

Research Article

Forecasting Host Cells for Recombinant Protein Expression

Hung Van Le* 

Biologics & Drug Targets, ProSci LLC, Rockaway, The United States

Abstract

Selection of an appropriate host cell is a critical determinant of success in recombinant protein expression. In practice, host choice is still largely guided by individual experience, ad hoc consultation of the literature, and intuitive decision-making, often resulting in suboptimal expression outcomes and costly cycles of experimental trial and error. Despite several decades of accumulated empirical knowledge in the field, there is currently no systematic, evidence-based framework for forecasting host cell suitability from protein sequence and structural characteristics. The purpose of this study was to develop predictive models that enable rational selection of host cells for recombinant protein expression based on intrinsic protein features. To achieve this, we leveraged collective experimental experience embedded in publicly available structural data. Protein entries from the Protein Data Bank were curated and analyzed, and logistic regression approaches were applied to relate expression outcomes to a range of protein attributes, including structural parameters, stability indices, predicted subcellular localization, and post-translational modification requirements. Using these variables, we constructed and validated statistical models capable of forecasting expression preferences across four commonly used host systems: *Escherichia coli*, insect cells, mammalian cells, and yeast. Model performance was assessed using internal validation procedures, demonstrating that distinct combinations of protein features are associated with differential expression success among host types. In conclusion, this work provides an evidence-based and quantitative framework for predicting suitable host cells for recombinant protein expression. By translating accumulated empirical knowledge into practical predictive tools, the proposed models reduce reliance on subjective judgment and trial-and-error experimentation. To facilitate broad adoption, the models, together with user guidance, have been implemented in a publicly accessible web server, offering a practical resource to improve experimental efficiency and success rates in protein expression studies.

Keywords

Recombinant Protein Expression, Host Cell Selection, Logistic Regression, Predictive Modeling

1. Introduction

Since the introduction of recombinant DNA technology [1-3], host cell selection for protein expression has largely relied on personal expertise and trial-and-error approaches. With the

proliferation of available host cell types, expression vectors, and commercial reagent kits, some degree of success can usu-

*Correspondence: Hung Van Le (hung.v.le@prosciconsulting.com)

Received: 15 January 2026; **Accepted:** 26 January 2026; **Published:** 6 February 2026



Copyright: © The Author(s), 2026. Published by Science Publishing Group. This is an **Open Access** article, distributed under the terms of the Creative Commons Attribution 4.0 License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

ally be achieved in recombinant protein expression laboratories. Nevertheless, obtaining optimized expression—defined as large amounts of soluble, biologically active protein with appropriate post-translational features suitable for downstream processing—often remains a daunting and costly iterative procedure.

Computational tools have been developed to address aspects of this challenge, primarily focusing on predicting protein solubility [4-6] or pairing vectors with protein sequences [7] for expression in *Escherichia coli*. Machine learning approaches have also been applied to predict expression potential in *Bacillus subtilis* [8], while others have focused on translation optimization strategies [9, 10]. These tools are valuable once an expression host has already been selected. However, none directly address the initial and critical task of pairing a protein sequence with an optimal expression host—a step that, if performed correctly, could eliminate or substantially reduce the need for subsequent optimization.

The vast amount of historical data accumulated over several decades warrants the development of a systematic, evidence-

based method to supplement or replace reliance on literature searches and intuitive decision-making. Here, we leverage information from the Protein Data Bank [11] (PDB; rcsb.org), specifically the reported expression systems used to produce proteins for structural determination. Logistic regression [12] was applied to identify structural, functional, and subcellular localization attributes associated with successful expression in specific host cell types. Host-specific logit models were derived and validated for calculating probabilities of expression and for rank-ordering host preferences. We discuss the utility of this statistical approach as a supplement to traditional host selection strategies.

2. Methods

Building predictive models to determine optimal pairing between expression hosts and protein sequences was carried out according to the workflow outlined below.

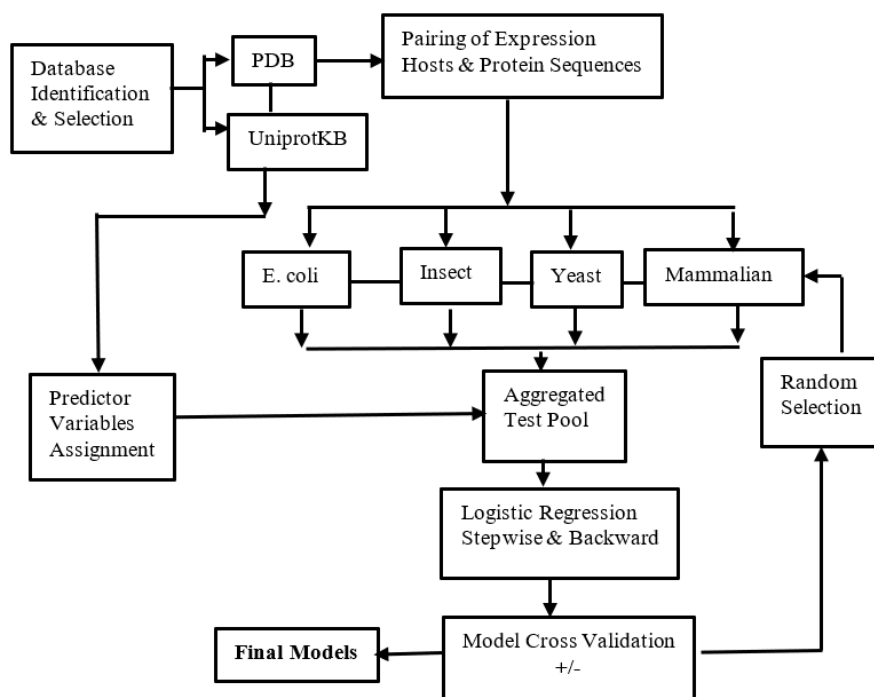


Figure 1. Predictive modeling workflow.

2.1. Building Training and Test Sets

Datasets consisting of protein sequences paired with specific expression host systems were constructed primarily using the Protein Data Bank (PDB) via its advanced search module. Random sampling was performed in Microsoft Excel using either the Random Sampling or Random Number Generation functions of the XLMiner Analysis ToolPak (Microsoft Corporation, Redmond, WA).

2.2. Predictor Variable Assignments

Protein sequence-derived variables—including total number of amino acid residues (r), isoelectric point (pI), Kyte-Doolittle grand average of hydropathicity (GRAVY) index [13] (hp), and instability index [14] (ii)—were calculated by submitting amino acid sequences of the corresponding PDB entries to the ProtParam tool [15] available on the ExPASy server maintained by the Swiss Institute of Bioinformatics

(SIB).

Tertiary structural classification was based on PDB annotations when available; otherwise, structures were examined by direct visualization using the 3DView module in Mol* [16]. When experimentally determined structures were unavailable, tertiary structural classes were inferred from predicted structures of the exact protein sequence, or, when necessary, from the closest homolog or ortholog obtained from the AlphaFold Protein Structure Database [17]. Structural classification followed the SCOP framework, with minor modifications.

Annotations from the PDB and UniProtKB [18] were primarily used for quaternary structural classification, determination of membrane association status, and assignment of molecular function. In cases where annotations were insufficient, membrane association was inferred using multiple prediction approaches: identification of transmembrane helices using TMHMM [19, 20], (DTU Health Tech) and Phobius [21, 22], (Stockholm Bioinformatics Center); prediction of lipid anchorage sites using GPS-Lipid [23] (S-palmitoylation, N-myristoylation, S-farnesylation, and S-geranylgeranylation) and the big-PI Predictor [24] (for GPI anchors); and prediction of comprehensive subcellular localization using DeepLoc [25] (DTU Health Tech).

The predicted number of N-linked glycosylation sites was obtained by submitting protein sequences—preferably including signal peptides when known—to NetNGlyc-1.0 [26].

2.3. Logistic Regression and Model Validation

All logistic regression analyses were performed in Microsoft Excel using the logistic regression module of the XLMiner Analysis ToolPak (Microsoft Corporation, Redmond, WA). Model validation was conducted using cross-validation approaches. For each expression host, average expression probabilities were calculated from random samplings of 25 protein sequences drawn from the training dataset and compared against a reserved test set. Statistical significance was assessed using analysis of variance (ANOVA), implemented via the ANOVA module of the XLMiner Analysis ToolPak.

3. Results

3.1. Construction of Training and Validation

Test Sets

The Protein Data Bank (PDB) served as the primary source of information for identifying protein sequences expressed in different host cell types. At the time of data collection (09/20/2021), the PDB contained 156,140 entries (including redundant records) corresponding to proteins expressed in both prokaryotic (87%) and eukaryotic (13%) systems (Table 1).

Table 1. Distribution of expression host/protein sequence pairs in the PDB.

Expression host distribution in PDB*	156,140
Prokaryote expression	135,653
E. coli	134,513
Others	1,140
Eukaryote expression	20,487
Insect	9,637
Mammalian	7,186
Yeast	3,165
Fungus	418
Plant	56
Protozoa	25

* The information captured on 09/20/2021 was divided into two large groups of prokaryote and eukaryote hosts, and further subdivided into 2 subgroups of prokaryotes and 6 subgroups of eukaryotes.

The predominance of prokaryotic expression systems is expected, given their ease of cultivation, broad availability of expression vectors, and cost effectiveness. Among prokaryotes, *Escherichia coli* strains accounted for approximately 99% of all entries. Within eukaryotic systems, insect (47%), mammalian (35%), and yeast (15%) hosts dominated, whereas fungal, plant, and protozoan systems collectively accounted for less than 2.5% of entries.

Based on this distribution, subsequent analyses focused on the four most extensively represented host systems: *E. coli*, insect, mammalian, and yeast cells. Initial random samplings of 50 protein sequences per host type were performed and combined for logistic regression analysis using backward variable selection to identify the most informative predictor variables. This preliminary training dataset, comprising approximately 200 protein-host pairs, was iteratively expanded through four additional sampling cycles, yielding stable sets of predictor variables and associated regression coefficients (β values). In parallel, independent validation datasets consisting of 25 non-overlapping protein sequences per host were randomly selected and reserved for model validation.

3.2. Regression Parameters Consideration

Protein sequences (dependent variables) were coded dichotomously (1, 0) to indicate positive expression or no expression reported in the PDB for a given host system. Each sequence was characterized by nine independent (predictor) variables describing structural properties, stability, subcellular localization, and molecular function.

Three continuous predictors were included to capture pri-

mary sequence characteristics: total number of amino acid residues (r), isoelectric point (pI), and the Kyte-Doolittle grand average of hydropathicity (GRAVY) index [13] (hp). Five categorical predictors were used to represent tertiary structural class (t), quaternary structural class (q), membrane association status (m), extent of N-glycosylation (ng), and molecular function (mf). In vivo protein stability was represented by the instability index (ii) as defined by Guruprasad, Bhasker Reddy, and Pandit [14].

Tertiary structural classification was based primarily on SCOP 2 [27] and comprised five categories: small protein (sp), all- α , all- β , alternating α/β , and segregated $\alpha+\beta$ domains. These classes were numerically coded from 1 to 5, respectively. Large multidomain proteins exhibiting combinations of α , β , and α/β architectures were not assigned weighted averages; instead, they were uniformly coded as 5 ($\alpha+\beta$).

Table 2. Numerical coding scheme for structural classes and membrane association.

Structural and membrane class definition*	Numerical coding
Tertiary structural classes (t)	
Small protein (sp)	1
All alpha (α)	2
All beta (β)	3
Alternating alpha/beta (α/β)	4
Segregated alpha and beta ($\alpha+\beta$)	5
Quaternary structural classes (q)	
Monomer	1
Homopolymer, nth order	2
Homopolymer, undefined size	3
Heteropolymer, nth order	4
Heteropolymer, undefined size	5
State of membrane association (m)	
None	0
Transmembrane helices, nth order	1
Membrane associated	2

*Tertiary structural classes were defined according to SCOP 2 with minor modifications. Quaternary structural classes and membrane association states were derived from annotations in the PDB, Alpha Fold Protein Structure Database and UniprotKB.

Quaternary structure was similarly divided into five categories, numerically coded from 1 to 5: monomer; homopolymer of nth order; homopolymer aggregate of undefined size; het-

eropolymer of nth order; and heteropolymer aggregate of undefined size. Membrane association was represented by three states: soluble proteins with no known membrane association (0), proteins embedded in the membrane via one or more transmembrane helices (1), and proteins associated with the cell surface without transmembrane helices (2) (Table 2).

Because N-glycosylation status was not experimentally determined for all proteins included in this study, glycosylation was instead represented by an index corresponding to the total number of potential N-linked glycosylation sites, as inferred from the presence of canonical N-X-S/T sequence motifs [28]. Molecular function was assigned based on the Gene Ontology (GO) molecular function classification (version 2017-09-30) [29, 30], which comprises 15 functional categories, each numerically coded as shown in Table 3.

Table 3. Numerical coding scheme for molecular functions.

Molecular function*	Numerical coding
Antioxidant	1
Binding	2
Catalytic	3
Hijacked molecular function	4
Molecular carrier activity	5
Molecular function regulator	6
Molecular transducer activity	7
Nutrient reservoir activity	8
Protein tag	9
Signal transducer activity	10
Structural molecule activity	11
Toxin activity	12
Transcription regulator activity	13
Translation regulator activity	14
Transporter activity	15

* Gene product molecular functions were assigned according to the Gene Ontology Project (version 2017-09-30).

3.3. Logistic Regression and Modeling

Proteins included in the database were randomly selected from the PDB based on documented positive expression outcomes in four host cell types selected for this study. The final training dataset comprised 953 proteins, including 260 expressed in E. coli, 246 in insect cells, 215 in yeast, and 229 in mammalian cells (Appendix I).

Stepwise logistic regression was performed in a backward

elimination mode, iteratively testing the statistical significance of predictor variables for inclusion or removal from the model. The final models retained six statistically significant β coefficients ($p \leq 0.05$) for positive expression outcomes in *E.*

coli and yeast, seven for insect cells, and eight for mammalian cells (Tables 4-7). Although each host-specific model contained a distinct set of predictor variables, partial overlap among predictors was observed across host types.

Table 4. Coefficients and associated statistics* for logistic regression on expression in *E. coli*.

	Coefficient	SE	P-value	OR	OR 95% CI	
					Lower	Upper
Intercept	0.822	0.443	0.064	2.275	0.954	5.423
r	-0.003	0.001	2.1E-08	0.997	0.996	0.998
pI	-0.128	0.049	0.009	0.880	0.780	0.969
t	0.238	0.064	4.0E-04	1.268	1.118	1.438
q	-0.352	0.063	2.8E-08	0.703	0.621	0.796
ng	-0.209	0.054	1.1E-04	0.811	0.730	0.902
mf	0.048	0.023	0.036	1.049	1.003	1.097

*The model fitted from the data in the table is $\text{logit}(p) = 0.822 - 0.003r - 0.128pI + 0.238t - 0.352q - 0.209ng + 0.048mf$, where p is the probability of expression, r the number of amino acid residues, pI the isoelectric point, t the tertiary structural class, q the quaternary structural class, ng the degree of predicted N-glycosylation and mf the molecular function. SE, Standard Error; OR, Odd Ratio; CI, Confidence Interval. Goodness of fit was ascertained from χ^2 statistics of 149 for sample size $N=953$.

Table 5. Coefficients and associated statistics* for logistic regression on expression in insect cells.

	Coefficient	SE	P-value	OR	OR 95% CI	
					Lower	Upper
Intercept	-3.204	0.480	2.5E-11	0.041	0.016	0.104
r	0.001	2.0E-04	0.002	1.001	1.000	1.001
hp	1.128	0.323	5.0E-04	3.089	1.641	5.812
ii	0.038	0.008	3.9E-06	1.039	1.022	1.056
t	0.323	0.070	3.9E-06	1.381	1.204	1.583
m	0.381	0.167	0.023	1.463	1.054	2.031
q	-0.158	0.062	0.011	0.854	0.756	0.965
mf	-0.101	0.030	0.001	0.904	0.852	0.960

*The model fitted from the data in the table is $\text{logit}(p) = -3.204 + 0.001r + 1.128hp + 0.038ii + 0.323t + 0.381m - 0.158q - 0.101mf$, where p is the probability of expression, r the number of amino acid residues, hp the Grand Average Hydropathicity (GRAVY) index, ii the instability index, t the tertiary structural class, m the state of membrane association, q the quaternary structural class and mf the molecular function. SE, Standard Error; OR, Odd Ratio; CI, Confidence Interval. Goodness of fit was ascertained from χ^2 statistics of 84 for sample size $N=953$.

Table 6. Coefficients and associated statistics* for logistic regression on expression in mammalian cells.

	Coefficient	SE	P-value	OR	OR 95% CI	
					Lower	Upper
Intercept	-1.977	0.567	0.001	0.138	0.046	0.421
r	-0.002	4.0E-04	2.0E-04	0.999	0.998	0.999
pI	0.125	0.054	0.020	1.133	1.020	1.259
ii	0.025	0.008	0.002	1.025	1.009	1.041
t	-0.334	0.072	3.6E-06	0.716	0.621	0.825
m	-0.834	0.282	0.003	0.434	0.250	0.755
q	0.343	0.064	8.0E-08	1.409	1.243	1.598
ng	0.243	0.039	2.8E-10	1.275	1.182	1.375
mf	-0.152	0.039	9.1E-05	0.859	0.797	0.927

*The model fitted from the data in the table is $\text{logit}(p) = -1.977 - 0.002r + 0.125pI + 0.025ii - 0.334t - 0.834m + 0.343q + 0.243ng - 0.152mf$, where p is the probability of expression, r the number of amino acid residues, pI the isoelectric point, ii the instability index, t the tertiary structural class, m the state of membrane association, q the quaternary structural class, ng the degree of predicted N-glycosylation and mf the molecular function. SE, Standard Error; OR, Odd Ratio; CI, Confidence Interval. Goodness of fit was ascertained from χ^2 statistics of 162 for sample size N=953.

Table 7. Coefficients and associated statistics* for logistic regression on expression in yeast cells.

	Coefficient	SE	P-value	OR	OR 95% CI	
					Lower	Upper
Intercept	-0.798	0.352	0.023	0.450	0.226	0.898
r	0.002	3.0E-04	5.1E-06	1.002	1.001	1.002
ii	-0.046	0.009	1.1E-07	0.955	0.939	0.972
m	0.392	0.167	0.019	1.480	1.068	2.051
q	0.171	0.065	0.009	1.186	1.045	1.347
ng	-0.083	0.041	0.042	0.920	0.849	0.997
mf	0.096	0.023	1.8E-05	1.101	1.054	1.151

*The model fitted from the data in the table is $\text{logit}(p) = -0.798 + 0.002r - 0.046ii + 0.392m + 0.171q - 0.083ng + 0.096mf$, where p is the probability of expression, r the number of amino acid residues, ii the instability index, m the state of membrane association, q the quaternary structural class, ng the degree of predicted N-glycosylation and mf the molecular function. SE, Standard Error; OR, Odd Ratio. Goodness of fit was ascertained from χ^2 statistics of 108 for sample size N=953.

3.3.1. Common Predictor Variables

Three predictor variables were common to all four expression host models: total number of amino acid residues (r), quaternary structural class (q), and molecular function (mf). The odds ratios (ORs) associated with r were close to unity for all hosts (95% CI, $0.997 < OR < 1.002$), indicating that protein length did not significantly influence expression outcomes.

In contrast, the effects of quaternary structure varied by host system. Odds ratios for q were less than 1 for E. coli (OR = 0.703) and insect cells (OR = 0.854), but greater than 1 for mammalian (OR = 1.409) and yeast cells (OR = 1.186). These results suggest that multimeric proteins (homo- or hetero-oligomeric) are more favorably expressed in mammalian and yeast systems, whereas they exhibit reduced expression odds in E. coli and insect cells.

Molecular function (mf) exerted a modest but measurable influence on expression outcomes. Odds ratios slightly exceeded 1 for *E. coli* (OR = 1.049) and yeast cells (OR = 1.101), while values below 1 were observed for insect (OR = 0.904) and mammalian cells (OR = 0.859). Consequently, proteins classified as transporters, transcription regulators, or translation regulators were marginally favored for expression in *E. coli* and yeast, but slightly disfavored in insect and mammalian hosts.

3.3.2. Distinctive Predictor Variables for *E. coli* Expression

Predictor variables unique to the *E. coli* expression model included isoelectric point (pI; OR = 0.780), tertiary structural class (t; OR = 1.268), and predicted number of N-glycosylation sites (ng; OR = 0.730). Two of these variables (pI and ng) exhibited odds ratios below 1, indicating decreasing odds of expression as their values increased.

In contrast, the positive odds ratio associated with tertiary structural class suggests a modest preference for more complex fold types, such as α/β and $\alpha+\beta$ architectures, in *E. coli* expression systems.

3.3.3. Distinctive Predictor Variables for Insect Cell Expression

The insect cell expression model was characterized by four distinctive predictor variables: GRAVY index (hp; OR = 3.089), instability index (ii; OR = 1.039), tertiary structural class (t; OR = 1.381), and membrane association status (m; OR = 1.463). Three of these variables exhibited odds ratios greater than 1, indicating increased odds of expression with increasing predictor values.

Notably, the strongest effect was observed for the GRAVY index, suggesting that proteins with higher overall hydrophobicity have substantially greater odds of successful expression in insect cell systems.

3.3.4. Distinctive Predictor Variables for Mammalian and Yeast Cell Expression

Five predictor variables were distinctive for mammalian cell expression: isoelectric point (pI; OR = 1.133), instability index (ii; OR = 1.025), tertiary structural class (t; OR = 0.716), membrane association status (m; OR = 0.434), and predicted number of N-glycosylation sites (ng; OR = 1.275). While higher predicted N-glycosylation was, as expected, associated with increased odds of successful expression, membrane association showed a negative effect. This observation likely reflects practical limitations in culturing mammalian cells to sufficiently high densities to efficiently express membrane-associated proteins.

For yeast expression, three distinctive predictor variables were identified: instability index (ii; OR = 0.955), membrane association status (m; OR = 1.480), and predicted number of N-glycosylation sites (ng; OR = 0.920). Similar to insect cells, increased membrane association was associated with higher odds of expression in yeast. In contrast, a higher degree of predicted N-glycosylation was associated with a modest reduction in expression odds.

3.4. Model Validation

The four logistic regression models developed to predict recombinant protein expression in *Escherichia coli*, insect, mammalian, and yeast cells are summarized in Table 8.

Table 8. Summary of logit models for recombinant protein expression in four principal hosts.

Expression host	Logit model
<i>E. coli</i>	$\text{logit}(p) = 0.822 - 0.003r - 0.128pI + 0.238t - 0.352q - 0.209ng + 0.048mf$
Insect	$\text{logit}(p) = -3.204 + 0.001r + 1.128hp + 0.038ii + 0.323t + 0.381m - 0.158q - 0.101mf$
Mammalian	$\text{logit}(p) = -1.977 - 0.002r + 0.125pI + 0.025ii - 0.334t - 0.834m + 0.343q + 0.243ng - 0.152mf$
Yeast	$\text{logit}(p) = -0.798 + 0.002r - 0.046ii + 0.392m + 0.171q - 0.083ng + 0.096mf$

All four models were subjected to cross-validation by comparing the average predicted probabilities of expression for random subsets of 25 proteins drawn from the training dataset with those obtained for corresponding non-overlapping test sets selected independently from the PDB (Appendix II). As

summarized in Table 9, no statistically significant differences were observed between the mean predicted probabilities for training and test datasets for any of the four host-specific models, based on analysis of variance (ANOVA) using F-statistics at a significance threshold of $p = 0.05$.

Table 9. Cross validation of prediction models for recombinant protein expression in four principal hosts.

Expression host	Probability average (Variance)		F-Statistics*	
	Training set*	Test set*	F	P value
E. coli	0.3548 (0.0260)	0.4116 (0.0243)	1.6001	0.2120
Insect	0.3316 (0.0177)	0.2956 (0.0206)	0.8460	0.3623
Mammalian	0.3568 (0.0302)	0.3108 (0.0259)	0.9422	0.3366
Yeast	0.3468 (0.0412)	0.4284 (0.0542)	1.7423	0.1931

* Protein composition of the training and test datasets is detailed in Appendix B. Predicted probabilities were calculated using the corresponding logit models shown in Table 8. Variance analyses were performed using the single-factor ANOVA module of the XLMiner ToolPak in Microsoft Excel. Values in parentheses represent the variances of the average probabilities of expression.

3.5. Practical Applications

The logit models summarized in Table 8 were applied to predict host cell preferences for specific protein sequences. Table 10 presents the results obtained by applying these models to five protein sequences selected from recent UniProtKB

entries (10/04/2021): the mating pheromone Er-23 from *Euplotes raikovi* (P58547), the methionine repressor MetJ from *Escherichia coli* (B7UNQ8), a rat sodium-dependent glucose transporter (Mfsd4b/Naglt1; Q80T22), homoserine O-succinyltransferase from the marine bacterium *Shewanella pealeana* (A8H6B7), and mitogen-activated protein kinase 1 from *Xenopus laevis* (P26696).

Table 10. Probability of expression and preference rank order* for five candidate proteins.

Expression host	MER23_	METJ_	Mfsd4	metAS	mapk1
	EUPRA	ECO27	Naglt1	Spea	mpk1
E. coli	0.50 (1)	0.55 (1)	0.18 (1)	0.35 (1)	0.21 (1)
Insect	0.13 (0.26)	0.03 (0.05)	0.24 (1.39)	0.28 (0.80)	0.26 (1.28)
Mammalian	0.31 (0.62)	0.07 (0.12)	0.03 (0.17)	0.17 (0.48)	0.19 (0.90)
Yeast	0.11 (0.21)	0.36 (0.65)	0.48 (2.72)	0.14 (0.40)	0.28 (1.37)

* Probabilities of expression were calculated using the corresponding logit models presented in Table 8. Preference rank order (shown in parentheses) was calculated by normalizing probabilities to those obtained for *E. coli*.

MER23_EUPRA: mating pheromone Er-23 (*Euplotes raikovi*, P58547); METJ_ECO27: methionine repressor MetJ (*E. coli*, B7UNQ8); Mfsd4b_Naglt1: rat sodium-dependent glucose transporter 1 (Q80T22); metAS_Spea: homoserine O-succinyltransferase (*Shewanella pealeana*, A8H6B7); Mapk1_mpk1: mitogen-activated protein kinase 1 (*Xenopus laevis*, P26696).

These proteins span a wide range of sequence lengths, molecular functions, and taxonomic origins. The predicted probabilities and rank ordering reveal distinct host preferences for each protein. As expected, proteins of bacterial origin—such as the methionine repressor (METJ_ECO27) and homoserine O-succinyltransferase (metAS_Spea)—showed a strong preference for *E. coli* expression. The mating pheromone Er-23

(MER23_EUPRA), although derived from a unicellular eukaryote, also exhibited a higher predicted probability of expression in *E. coli* than in eukaryotic host systems.

In contrast, the rat sodium-dependent glucose transporter (Mfsd4b_Naglt1) showed a clear preference for yeast and, to a lesser extent, insect cell expression systems. Importantly, the rank ordering also highlights proteins that may be broadly tolerant of multiple hosts. For example, *Xenopus* Mapk1 exhibited relatively low predicted probabilities across all four systems, with only marginal preference for yeast or insect cells over *E. coli* and mammalian hosts.

Other proteins demonstrated much sharper discrimination among hosts. For METJ_ECO27 and Mfsd4b_Naglt1, the dif-

ference between the most and least preferred expression systems reached approximately 18-fold and 16-fold, respectively. Such pronounced differences provide actionable guidance for host selection and enable more informed experimental decisions when balanced against other practical considerations such as cost, scalability, and downstream processing requirements.

4. Discussion

In current practice, the selection of a host cell for recombinant protein expression is largely guided by laboratory experience, available literature, and practical project constraints. Literature searches aimed at identifying prior expression of the protein of interest—or of the closest homolog or ortholog identified by BLASTp [31]—remain the principal decision-making tool. Even when relevant information is available, there is no guarantee that the reported expression host or associated vectors represent the optimal choice. Practical experience has shown that even minor amino acid differences can lead to dramatically different expression outcomes [32].

Existing computational approaches [4–10] in recombinant protein expression primarily address optimization steps that occur after a host has been selected. In contrast, the approach presented here focuses on the initial host selection step by ranking the probability of expression across four commonly used host types, based on empirical evidence accumulated over approximately four decades of practice. This ranking facilitates rational host selection while accounting for additional project-specific considerations and before engaging in downstream optimization. This step is critical, as an appropriate initial host choice can render many optimization efforts unnecessary.

The recombinant protein expression literature is vast and fragmented across diverse disciplines, making comprehensive identification of relevant studies impractical. In contrast, the PDB, which currently contains more than 150,000 entries, provides a rich source of information beyond structural data alone. Of particular relevance is the documentation of expression systems used to produce proteins in quantities sufficient for structural determination—an outcome that implies both functionality and practical feasibility. The PDB is heavily skewed toward *E. coli* expression systems (Table 1) and potentially presents a bias towards overrepresentation of certain predictor variables. It nevertheless contains a sufficient number of entries for insect, yeast, and mammalian hosts to support comparative analysis. Through random selection and cumulative sampling, a representative dataset of 953 proteins with near-equal distribution among the four host types was assembled (Appendix A).

Predictor variables were selected from structural parameters and features related to subcellular localization, post-translational modifications, in vivo stability, and molecular function. While multiple parameters could have been chosen within each category, priority was given to features commonly

suspected by practitioners to influence expression outcomes and those that are well established and readily obtainable from online resources. These included protein length (*r*), isoelectric point (*pI*), GRAVY index (*hp*), and ternary and quaternary structural information encoded using the polychotomous indices *t* and *q* (Table 2).

Subcellular localization was limited to membrane association and coded into three categories: non-membrane-associated proteins and membrane-associated proteins, further subdivided based on the presence or absence of transmembrane helices. Membrane association is well known to influence expression yield and host selection.

Among known post-translational modifications [33], only N-glycosylation was selected as a representative feature. This modification is present across all three domains of life and is known to affect protein solubility and stability. Because not all N-glycosylation motifs (N-X-S/T) are experimentally confirmed and site-specific data are incomplete, the predicted number of potential glycosylation sites [26, 28] was used as a uniform surrogate measure. The final logistic models for mammalian cells and *E. coli* validated this choice. In the *E. coli* model, N-glycosylation (*ng*) was associated with a negative regression coefficient ($\beta = -0.209$) and an odds ratio (OR) < 1 (0.811), indicating reduced expression suitability. In contrast, in the mammalian cell model, *ng* displayed a positive coefficient ($\beta = 0.243$) and an OR > 1 (1.275), indicating enhanced expression suitability. Accordingly, the models favor mammalian cell systems for proteins with high levels of N-glycosylation, a well-established principle in recombinant protein expression.

Assigning molecular function posed challenges due to periodic updates to the Gene Ontology framework [30] and the frequent annotation of multiple functions for individual proteins in UniProtKB. To address this, only the core molecular function was retained. For example, S-adenosylmethionine transferases are annotated with ATP binding, Mg^{2+} binding, and catalytic activity; in this study, such proteins were classified simply as having “catalytic activity” and coded accordingly (Table 3).

Using the selected predictor variables, backward stepwise logistic regression was applied, retaining variables with *p* values ≤ 0.05 . The dataset was incrementally expanded in batches of 200 proteins, with redundant identical sequences removed at each iteration. Model stability was achieved at 953 proteins, beyond which further expansion did not materially alter predictor selection. Host-specific logit models (Table 8) were validated by demonstrating that the average predicted expression probability for randomly selected subsets of the training data did not differ significantly from that of an independent test set drawn from the PDB (Table 9).

These validated models enable estimation of expression probability and rank ordering of host preference for individual protein sequences (Table 10). Distinct and sequence-specific preference patterns were observed across host types. As such,

the logistic models presented here provide a practical, evidence-based tool for host selection in recombinant protein expression, to be used as an initial decision-making step prior to applying existing computational optimization tools [4-10] and conducting experimental work.

Future directions include periodic resampling of the expanding recombinant protein expression landscape, incorporation of additional predictor variables, exploration of alternative machine learning approaches, and extension of the framework to additional host cell types.

5. Conclusion

The present work offers a facile and practical method for predicting the most suitable host cell for the expression of novel protein sequences. This approach is based on regression analyses performed on a curated set of 953 protein sequences with known positive expression outcomes in four commonly used host systems: *Escherichia coli*, insect, mammalian, and yeast cells. Expression outcomes were linked to a defined set of predictor variables encompassing structural parameters, instability index, membrane association status, post-translational modification potential, and molecular functions.

The derived logit models were distinct for each expression host and incorporated unique sets of statistically significant predictor variables. These models were validated and shown to be effective in predicting the probability of expression of protein sequences in specific host systems, as well as in rank-ordering host preferences. Deploying these tools at the initial stage of expression system design has the potential to increase the likelihood of success in laboratory trials while reducing experimental effort, development time, and associated costs. The models and prediction tools described here are publicly accessible at: <http://www.prosciconsulting.com/expred-beta-v2.html>

Abbreviations

PDB	Protein Data Bank
GRAVY	Kyte-Doolittle Grand Average of Hydropathicity
SIB	Swiss Institute of Bioinformatics
SCOP	Structural Classification of Protein
UniProtKB	Universal Protein Knowledgebase
DTU	Technical University of Denmark
GPI	Glycosyl Phospho Inositol
ANOVA	Analysis of Variance
GO	Gene Ontology

Acknowledgments

Vincent Vinh-Hung, Catherine K. Smith and Andrew Prongay critically read the manuscript and provided valuable input.

Author Contributions

Hung Van Le is the sole author. The author read and approved the final manuscript.

Conflicts of Interest

The author declares no conflicts of interest.

Appendix

Appendix I: Training Sets

Randomly selected components of training sets identified by PDB numbers

E. coli: 2MMW, 2MNQ, 1IU5, 3I8Z, 2ZW0, 1A8O, 1CTF, 5E11, 2UZG, 2LWB, 1M9Z, 2FYG, 1VZI, 2D48, 1M48, 3BES, 2E7A, 1EXT, 1TNW, 3SE3, 1UEA, 2QPO, 1YV1, 1QLX, 3P92, 1U5K, 1OK3, 1VK6, 1BR6, 1DCQ, 1L6J, 4ENE, 1QQT, 1SVL, 3FFN, 1MSW, 5TFR, 5B7I, 6YAK, 6YAK, 6YIJ, 6YKG, 6VWC, 7KFL, 6U4U, 6V1A-4, 6V1A-5, 5QOA, 6PLJ, 6PLD, 4QDB, 6Q64, 6Q61, 6U36-2, 6U36_3, 6MVT, 5A1V_1, 5A4J, 5DG3, 6DNV, 5PGI, 6H5V, 5NM8, 4UUS_1, 4UUS_2, 2MJ9, 2MPU, 2KS1_1, 2KS1_2, 2LU3, 3LNV, 4E6A, 4FRF, 4AJ5_1, 4AJ5_2, 4AJ5_3, 3LDN, 3THN, 2QUX, 2P7C, 2IFW, 3H7N, 3GLI, 3GLI_2, 3GLI_3, 3GLI_4, 1M01, 1DVY, 1SAY, 6C71, 6ST4, 6JXV, 6OL8, 6RNM, 5CS6, 5E67, 2DT3, 5DSD, 5E4R, 5DU2, 5E53, 5DTQ, 5DU9, 5DVA, 5DTP, 5E4Q, 5DV8, 5DTE, 5SA9, 4GQC, 3V5H_1, 3V5H_2, 4IFV_1, 4IFV_2, 4EYW, 4FFN, 4M2X, 5T0W, 3P30, 3ZDV, 3K4Y, 3JUJ, 3KF9, 2B4V, 2IXF, 2I82, 3CXZ, 2POD, 2W4Q, 1ZXB, 3B8U, 2KHD, 1YEU_1, 1YEU_2, 1NMS, 1JGS, 1LX8, 1TF5, 1MSF, 1BD9, 1AXM, 1B56, 1J7H, 2R0H, 1G5I, 1JB3, 1X7X_1, 1X7X_2, 2Z8W_1, 2Z8W, 1RXT, 2AVS, 2C95, 2R2C, 2IPI, 3CRQ, 3D79, 2O93, 3CDV, 2LH8, 2XSR, 3HGV, 3KN2, 3NXS, 3NQA, 4DGZ, 2YEE, 4FI3, 4FI3_2, 4FI3_3, 4FOU_1, 4FOU_2, 4K0O, 3TSU, 6PG9, 6AIX, 6AIG, 6U2H_1, 6U2H_2, 6UGP, 5V6H_1, 5V6H_2, 6H5M, 6C1X, 5I5R, 5OO3, 6SOD, 6H2J, 6GV2, 5IAX, 5H11, 4XMD, 6SOR, 6JMD, 6KA2, 6KJ8_1, 6KJ8_2, 6KPH, 5LHM, 6HF4, 6H74_1, 6H74_2, 6EC7, 6EEJ, 5LBM, 6EHV, 5LFA, 5HT6, 5A3O, 5TNB, 6IYY, 5JJM, 6FUA_1, 6FUA_2, 4D4W, 6C37, 5LGW, 3UQA, 4FXV, 4LOB, 2L0I, 3RYI, 3RYI, 3NZ6, 3QS1, 2XBQ, 2VD1, 4GQL, 3U43, 3U43, 3TKP, 4IRP, 3GMV, 2BVZ, 2GB4, 1W6L, 2OMU_1, 2OMU_2, 3ETM, 2V2M, 2FL3, 2WBR, 3GSO_1, 1EI5, 1S2V, 1KAF, 1NIO, 1SNG, 1CKJ, 1KBC.

Insect: 1T50, 5MPO, 2VSD, 1Q8D, 1VYI, 3TJQ, 1Z92, 2HEW, 5MPO, 3SE3, 3N7O, 2ANW, 4H14, 3PE6, 3UG9, 5TIH, 2OU7, 1FO8, 3KJ6, 2V5M, 3SE3, 4GGA, 5UIG, 5X2M, 4O9R, 5X2M, 5T1A, 3NSJ, 4TNB, 4KRO, 1CK7, 1NUF, 3FVY, 2PGG, 5M05, 4KX7, 4XE0, 5UE8, 5ME3, 3S4Z, 6SA5, 6Q7E, 6TOU, 6S6Q, 6JOL, 6XY7, 6LKO, 6VHG, 6V2W_2, 6TQ4, 6U8Z, 7CRC_1, 7CRC_2, 7D4F_1,

7D4F_2, 7D4F_3, 6PZD, 4RWS_1, 4RWS_2, 4ZG7, 5CHT, 5D9A_1, 5D9A_2, 5D9A_3, 6DBO_1, 6DBO_2, 5O1O, 5TCD, 6N52, 6C4D, 5AFM_1, 6C7I, 6AKY, 4G56_1, 4G56_2, 4MRO, 4BTJ, 4KC3_2, 4CI8, 3PIX, 4Q7X, 4PXW, 2ZVA, 3B2U_3, 1YAJ, 1ZCB, 1ZRZ, 1JK8_1, 1JK8_2, 1N52_1, 1N52_2, 1AYU, 1IEA_1, 6K42_1, 6K42_2, 6K42_3, 6K42_4, 6K42_5, 6K42_6, 6K42_7, 6K42_8, 6K42_9, 6K42_10, 6K42_11, 6K42_12, 6K42_13, 6K42_14, 6K42_15, 6K42_16, 6K42_17, 6K42_18, 6K42_19, 6K42_20, 6K42_21, 6K42_22, 6K42_23, 6K42_24, 6K42_25, 6K42_26, 6K42_27, 6K42_28, 6K42_29, 6K42_30, 6K42_31, 6K42_32, 6K42_33, 6K42_34, 6K42_35, 6K42_36, 6K42_37, 6K42_38, 6K42_39, 6K42_40, 6K42_41, 6K42_42, 6K42_43, 6K42_44, 6K42_45, 6K42_46, 6K42_47, 6K42_48, 6K42_49, 6K42_50, 6K42_51, 6K42_52, 6K42_53, 6K42_54, 6K42_55, 6K42_56, 6K42_57, 6K42_58, 6K42_59, 6K42_60, 6K42_61, 6K42_62, 6K42_63, 6K42_64, 6K42_65, 6K42_66, 6K42_67, 6K42_68, 6K42_69, 6K42_70, 6K42_71, 6K42_72, 6K42_73, 6K42_74, 6K42_75, 6K42_76, 6K42_77, 6K42_78, 6K42_79, 6K42_80, 6K42_81, 6K42_82, 6K42_83, 6K42_84, 6K42_85, 6K42_86, 6K42_87, 6K42_88, 6K42_89, 6K42_90, 6K42_91, 6K42_92, 6K42_93, 6K42_94, 6K42_95, 6K42_96, 6K42_97, 6K42_98, 6K42_99, 6K42_100, 6K42_101, 6K42_102, 6K42_103, 6K42_104, 6K42_105, 6K42_106, 6K42_107, 6K42_108, 6K42_109, 6K42_110, 6K42_111, 6K42_112, 6K42_113, 6K42_114, 6K42_115, 6K42_116, 6K42_117, 6K42_118, 6K42_119, 6K42_120, 6K42_121, 6K42_122, 6K42_123, 6K42_124, 6K42_125, 6K42_126, 6K42_127, 6K42_128, 6K42_129, 6K42_130, 6K42_131, 6K42_132, 6K42_133, 6K42_134, 6K42_135, 6K42_136, 6K42_137, 6K42_138, 6K42_139, 6K42_140, 6K42_141, 6K42_142, 6K42_143, 6K42_144, 6K42_145, 6K42_146, 6K42_147, 6K42_148, 6K42_149, 6K42_150, 6K42_151, 6K42_152, 6K42_153, 6K42_154, 6K42_155, 6K42_156, 6K42_157, 6K42_158, 6K42_159, 6K42_160, 6K42_161, 6K42_162, 6K42_163, 6K42_164, 6K42_165, 6K42_166, 6K42_167, 6K42_168, 6K42_169, 6K42_170, 6K42_171, 6K42_172, 6K42_173, 6K42_174, 6K42_175, 6K42_176, 6K42_177, 6K42_178, 6K42_179, 6K42_180, 6K42_181, 6K42_182, 6K42_183, 6K42_184, 6K42_185, 6K42_186, 6K42_187, 6K42_188, 6K42_189, 6K42_190, 6K42_191, 6K42_192, 6K42_193, 6K42_194, 6K42_195, 6K42_196, 6K42_197, 6K42_198, 6K42_199, 6K42_200, 6K42_201, 6K42_202, 6K42_203, 6K42_204, 6K42_205, 6K42_206, 6K42_207, 6K42_208, 6K42_209, 6K42_210, 6K42_211, 6K42_212, 6K42_213, 6K42_214, 6K42_215, 6K42_216, 6K42_217, 6K42_218, 6K42_219, 6K42_220, 6K42_221, 6K42_222, 6K42_223, 6K42_224, 6K42_225, 6K42_226, 6K42_227, 6K42_228, 6K42_229, 6K42_230, 6K42_231, 6K42_232, 6K42_233, 6K42_234, 6K42_235, 6K42_236, 6K42_237, 6K42_238, 6K42_239, 6K42_240, 6K42_241, 6K42_242, 6K42_243, 6K42_244, 6K42_245, 6K42_246, 6K42_247, 6K42_248, 6K42_249, 6K42_250, 6K42_251, 6K42_252, 6K42_253, 6K42_254, 6K42_255, 6K42_256, 6K42_257, 6K42_258, 6K42_259, 6K42_260, 6K42_261, 6K42_262, 6K42_263, 6K42_264, 6K42_265, 6K42_266, 6K42_267, 6K42_268, 6K42_269, 6K42_270, 6K42_271, 6K42_272, 6K42_273, 6K42_274, 6K42_275, 6K42_276, 6K42_277, 6K42_278, 6K42_279, 6K42_280, 6K42_281, 6K42_282, 6K42_283, 6K42_284, 6K42_285, 6K42_286, 6K42_287, 6K42_288, 6K42_289, 6K42_290, 6K42_291, 6K42_292, 6K42_293, 6K42_294, 6K42_295, 6K42_296, 6K42_297, 6K42_298, 6K42_299, 6K42_300, 6K42_301, 6K42_302, 6K42_303, 6K42_304, 6K42_305, 6K42_306, 6K42_307, 6K42_308, 6K42_309, 6K42_310, 6K42_311, 6K42_312, 6K42_313, 6K42_314, 6K42_315, 6K42_316, 6K42_317, 6K42_318, 6K42_319, 6K42_320, 6K42_321, 6K42_322, 6K42_323, 6K42_324, 6K42_325, 6K42_326, 6K42_327, 6K42_328, 6K42_329, 6K42_330, 6K42_331, 6K42_332, 6K42_333, 6K42_334, 6K42_335, 6K42_336, 6K42_337, 6K42_338, 6K42_339, 6K42_340, 6K42_341, 6K42_342, 6K42_343, 6K42_344, 6K42_345, 6K42_346, 6K42_347, 6K42_348, 6K42_349, 6K42_350, 6K42_351, 6K42_352, 6K42_353, 6K42_354, 6K42_355, 6K42_356, 6K42_357, 6K42_358, 6K42_359, 6K42_360, 6K42_361, 6K42_362, 6K42_363, 6K42_364, 6K42_365, 6K42_366, 6K42_367, 6K42_368, 6K42_369, 6K42_370, 6K42_371, 6K42_372, 6K42_373, 6K42_374, 6K42_375, 6K42_376, 6K42_377, 6K42_378, 6K42_379, 6K42_380, 6K42_381, 6K42_382, 6K42_383, 6K42_384, 6K42_385, 6K42_386, 6K42_387, 6K42_388, 6K42_389, 6K42_390, 6K42_391, 6K42_392, 6K42_393, 6K42_394, 6K42_395, 6K4

Yeast: 1SS3, 1TCP, 5HPG, 3S64, 1BXM, 1SM7, 1GD6, 5T89, 2ZIB, 2KVA, 1XX9, 2FIN, 4WFE, 3FJU, 2O6X, 3SI1, 4N2Z, 3ZY2, 1H8L, 3N9K, 3BSG, 1WD3, 2XQR, 5GHK, 4AA1, 2VZ1, 6F7H, 5ED1, 6RW3, 5V5Z, 6SP2, 6Q81, 1E9T, 1EK6, 1EQC, 6WHG, 6Z3Y, 7KAQ_1, 7KAQ_2, 7KAQ_3, 7KAQ_4, 7KAQ_5, 7KAQ_6, 7KY8_1, 7KY8_2, 6RB2, 6UCV_1, 6UCV_2, 6UCV_3, 6UCV_4, 6UCV_5, 4QVM_1, 4QVM_2, 4QVM_3, 4QVM_4, 4QVM_5, 4QVM_6, 4QVM_7, 4QVM_8, 4QVM_9, 4QVM_10, 4QVM_11, 4QVM_12, 4QVM_13, 4QVM_14, 6FUF_1, 6FUF_2, 5VNY, 5XF8_1, 5XF8_2, 5XF8_3, 5XF8_4, 5XF8_5, 5XF8_6, 5XF8_7, 4ZR1, 6BM4_1, 6BM4_2, 6BM4_3, 6BM4_4, 6BM4_5, 6BM4_6, 6BM4_7, 6BM4_8, 6BM4_9, 6BM4_10, 6T8F, 6LCR_1, 6LCR_2, 6VFF, 4YB9, 5CO6, 5ESF, 6DZ7, 5VLJ_1, 5VLJ_2, 6GSA_1, 6GSA_2, 6GSA_3, 6GSA_4, 5KC2_1, 5KC2_2, 5FUF, 6P2R_1, 6P2R_2, 6DMU, 6I6J_1, 5M8C, 5V7V, 4ZRG, 6ROJ_1, 6ROJ_2, 3QVN, 4I5U, 3PLG, 2WfV, 3H2P, 2XA3, 2C32, 2VU8_1, 2FLP_3, 1X9Q, 1MM0, 1EX0, 1FTZ, 1GFT, 1GYR, 1J14, 1E78, 2DAA, 1BX7, 1CTE, 2IFF, 1KNT, 1DIX, 4WIS, 6PGM, 1QPG, 1QRK, 1CT5, 1A1S, 1LDT, 1A9W_1, 1A9W_2, 1FW8, 1R1J, 1GB5, 1H0K, 1J16, 1I8N, 1CC0_1, 1CC0_2, 1K9Z, 3CQQ, 1ZPU, 3UGX, 3W3X_1, 4K3G, 4AI6, 4A6K, 4EV5, 4C9Q, 6P25_1, 6SPEK, 6ROJ_1, 6E1K, 6FE8_3, 6U8A, 6H5Y, 6AFZ, 5AEZ, 5C1E, 5M5G_1, 5M5G_2, 5I6C, 5IFP, 6S4I, 4ZDY, 5SYT, 6PSY_1, 6PSY_2, 6CX0, 5WFD_2, 5OJS, 2WYT, 2LLI, 4AI6, 3OLE, 3OB8, 4AKJ, 2OKJ, 3FP7, 2IWH, 1GFK, 1OF6, 1N9G, 1OXM, 2WFJ, 3OEE_1, 3OEE_2, 3OEE_3, 3OEE_4,

3OEE_5, 1SJX, 1FMI, 4YBQ, 3O8O, 3O8O, 1DOR, 4A01, 2DKH, 3H8C, 1FBR, 5V6P, 4WJS, 1UL9.

Mammalian: 1ERG, 3T8X, 1BQH, 5EN2, 1UEA, 1PEX, 1LOY, 1SBB, 2ASU, 1KB5, 1FGX, 3T8X, 5LGJ, 3VI3, 3ZE2, 3S88, 3ZE2, 4XNU, 3V4V, 3HN3, 3VI3, 5NJ3, 4DB1, 4ZXB, 6V7M, 6PE8, 6HGA, 6USC, 1PV7, 6UMX, 1IJQ, 1IJY, 1IMV, 1P53, 1P8J, 6WGB_1, 6WGB_2, 6VWG_1, 7DPM_3, 7DPM_1, 7DPM_2, 7DUO_2, 7DUO_1, 7DUO_3, 7KMG_1, 7KMG_2, 7AH1, 5VSI_1, 5VSI_2, 6CH8_1, 6CH8_4, 6GL7_2, 6GL7_3, 5FMV, 4ZMA, 4YCI, 5HYS_1, 5HYS_2, 5A2E, 5A2F, 3ZHG_2, 4AKM, 3R08_3, 4EWS, 2JJW, 2DRU, 2V5N, 3IG_1, 3IG_2, 1DX5_2, 1T7W, 1N26, 1AU1, 3PP3_1, 3PP3_2, 3T3M_3, 3T3M_4, 4WUU, 5FN8, 5FCS_1, 5FCS_2, 4Z2A, 2UZX_2, 2VSC, 4F2M, 6U2F, 6TT1, 6VRT_1, 6VRT_2, 6XTA, 4R90_1, 4R90_2, 5KU6, 5WNA_1, 5WNA_2, 5KZW, 6IAS_1, 6IAS_2, 6IBT, 6OKP1_1, 6OKP_2, 6OKP_3, 6OKP_4, 6OKP_5, 6OKP_6, 6OML, 6UVO_1, 6UVO_2, 6EYJ, 5ANM_1, 5ANM_2, 5ANM_3, 5VH4_1, 5VH4_2, 5VH5, 4JDA, 2XFK, 4OGA_5, 3KWZ, 3REZ, 3RJR, 2NYY, 3E6P, 3HH2_1, 3HH2_2, 3EDY, 11N8Z, 1OYH, 1RF0_1, 1RF0_2, 1RF0_3, 1I8L_1, 1I8L_2, 1J89, 1F42, 1PHM, 1CKL, 1CDU, 1EWF, 1OLZ, 1O86, 1M4U, 1JQF, 2F61, 2VXQ_2, 2VXQ_3, 2WNG, 1ZKZ, 2AM4, 1YIP, 1W0Y, 2V10, 2CF9, 6EHO, 4PO7, 3RM9, 2XMD, 2LDB_3, 2LDB_2, 4BDV_1, 4BDV_2, 4F5C_1, 4F5C_2, 4CSY, 3U7U, 5BO1_2, 5BO1_3, 5ALC_1, 5ALC_2, 5JNC, 5K33_1, 5K33_2, 6GSI_2, 5U6A_1, 5U6A_2, 5VR9_1, 5VR9_2, 5VQP, 6OMO, 6QB3_2, 6UVO_1, 6UVO_2, 6VRT_1, 6VRT_2, 6SRX_2, 6SRX_1, 6EDU_3, 3O9M, 2HWL, 1FRT_1, 1FRT_2, 6O9H, 6MF0, 3K71, 1SZH, 1MMP, 4LEO_3, 6LIQ_1, 1OQO_1, 1OQO_2, 5LY6, 2VNN, 1LTJ_3, 3NK3, 5VH4_1, 5VH4_2, 6MNF_2, 6MNF_1, 4BC0, 5XJE_2, 5XJE_1, 2V11, 6PYC_1, 6PYC_2, 2X8B, 5DTF_1, 5DTF_2, 3N2Z, 6DJP_1, 6DJP_2, 6SS2, 6U2F, 6U6U, 6UE7_1, 6UE7_2, 6UE7_3.

Randomly selected components of training sets identified by UniprotKB numbers:

E. coli: Q75NH7, P09603, P03069, P07951, P45379-6, P00797, P01009, P00970, P04535, P09922.

Mammalian: P60568, O90495, P062133V4V.

Appendix II: Validation Sets

Randomly selected components of training subsets spared for cross validation studies

E. coli: Q75NH7, 1TNW, P03069, 5A4J, 2KSI_2, 3H7N, 3GLI_4, 3ZDV, 2POD, 1ZXB, 1B56, 1J7H, 1X7X_2, 2IPI, 3NQA, 4DGZ, 4FI3_1, 5V6H_1, 5IAX, 5H11, 6C37, 3QSI, 3U43, 4GOL, 1S2V.

Insect: 1Q8D, 4H14, 3UG9, 6Q7E, 5CHT, 4PXW, 1ZCB, 6THX, 6U5O_1, 6CX9_2, 5FZF, 3L9L_1, 1LO6, 1INP, 4RJ5, 4GRW_1, 4KBA, 3JBY_2, 5FDX, 6S1L, 4YVC, 6ACJ_2, 3OPM, 4LL0, 2O6S.

Yeast: 5HPG, 1XX9, 4WFE, 3BSG, 2XQR, 6F7H, 7KY8_2, 6UCV 2, 4QVM 6, 4QVM 13, 6BM4 1, 6BM4 7,

6BM4_8, 6BM4_9, 6VFF, 5VLJ_1, 6GSA_2, 5V7V, 3H2P, 1FTZ, 1GFT, 1E78, 1CT5, 1A1S, 1FW8.

Mammalian: 3T8X, Q90495, P06213, 7DPM_2, 3I9G_2, 3T3M_3, 3T3M_4, 6UVO_1, 5ANM_3, 1RF0_1, 1RF0_2, 1RF0_3, 4F5C_1, 4F5C_2, 5BO1_3, 6GSI_2, 5U6A_2, 5VR9_15VQP, 6VRT_2, 1OQO_1, 1OQO_2, 6MNF_1, 5DTF_2, 6U2F.

Randomly selected components of test sets selected from the PDB for cross validation studies

E. coli: 1P09, 1NCT, 1H7Q, 1ZUH, 4XZD, 1QSA, 1IJG, 2E7J, 2LX3, 4ZXS_1, 4ZXS_2, 1K0K, 2G0R, 5D45, 1K6O_3, 1K6O_4, 2HBT, 3LGW, 5ET0_1, 5ET0_2, 1L3L, 2PVG_1, 2PVG_2, 2PVG_3, 3MA9_1.

Insect: 1Q5K, 2YIY, 4MYW, 4Y5X_1, 4Y5X_2, 4Y5X_3, 6BKO_1, 6BKO_2, 6E7R_1, 6E7R_2, 6P7V_1, 6P7V_2, 6P7V_3, 6QWS, 7KN6_1, 7KNT_1, 7KNT_2, 2H2T, 3EU7, 3K7W, 4IB4, 4TYE, 4UXI, 6CZS, 5JFW.

Yeast: 1B7N, 1BX8, 1EAX, 1GAZ, 1H6M, 1I3D, 1UXL, 2E3F, 2J7O, 3L4K, 4BKE, 4FM9, 4QV3, 4W8F, 6ND1, 6QG3_1, 6QG3_2, 6QG3_3, 6QG3_4, 6QG3_5, 6QG3_6, 6QG3_7, 6QG3_8, 6T9J, 7OP1.

Mammalian: 1OAK_1, 1OAK_2, 1JMY, 3QRG_1, 3QRG_2, 4C8F, 2VXT_1, 2VXT_2, 2VDP_2, 4FTN, 2W0F_1, 2W0F_2, 3U9U_3, 4J1U_1, 4J1U_2, 4X96, 3QG7_1, 3QG7_2, 5EGH, 4X4M_2, 5L7I, 4LU5_2, 4LU5_3, 5UJZ_3, 6HYG_1.

References

- [1] Cohen SN, Chang ACY, Boyer HW, Helling RB, Construction of biologically functional bacterial plasmids in vitro, *Proc Natl Acad Sci USA* 70: 3240-3244, 1973. <https://doi.org/10.1073/pnas.70.11.3240>
- [2] Chang ACY, Cohen SN, Genome construction between bacterial species in vitro: Replication and expression of *Staphylococcus* plasmid genes in *Escherichia coli*, *Proc Natl Acad Sci USA* 71: 1030-1034, 1974. <https://doi.org/10.1073/pnas.71.4.1030>
- [3] Morrow JF, Cohen SN, Chang ACY, Boyer HW, Goodman HM, Helling RB, Replication and transcription of eukaryotic DNA in *Escherichia coli*, *Proc Natl Acad Sci USA* 71: 1743-1747, 1974. <https://doi.org/10.1073/pnas.71.5.1743>
- [4] Chang CCH, Song J, Tey BT, Ramanan RN, Bioinformatics approaches for improved recombinant protein production in *Escherichia coli*: protein solubility prediction, *Brief Bioinform* 15: 953-962, 2013. <https://doi.org/10.1093/bib/bbt057>
- [5] Habibi N, Hashim SZM, Norouzi A, Samian MR, A review of machine learning methods to predict the solubility of overexpressed recombinant proteins in *Escherichia coli*, *BMC Bioinform* 15: 134, 2014. <https://doi.org/10.1186/1471-2105-15-134>
- [6] Bhandari BK, Gardner PP, Lim CS, Solubility-Weighted Index: fast and accurate prediction of protein solubility, *Bioinform* 36: 4691-4698, 2020. <https://doi.org/10.1093/bioinformatics/btaa578>
- [7] Chan W-C, Liang P-H, Shih Y-P, Yang U-C, Lin C-C, Hsu C-N, Learning to predict expression efficacy of vectors in recombinant protein production. 2010. *BMC Bioinform* 11 (Suppl 1): S21, 2010. <https://doi.org/10.1186/1471-2105-11-S1-S21>
- [8] Martiny H-M, Armenteros JJA, Johansen AR, Salomon J, Nielsen H, Deep protein representations enable recombinant protein expression prediction. *Comput Biol Chem* 95: 107596, 2021. <https://doi.org/10.1016/j.compbiolchem.2021.107596>
- [9] Bhandari BK, Lim CS, Gardner PP, TISIGNER.com: web services for improving recombinant protein production, *Nucleic Acids Res* 49: W654-W661, 2021. <https://doi.org/10.1093/nar/gkab175>
- [10] Bhandari BK, Lim CS, Remus DM, Chen A, van Dolleweerd C, Gardner PP, Analysis of 11,430 recombinant protein production experiments reveals that protein yield is tunable by synonymous codon changes of translation initiation sites, *PLoS Comput Biol* 17: e1009461, 2021. <https://doi.org/10.1371/journal.pcbi.1009461>
- [11] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE, The Protein Data Bank, *Nucleic Acids Res* 28: 235-242, 2000. <https://doi.org/10.1093/nar/28.1.235>
- [12] Hosmer DW, Lemeshow S, *Applied Logistic Regression*, John Wiley & Sons, New York, 2000.
- [13] Kyte J, Doolittle RF, A simple method for displaying the hydropathic character of a protein, *J Mol Biol* 157: 103-132, 1982. [https://doi.org/10.1016/0022-2836\(82\)90515-0](https://doi.org/10.1016/0022-2836(82)90515-0)
- [14] Guruprasad K, Bhasker Reddy BV, Pandit MW, 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* 4: 155-161, 1990. <https://doi.org/10.1093/protein/4.2.155>
- [15] Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A, Protein Identification and Analysis Tools on the ExPASy Server, in Walker JM (ed.), *The Proteomics Protocols Handbook*, Humana Press, Totowa, NJ, pp 571-607, 2005. <https://doi.org/10.1385/1-59259-890-0:571>
- [16] Sehnal D, Bittrich S, Deshpande M, Svobodová R, K. Berka K, V. Bazgier V, S. Velankar S, S. K. Burley SK, J. Koča J, A. S. Rose AJ, Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures, *Nucleic Acids Res* 49 (W1): W431-W437, 2021. <https://doi.org/10.1093/nar/gkab314>
- [17] Jumper J et al., Highly accurate protein structure prediction with AlphaFold, *Nature* 596: 583-589, 2021. <https://doi.org/10.1038/s41586-021-03819-2>
- [18] The UniProt Consortium, UniProt: the universal protein knowledgebase in 2021, *Nucleic Acids Res* 49: D480-D489, 2021. <https://doi.org/10.1093/nar/gkaa1100>
- [19] Krogh A, Larsson B, von Heijne G, Sonnhammer ELL, Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes, *J Mol Biol* 305: 567-580, 2001. <https://doi.org/10.1006/jmbi.2000.4315>

- [20] Sonnhammer ELL, von Heijne G, Krogh A, A hidden Markov model for predicting transmembrane helices in protein sequences, in Glasgow J, Littlejohn T, Major F, Lathrop R, Sankoff D, Sensen C (eds.) Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology, AAAI Press, Menlo Park, CA, pp 175-182, 1998.
- [21] Käll L, Krogh A, Sonnhammer ELL, A Combined Transmembrane Topology and Signal Peptide Prediction Method, J Mol Biol 338: 1027-1036, 2004.
<https://doi.org/10.1016/j.jmb.2004.03.016>
- [22] Käll L, Krogh A, Sonnhammer ELL, Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server, Nucleic Acids Res 35: W429-W432, 2007.
<https://doi.org/10.1093/nar/gkm256>
- [23] Xie Y, Zheng, Y, Li, H et al., GPS-Lipid: a robust tool for the prediction of multiple lipid modification sites, Sci Rep 6: 28249, 2016. <https://doi.org/10.1038/srep28249>
- [24] Eisenhaber B, Bork P, Eisenhaber F, Prediction of potential GPI-modification sites in proprotein sequences, J Mol Biol 292: 741-758, 1999. <https://doi.org/10.1006/jmbi.1999.3069>
- [25] Armenteros JJA, Sønderby CK, Sønderby SK, Nielsen H, Winther O, DeepLoc: prediction of protein subcellular localization using deep learning, Bioinform 33: 3387-3395, 2017.
<https://doi.org/10.1093/bioinformatics/btx431>
- [26] Gupta R, Brunak S, Prediction of glycosylation across the human proteome and the correlation to protein function, Pac Symp Biocomput 7: 310-322, 2001.
https://doi.org/10.1142/9789812799623_0029
- [27] Murzin AG, Brenner SE, Hubbard T, Chothia C, SCOP: a structural classification of proteins database for the investigation of sequences and structures, J Mol Biol 247: 536-540, 1995.
[https://doi.org/10.1016/S0022-2836\(05\)80134-2](https://doi.org/10.1016/S0022-2836(05)80134-2)
- [28] Welply JK, Shenbagamurthi P, Lennarz WJ, Naid F, Substrate recognition by oligosaccharyltransferase: Studies on glycosylation of modified asn-x-thr/ser tripeptides, J Biol Chem 258: 11856-11863, 1983.
[https://doi.org/10.1016/S0021-9258\(17\)44311-0](https://doi.org/10.1016/S0021-9258(17)44311-0)
- [29] Ashburner et al., Gene ontology: tool for the unification of biology, Nat Genet 25: 25-29, 2000.
<https://doi.org/10.1038/75556>
- [30] The Gene Ontology Consortium, The Gene Ontology resource: enriching a GOLD mine, Nucleic Acids Res 49: D325-D334, 2021. <https://doi.org/10.1093/nar/gkaa1113>
- [31] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool, J Mol Biol 215: 403-410, 1990.
[https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)
- [32] Price II WN et al., Large-scale experimental studies show unexpected amino acid effects on protein expression and solubility in vivo in E. coli, Microb Inform Exp 1: 6, 2011.
<https://doi.org/10.1186/2042-5783-1-6>
- [33] Huang KY, Su MG, Kao HJ, Hsieh YC, Jhong JH, Cheng KH, Huang HD, Lee TY, dbPTM 2016: 10-year anniversary of a resource for post-translational modification of proteins, Nucleic Acids Res 44: D435-D446, 2016.
<https://doi.org/10.1093/nar/gkv1240>