical_avg_molecular_weight | Molecular weight normalized by the sequence length.                         |
| physico_chemical_aromaticity      | Fraction of aromatic amino acids (Y, W, F)                                 |
| physico_chemical_flexibility      | Flexibility according to (Vihinen et al., 1994)                            |
| physico_chemical_gravy           | Grand average of hydropathy according to (Kyte and Doolittle, 1982)        |
| physico_chemical_isoelectric_point | Isoelectric point using methods of Bjellqvist (Bjellqvist et al., 1993, 1994) |
| physico_chemical_instability_index | Instability index according to (Guruprasad et al., 1990)                   |
Table S5. Sequence features and their importance in the final SoluProt model.

| #  | Feature                                | Importance |
|----|----------------------------------------|------------|
| 1  | ecoli_usearch_identity_identity        | 14.11%     |
| 2  | physico_chemical_isoelectric_point     | 6.20%      |
| 3  | monomers_K                            | 3.87%      |
| 4  | tmhmm_first_60                        | 3.43%      |
| 5  | monomers_Q                            | 3.31%      |
| 6  | monomers_E                            | 2.02%      |
| 7  | monomers_M                            | 1.94%      |
| 8  | physico_chemical_aa_helix             | 1.87%      |
| 9  | dimers_comb_DK                        | 1.77%      |
| 10 | physico_chemical_molecular_weight      | 1.56%      |
| 11 | dimers_comb_EN                        | 1.53%      |
| 12 | dimers_comb_AA                        | 1.49%      |
| 13 | monomers_Y                            | 1.39%      |
| 14 | monomers_C                            | 1.37%      |
| 15 | dimers_comb_EK                        | 1.25%      |
| 16 | dimers_comb_AI                         | 1.14%      |
| 17 | dimers_comb_DT                        | 1.11%      |
| 18 | dimers_comb_DR                        | 1.09%      |
| 19 | dimers_comb_RR                        | 1.09%      |
| 20 | monomers_W                            | 1.07%      |
| 21 | dimers_comb_IS                        | 1.05%      |
| 22 | dimers_comb_PQ                        | 1.02%      |
| 23 | dimers_comb_GK                        | 1.02%      |
| 24 | dimers_comb_EI                        | 1.01%      |
| 25 | dimers_comb_DI                        | 0.95%      |

| #  | Feature                                | Importance |
|----|----------------------------------------|------------|
| 26 | dimers_comb_EE                        | 0.95%      |
| 27 | dimers_comb_LT                        | 0.93%      |
| 28 | dimers_comb_EM                        | 0.90%      |
| 29 | dimers_comb_LL                        | 0.89%      |
| 30 | dimers_comb_MV                        | 0.89%      |
| 31 | monomers_F                            | 0.87%      |
| 32 | dimers_comb_AQ                        | 0.86%      |
| 33 | dimers_comb_IL                        | 0.85%      |
| 34 | dimers_comb_LQ                        | 0.85%      |
| 35 | dimers_comb_GN                        | 0.84%      |
| 36 | dimers_comb_FP                        | 0.82%      |
| 37 | dimers_comb_KQ                        | 0.82%      |
| 38 | dimers_comb_QT                        | 0.80%      |
| 39 | dimers_comb_GL                        | 0.79%      |
| 40 | dimers_comb_FT                        | 0.78%      |
| 41 | dimers_comb_AM                        | 0.78%      |
| 42 | dimers_comb_TY                        | 0.77%      |
| 43 | dimers_comb_EV                        | 0.76%      |
| 44 | dimers_comb_EL                        | 0.75%      |
| 45 | dimers_comb_EP                        | 0.75%      |
| 46 | dimers_comb_VY                        | 0.75%      |
| 47 | dimers_comb_QV                        | 0.72%      |
| 48 | dimers_comb_LN                        | 0.71%      |
| 26 | dimers_comb_EE                        | 0.95%      |
| 27 | dimers_comb_LT                        | 0.93%      |
| #  | Feature                   | Importance |
|----|---------------------------|------------|
| 49 | dimers_comb_DE            | 0.71%      |
| 50 | dimers_comb_SV            | 0.69%      |
| 51 | dimers_comb_GG            | 0.68%      |
| 52 | dimers_comb_DM            | 0.67%      |
| 53 | monomers_H                | 0.67%      |
| 54 | physico_chemical_fracnumcharge | 0.66%            |
| 55 | dimers_comb_IT            | 0.65%      |
| 56 | dimers_comb_FI            | 0.65%      |
| 57 | dimers_comb_AC            | 0.65%      |
| 58 | dimers_comb_KV            | 0.63%      |
| 59 | dimers_comb_AV            | 0.63%      |
| 60 | dimers_comb_CP            | 0.63%      |
| 61 | dimers_comb_MN            | 0.62%      |
| 62 | dimers_comb_FL            | 0.62%      |
| 63 | dimers_comb_RS            | 0.61%      |
| 64 | dimers_comb_GH            | 0.57%      |
| 65 | dimers_comb_EF            | 0.55%      |
| 66 | dimers_comb_AK            | 0.55%      |
| 67 | dimers_comb_MW            | 0.54%      |
| 68 | dimers_comb_AG            | 0.54%      |
| 69 | dimers_comb_NY            | 0.52%      |
| 70 | dimers_comb_CI            | 0.52%      |
| 71 | dimers_comb_HK            | 0.51%      |

| #  | Feature                   | Importance |
|----|---------------------------|------------|
| 74 | dimers_comb_CG            | 0.49%      |
| 75 | dimers_comb_KM            | 0.48%      |
| 76 | dimers_comb_RW            | 0.48%      |
| 77 | dimers_comb_AN            | 0.47%      |
| 78 | dimers_comb_HT            | 0.47%      |
| 79 | dimers_comb_EH            | 0.46%      |
| 80 | dimers_comb_GM            | 0.46%      |
| 81 | dimers_comb_CY            | 0.46%      |
| 82 | dimers_comb_DW            | 0.44%      |
| 83 | dimers_comb_HL            | 0.43%      |
| 84 | dimers_comb_IY            | 0.42%      |
| 85 | dimers_comb_PW            | 0.41%      |
| 86 | dimers_comb_CS            | 0.39%      |
| 87 | dimers_comb_KR            | 0.37%      |
| 88 | dimers_comb_FM            | 0.37%      |
| 89 | dimers_comb_FH            | 0.32%      |
| 90 | dimers_comb_GT            | 0.30%      |
| 91 | dimers_comb_MY            | 0.29%      |
| 92 | dimers_comb_CC            | 0.27%      |
| 93 | dimers_comb_HW            | 0.25%      |
| 94 | dimers_comb_MM            | 0.24%      |
| 95 | dimers_comb_WW            | 0.12%      |
| 96 | tmhmm_pred_hel            | 0.01%      |
Table S6. Optimized hyperparameters of the Gradient Boosting classifier. In each stage, one or two parameters were optimized while the other parameters were left either at their final values from previous stages or at their default values if they had not been optimized previously. The parameters were first optimized using a large step size. Smaller steps were then used for refinement. The learning rate was lowered from the default value of 0.1 to 0.01 before optimizing the number of estimators. Parameters not mentioned here were left at their default values.

| Stage | Parameter                | Range       | Step       | Final value |
|-------|--------------------------|-------------|------------|-------------|
| 1     | n_estimators             | 20-100      | 10         | .a          |
| 2     | max_depth                | 3-17        | 2, 1       | 6           |
|       | min_samples_split        | 100-1400    | 100, 50    | 1250        |
| 3     | min_samples_leaf         | 1-160       | 10, 5      | 6           |
| 4     | max_features             | 5-96        | 5          | 40          |
| 5     | subsample                | 0.5-1       | 1/40       | 0.525       |
| 6     | learning_rate            | .b          | .b         | 0.01        |
| 7     | n_estimators             | 200-1800    | 200, 50    | 1500        |

a The parameter was optimized again in the 7th stage, after which its final value was determined; b The learning rate was set to a fixed value; The final set of parameters was as follows: criterion='friedman_mse', init=None, learning_rate=0.01, loss='deviance', max_depth=6, max_features=40, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=6, min_samples_split=1250, min_weight_fraction_leaf=0.0, n_estimators=1500, n_iter_no_change=None, presort='auto', random_state=9, subsample=0.525, tol=0.0001, validation_fraction=0.1, verbose=0, warm_start=False.

Table S7. Class disagreements between available training sets and the SoluProt test set when applying different binarization thresholds.

| Dataset    | FP1 | FP2 | FP3 | FP4 | FP5 | FN1 | FN2 | FN3 | FN4 | FN5 | E1 | E2 | E3 | E4 | E5 |
|------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|----|----|----|----|----|
| PROSO II   | 49  | 55  | 188 | 384 | 509 | 508 | 377 | 309 | 204 | 149 | 557| 432| 497| 588| 658|
| initial    |     |     |     |     |     |     |     |     |     |     |    |    |    |    |    |
| DeepSol/   | 66  | 76  | 183 | 324 | 426 | 356 | 252 | 205 | 129 | 99  | 422| 328| 388| 453| 525|
| SKADE      |     |     |     |     |     |     |     |     |     |     |    |    |    |    |    |
| SWI        | 43  | 94  | 163 | 256 | 339 | 16  | 11  | 7   | 5   | 2   | 59 | 105| 170| 261| 341|
| SOLpro     | 39  | 40  | 46  | 83  | 106 | 156 | 132 | 87  | 52  | 35  | 195| 172| 133| 135| 141|
| SoluProt   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0  | 0  | 0  | 0  | 0  |

FP – false positives, FN – false negatives, E – total number of errors (FP + FN). The numerical suffix denotes the binarization threshold used for the SoluProt test set. For example, a binarization threshold of 2 means that all sequences with solubility scores of 2 or above are considered soluble, and all others are considered insoluble.
References

Bjellqvist, B. et al. (1994) Reference points for comparisons of two-dimensional maps of proteins from different human cell types defined in a pH scale where isoelectric points correlate with polypeptide compositions. *Electrophoresis*, **15**, 529–539.

Bjellqvist, B. et al. (1993) The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. *Electrophoresis*, **14**, 1023–1031.

Cock, P. J. A. et al. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.

Guruprasad, K. et al. (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 Eng Des Sel*, **4**, 155–161.

Kyte, J. and Doolittle, R. F. (1982) A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, **157**, 105–132.

Smialowski, P. et al. (2012) PROSO II - a new method for protein solubility prediction. *FEBS J.*, **279**, 2192–2200.

Vihinen, M. et al. (1994) Accuracy of protein flexibility predictions. *Proteins: Structure, Function, and Bioinformatics*, **19**, 141–149.
