Trends in Biotechnology

CellPress OPEN ACCESS

Feature Review

Can protein expression be ‘solved’?

Catherine Baranowski [1], Hector Garcia Martin [2,3,4,13], Diego A. Oyarzún [5,6,13], Aviv Spinner [1,13], Bijoy Desai [1], Christopher J. Petzold [2,3,4], Evangelos-Marios Nikolados [7], Sebastian Jaaks-Kraatz [8], Aljaž Gaber [9], Robert J. Chalkley [10], Devin Scannell [11], Rachel Sevey [1], Michael C. Jewett [12], Peter J. Kelly [1], and Erika A. DeBenedictis [1,*]

Recombinant protein expression is central to biotechnology’s application. However, not all proteins can be expressed in all organisms, and, given the vast experimental space, it can be challenging to identify the conditions that will yield successful protein expression. The field lacks a predictive model of soluble protein expression that could replace laborious experimental trial and error. Here, we discuss the state of the field and identify the lack of large, high-fidelity datasets as the primary bottleneck to progress. We outline a proposed path toward an extensible experimental platform for collecting soluble overexpression data across organisms. We suggest that the resulting data should be used to train predictive models of protein expression toward answering the question: can protein expression be solved?

Highlights

Heterologous protein expression is a fundamental technique used frequently in modern day biology. It enables scientific exploration of protein function as well as development of lifesaving medicines and economically impactful industrial products.

Protein expression experiments primarily remain an experience-guided trial and error situation, even though it is an approach used by nearly all biologists.

Generating an openly available, large-scale protein expression dataset that spans organisms and uses a standard experimental approach would provide the machine learning community with a foundation for building a multispecies predictive model of expression.

A predictive model of protein expression would have a profound commercial impact and could replace countless hours of experimentation with a higher-probability directed approach.

Why do we need a predictive model for protein expression?

AlphaFold has ushered in a new era in biology in which predictive models complement or abrogate the need for time-consuming and expensive laboratory work in protein structure elucidation. Predicting a protein structure from sequence has generated successes across the industry – from basic scientists who can now study proteins on the basis of structural rather than sequence similarity to companies that can take new approaches to drug design aided by accurate *in silico* structure prediction. The huge impact of AlphaFold2 has inspired a search for the next prediction task that is feasible with today’s **machine learning (ML)** (see Glossary; see also https://www.nnlm.gov/guides/data-glossary/machine-learning) technology and that is equally impactful to industry.

Soluble protein expression impacts all corners of the scientific community from basic scientific discovery and protein engineering to biomanufacturing. In this review, we are using the term ‘soluble protein expression’ to generally describe the amount of protein in the soluble fraction of the cell lysate rather than the amount of a specific protein in a ‘pure’ solution after rigorous purification and quality assessment. There is a tremendous opportunity to save significant amounts of time and money by increasing the success rate of soluble protein expression, an endeavor that frequently leads to failure. Below, we highlight a few areas where a predictive model of protein expression would be game-changing (Box 1). We then dive into factors that impact protein expression. Following this, we summarize existing protein expression datasets and models generated from these data. Next, we outline both an experimental approach and an ML strategy that could be used to generate a predictive model. Last, we discuss major hurdles and a path for dataset growth beyond the first steps suggested in this review.

Factors that impact soluble protein expression

A major challenge in studying protein expression is that there are many steps that impact the probability of successful soluble protein expression in an **expression host**

[1] The Align Foundation, Covina, CA, USA
[2] DOE Agile BioFoundry, Emeryville, CA, USA
[3] The Joint BioEnergy Institute, Lawrence Berkeley National Laboratory, Emeryville, CA, USA
[4] Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[5] School of Informatics, University of Edinburgh, Edinburgh, UK
[6] School of Biological Sciences, University of Edinburgh, Edinburgh, UK
[7] Myria Biosciences AG, Basel, Switzerland
[8] Friedrich Miescher Institute for Biomedical Research (FMI), Basel, Switzerland
[9] Faculty of Chemistry and Chemical Technology, University of Ljubljana, Ljubljana, Slovenia

Box 1. Research areas impacted by protein expression

Basic scientific research

Soluble protein expression is necessary in basic research to probe structure and function and to develop new tools. One example is the discovery and subsequent development of the CRISPR/Cas system. It took researchers decades after the CRISPR locus was first discovered before *S. thermophilus* CRISPR/Cas could be heterologously expressed and was functional in *E. coli* [55]. Expression of mutant Cas9 abrogated specific activities, and it was discovered that Cas9 could be directed to a specific site for action [56,57]. Shortly thereafter, heterologous expression of the system was used for genome editing in mammalian cells [58,59]. This very brief history of the CRISPR/Cas system highlights the impact of protein expression in heterologous systems.

Protein engineering

Protein expression is a major bottleneck for *de novo* protein investigations. Failure to express a designed protein can make validation and further study difficult or impossible. Many studies have been conducted to improve factors that impact expression of designed proteins. Examples of this include high-throughput methods for synthesizing and screening libraries of *de novo* protein domain stability [38,60]. Other protein engineering techniques, such as **directed evolution**, are similarly fundamentally limited by the need to retain intrinsic protein properties, which limits the speed with which proteins can be diversified [61–63].

Protein engineering can also use diverse sequences found in metagenomic samples, such as from the relatively unexplored '**microbial dark matter**' [64–66]. Although sequencing and analysis technologies have allowed discovery of new protein families, expression of these genes remains a challenge and prevents efficient exploration of this vast protein space [67,68].

Biomanufacturing and pharmaceuticals

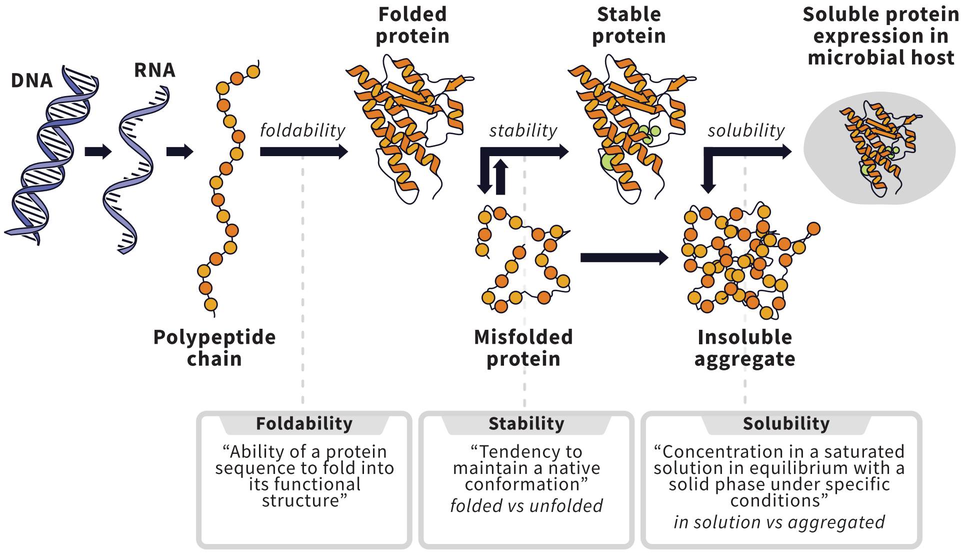
Soluble protein expression in a heterologous host is an important avenue to generating enough protein for downstream investigation or for large-scale production for industrial purposes such as pharmaceuticals (insulin, Herceptin, monoclonal antibodies), industrial enzymes (carbohydrases, proteases), food proteins [69], and polymeric materials [70]. Purification of soluble protein from a microbial system is less expensive and less time-consuming than other methods. If a protein is insoluble, recovery of usable protein and process efficiency are significantly reduced, impacting economic feasibility of production [71,72]. Although the first FDA-approved recombinant protein therapeutic was introduced decades ago, achieving industrial production-ready titers still requires millions of dollars of investment along with substantial strain engineering and process development efforts [73]. Despite the challenges, recombinant proteins and protein drugs are a growing and valuable industry, valued at $2.8 billion in 2022 [72] with revenue of over $550 million in 2021 [73]. Unfortunately, projects are shelved if the protein cannot be expressed at adequate titers because of issues with scale-up and can lead to eventual abandonment of assets.

(Figure 1). Although discussing the steps involved in protein synthesis from nucleic acid through functional folded protein are beyond the scope of this review, we discuss a few specific examples as they pertain to common experimental variables in protein expression experiments. Broadly, the factors that impact expression can be divided into two categories: intrinsic and extrinsic factors [1,2]. Intrinsic characteristics of a protein are fundamentally determined by the amino acid sequence of the protein. The sequence determines post-translational modifications, necessity of cofactors, and formation of disulfide bonds. These in turn impact properties such as the isoelectric point, surface charge, and hydrophobicity. The cascade of intrinsic information starting at the sequence funnels to broader outcomes in heterologous systems such as **foldability** [3], **solubility** [1,2], and **stability** [4] (Figure 1). Meanwhile, extrinsic factors represent variables determined by experimental choices in how to express the protein – host organism, host genotype, **expression cassette architecture**, codon choice, coexpression of chaperones and foldases, media composition, cultivation and lysis conditions, and so forth – that can be modified in the hope of better results (Figure 2).

We suggest that a dataset foundation for a predictive model must explore both extrinsic (e.g., different expression hosts) and intrinsic factors (e.g., different amino acid sequences). Ideally, the impact of both intrinsic and extrinsic factors could be predicted using ML models that are trained on the resulting dataset.

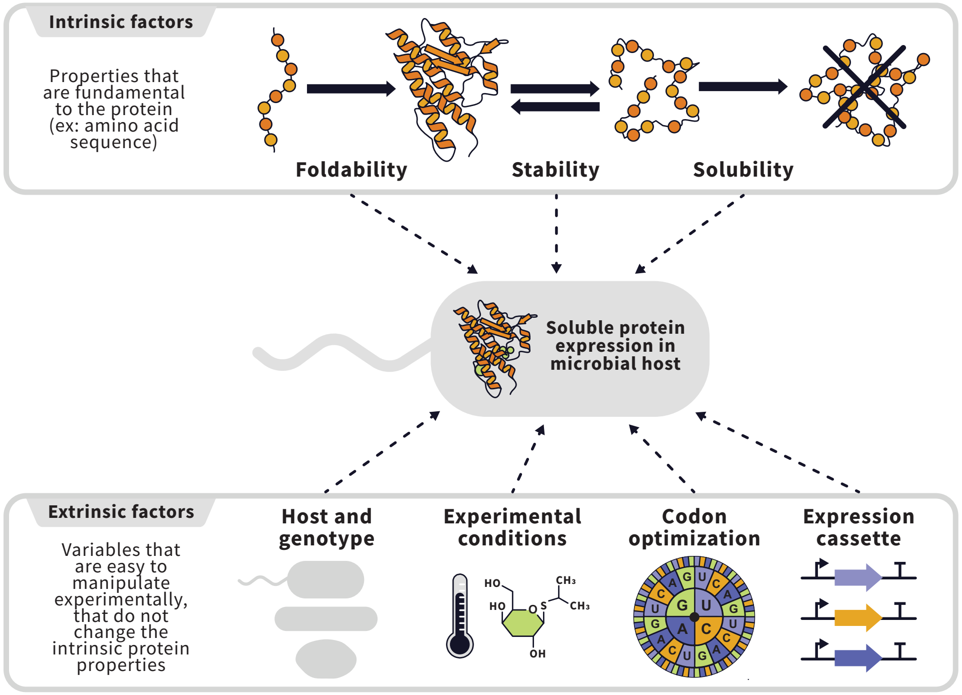
[10]Department of Pharmaceutical Chemistry, University of California, San Francisco, CA, USA
[11]Decade, Inc., Mountain View, CA, USA
[12]Department of Bioengineering, Stanford University, Stanford, CA, USA
[13]These authors contributed equally

*Correspondence:
erika@alignbio.org (E.A. DeBenedictis).


DNA
RNA
Folded protein
Stable protein
Soluble protein expression in microbial host
foldability
stability
solubility
Polypeptide chain
Misfolded protein
Insoluble aggregate
Foldability
"Ability of a protein sequence to fold into its functional structure"
Stability
"Tendency to maintain a native conformation"
folded vs unfolded
Solubility
"Concentration in a saturated solution in equilibrium with a solid phase under specific conditions"
in solution vs aggregated

Trends in Biotechnology

Figure 1. Steps in the recombinant protein expression process. Successful soluble protein expression is the outcome of many processes, including transcription, translation, protein folding, and protein degradation. The sequence of the protein determines biophysical properties that interact with the environment to impact the probability of foldability, stability, and solubility in an expression host. Figure adapted from Bhatwa and colleagues [74], licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).


Intrinsic factors
Properties that are fundamental to the protein (ex: amino acid sequence)
Foldability
Stability
Solubility
Soluble protein expression in microbial host
Extrinsic factors
Variables that are easy to manipulate experimentally, that do not change the intrinsic protein properties
Host and genotype
Experimental conditions
Codon optimization
Expression cassette

Trends in Biotechnology

Figure 2. Intrinsic and extrinsic factors contributing to soluble protein expression. Intrinsic factors, those 'hard-coded' into the protein via its sequence, and extrinsic factors, those that can be changed experimentally, are both important to explore in a protein expression dataset.

Glossary

Benchmarking: in machine learning and artificial intelligence, benchmarking is the process of systematically comparing performance of different models against predefined metrics and standards to identify strengths and weaknesses for each of the models.

Cell-free expression system: a soup of purified or crude components that can carry out protein expression in the absence of a cell.

Codon optimization: synonymous changes to the codons of a sequence to improve expression.

Directed evolution: changing a specific trait through rounds of screening/selection after changing the DNA of an organism.

Expression cassette architecture: the combination and organization of genetic elements controlling or augmenting the expression of the target protein. These components can be expressed from plasmids or integrated into the host genome. Some examples of expression cassette components include gene promoter, origin of replication, plasmid background, integration site, solubility enhancing fusion proteins, and tags required for purification.

Expression host: a microbe or cell line used to express heterologous protein. Examples include bacteria (*Escherichia coli*, *Bacillus subtilis*), fungi (*Pichia pastoris*), mammalian cell lines (CHO, HEK293), and insect cell lines (*Spodoptera frugiperda*, *Trichoplusia ni)*.

Foldability: ability of a protein sequence to fold into its functional structure.

Inclusion bodies: aggregate of proteins, usually misfolded but sometimes still functional.

Machine learning (ML): training algorithms on existing data so they can identify patterns and predict the results of new, untested data or experiments.

Microbial dark matter: unexplored microbial diversity.

Pooled expression assay: a multiplexed measurement of protein quantity in a mixed population of cells, each expressing a distinct ORF.

Reproducible: samples run three times and show the same measurement within an acceptable error range. The method can be run successfully at another laboratory, and data would reproduce. This is often helped by using automation and by providing detailed and freely

As discussed above, soluble protein expression is the compound outcome of numerous upstream processes and is fundamentally constrained by the biophysical characteristics of the protein. For this reason, soluble protein expression is a valuable single readout that summarizes many relevant inputs. It is important to note that not all soluble proteins are properly folded or functional. Similarly, aggregates can appear in the soluble fraction [5]. Although proteins in aggregates or **inclusion bodies** may be functional, we focus our attention on soluble proteins in this review.

While soluble expression is the desired goal, understanding where in the process expression fails is also valuable; however, probing these details may be secondary to gathering soluble expression data. Failure to express can be tackled with a bottom-up or top-down approach by attempting to diagnose where expression has failed or by finding conditions that work and then deducing why expression failed initially. Did expression fail at the transcriptional level? An example of a bottom-up approach would be to monitor RNA levels of the gene of interest over the course of the expression experiment. Low levels of RNA after induction may be caused by a weak promoter or poor mRNA stability. To investigate this same issue via a top-down strategy, one could change the strength of the promoter driving the gene of interest or choose a strain with less RNase activity [e.g., BL21 Star (DE3); https://www.thermofisher.com/order/catalog/product/C601003] and then assess expression. Did expression fail at the protein folding step? Expression can be tested in genetically modified strains with different protein folding chaperones or with increased ability to form disulfide bonds (e.g., SHuffle T7 Express Competent *E. coli*; https://www.neb.com/en-us/products/c3029-shuffle-t7-express-competent-e-coli). If the chaperone-enhanced strain improved expression, one could hypothesize that the missing chaperone was what caused poor expression in the first experiment.

Did expression fail because of protein stability/solubility issues? These properties are challenging to assess in lysate; however, using a purified sample, one can assay homogeneity (aggregation) using dynamic light scattering [6]. Alternatively, differential scanning fluorimetry [7] could be used to identify samples that do not contain a folded protein. Expression at a lower temperature after induction or expression of a protein with the addition of a solubility tag can improve outcomes.

Both bottom-up and top-down approaches add value to the larger soluble expression dataset and to optimizing expression of a protein of interest. A top-down strategy could provide a solution to poor expression and would build the dataset by adding alternate conditions that would be stored as metadata, while a bottom-up approach adds diverse data types and targeted avenues for troubleshooting expression failure per open reading frame (ORF). Critically, soluble protein expression as a final readout captures the success (or failure) of all steps in the process and would be the recommended first data type to collect.

Where are we now, and where do we need to go?

There is currently no predictive model to understand which combination of factors will be most successful for soluble protein expression. Although there are available protein expression datasets, many are collected using different experimental methods, test expression in a single host, were not collected with ML applications in mind, or focus on a small subset of proteins or protein domains. Similarly, most existing models focus on a single variable of expression in a single host. The ideal dataset that would serve as a foundation for a predictive model of protein expression would be large in scale, cover diverse organisms, use consistent experimental protocols, be freely available, and be designed with ML utility in mind.

The creation of this type of protein expression dataset should be coupled with its use for **benchmarking** existing and future models, both for predicting protein expression, given an ORF

available protocols (shareable) and well-defined and described internal standards.

Scalable: the assay can be run at high throughput, and it is economically feasible to do so.

Shareable: detailed assay methods (technical and analytical) are made freely available.

Singleplex expression assay (SPX): an arrayed measurement of protein expression measured in cells or cell lysate. Results from singleplex expression measurements can be used to calibrate large-scale, pooled expression measurements or on their own as standalone quantification. Examples include HiBiT, split fluorescent protein (split-FP), and Pierce BCA.

Solubility: the protein's concentration in a saturated solution in equilibrium with a solid phase under specific conditions (e.g., proteins in solution vs. aggregated).

Soluble protein expression: measurable recombinant protein in the soluble fraction of cell lysate produced in the context of an expression system.

Stability: tendency to maintain a native conformation (e.g., folded vs. unfolded proteins).

sequence and host organism, and for generating ORF sequences, given a particular host organism and desired expression level. This can be facilitated by holding regular benchmarking competitions using the collected protein expression datasets, in which the community is challenged with defined predictive and generative tasks centered on protein expression and resulting model performance is compared against existing standards [8–10]. The use of datasets with benchmarking in mind not only informs the field of the strengths and limitations in current modeling approaches [8] but also reveals gaps in the dataset that can inform future data collection to bridge these identified gaps.

Benchmarking competitions such as Critical Assessment of Structure Prediction (CASP) continue to play a key role in evaluating the strengths and weaknesses of structure prediction models and have highlighted the capability of models such as AlphaFold and AlphaFold2 in accurately predicting protein structures from amino acid sequences [11,12]. CASP accomplished this by challenging the computational structure modeling practitioners each year to predict the structures of proteins that experimentalists had determined recently but had not yet deposited in the public domain, as well as by performing a side-by-side comparison of the predictive models using well-defined evaluation metrics. Moreover, at the end of each year's competition, CASP organizers also provided an appraisal of the state of structure prediction modeling in a retrospective report to the community. In recent years, similar benchmarking competitions have been designed for both predictive and generative models for protein engineering and protein–small molecule binding [13,14]. Benchmarking competitions could play a similar role in assessing the strength of predictive and generative models for soluble protein expression. The benchmarking competitions can be designed around defined predictive and generative tasks revolving around real-world protein expression problems, such as maximizing soluble expression in a particular heterologous host, while maintaining specific activity. The performance of the models can be compared using evaluation metrics that would meaningfully assess success in achieving these tasks and reporting a critical assessment of each year's competition around the state of the predictive and generative modeling for soluble protein expression.

Current approach to soluble recombinant protein expression

Today, optimizing protein expression is primarily an exercise in experimental trial and error, guided by many qualitative predictors. The first step in a protein expression experiment is to choose an expression host. An excellent in-depth review of choosing an expression system has recently been published [15]. Important factors that drive the decision for selecting a host include whether the target protein is a prokaryotic or eukaryotic protein, whether the downstream use of the protein requires post-translational modification, if the protein has disulfide bonds, the size and complexity of the protein (e.g., membrane protein), and downstream applications [15]. One must then decide on a number of experimental variables (extrinsic factors), such as where to express the protein (extracellular or intracellular), the expression cassette details, which specific expression strain to use (genetic modifications to improve expression of proteins with specific characteristics), and what conditions to express the protein in (media, temperature, time, type of induction, etc.).

Today, qualitative predictors of protein expression exist and are used to improve expression, often through experimental trial and error. A quantitative ML model of protein expression based on a large, open dataset, coupled to existing guidance and experience, would enhance the predictability of soluble expression and would be a widely used resource in both academia and industry.

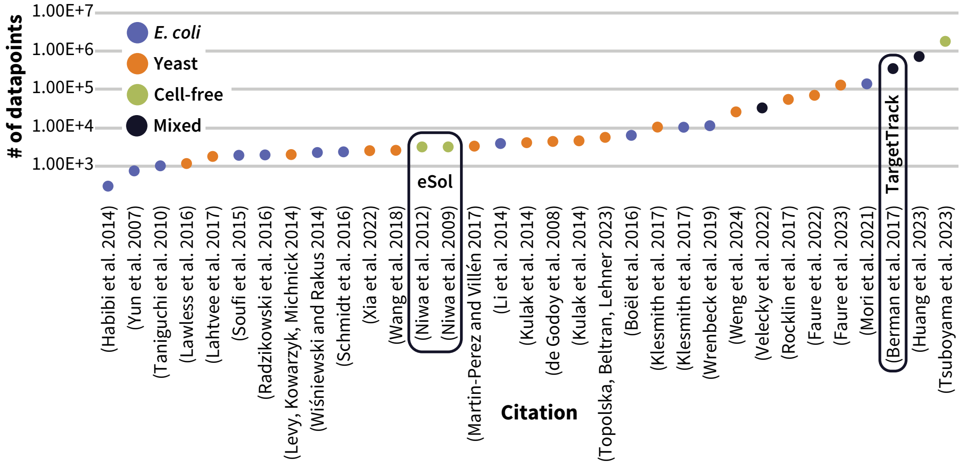
Available protein expression datasets

ML models for generalized prediction of soluble protein expression will require new, large-scale datasets that do not exist today. Existing soluble protein expression datasets are small or highly focused, or they have been collected from proteins expressed *in vitro* rather than in living organisms. Larger datasets are typically narrow in scope (single proteins or domains) or aggregations of existing smaller datasets. They use data collected under differing conditions or protocols and with inconsistencies in data analysis (Figure 3). This creates gaps in usable data and metadata to help build ML models [16]. Finally, existing studies have typically focused on one expression host. We suggest that a valuable dataset would use a unified experimental setup rather than database mining and would focus on expression of ORFs in diverse expression hosts (e.g., *Escherichia coli*, *Pichia pastoris*, *Bacillus subtilis*, CHO, HEK293T, *Aspergillus niger*, *Saccharomyces cerevisiae*). Building a standardized dataset of diverse proteins captured in a consistent experimental setting across various hosts would be of incredible value on its own, but its value would increase exponentially if used to develop highly predictive models of protein expression.

Available protein expression models

> ‘The biological principles for recombinant protein expression are well established; however, the ability to distinguish protein targets that express well from those that express poorly is still considered a ‘black box’ process that often requires screening many conditions to obtain a soluble product.’ [17]

Models today focus on making predictions of one component of the multistep protein expression process, such as codon use or protein biochemical solubility, rather than making predictions of the overall process. Some models rely on gathering data on the same protein encoded with different codons to predict improved expression. One example of ubiquitously used models is those for **codon optimization**. When expressing a heterologous gene, it is often without a second thought that a scientist will use codon optimization to improve expression – these models have become synonymous with ordering synthetic genes for recombinant expression. Although the concept of codon optimization was first proposed nearly 40 years ago, there is little consensus between models when optimizing the same ORFs for expression in the same host [17]. Furthermore, despite substantial efforts to develop effective codon use optimization algorithms,



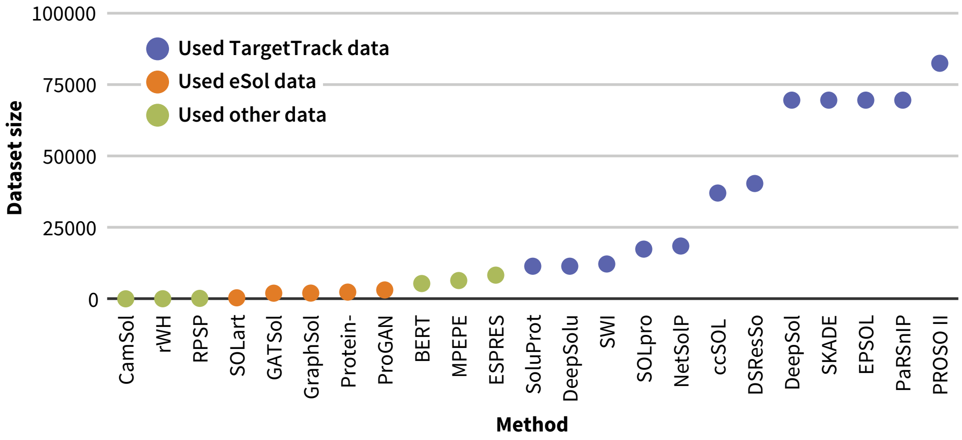
Trends in Biotechnology

Figure 3. Summary of existing protein expression datasets. The expression datasets are ordered from smallest to largest by number of datapoints (*y*-axis). The color of the points represents the expression system that was used to gather the data. The *x*-axis is first author and year for the dataset paper citation. Table S1 in the supplemental information online includes full citations, an assay summary, and annotation for whether the data represent native protein expression or ectopic recombinant protein expression.

optimized sequences continue to show differences in expression level [17]. This example highlights the intricate nature of protein expression, even when focusing on a single variable such as codon use.

Most existing models related to expression focus on protein solubility. Two datasets are highlighted in Figure 3 (eSol and TargetTrack) because they were used by multiple protein solubility models as input data. The eSol data are composed of ~3200 *E. coli* proteins translated in a **cell-free expression system** with and without chaperones [18,19]. TargetTrack was a large protein expression effort driven by the Protein Structure Initiative (PSI) and was a compilation of several of their databases (PSI, TargetDB, PepcDB). This effort ended in 2017, and the data are available online to parse [20]. TargetTrack was an enormous effort spanning 35 centers and over 100 investigators, and expression, purification, and downstream structure determination of over 300 000 proteins was attempted. This was an impressive effort whose data have been used repeatedly for model building (Figure 4), although some characteristics of the dataset could be improved. For example, Hon and colleagues note, ‘A major limitation of this database is the low quality of its annotations…. 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’ [1]. We suggest that a future dataset ‘stand on the shoulders of giants’ in order to expand open data available for model building and to leverage learnings from previous dataset structure and collection.

Protein variant effect prediction (VEP) models assess how mutations in a protein may affect protein function [21]. In a similar way that a model of protein expression would allow researchers to move faster with fewer experiments, these models are of high interest because they can predict the consequences of mutations to a sequence without having to test the new proteins. Many VEP models have been benchmarked on predicting protein expression data in hundreds of publicly available deep mutational scanning and human clinical variant datasets in ProteinGym (>2.5 million data points, >200 different datasets). ProteinGym evaluates both supervised and unsupervised VEP models tested on different proteins expressed in a range of organisms. One



Trends in Biotechnology

Figure 4. Summary of existing solubility prediction models. The protein solubility prediction models are ordered on the basis of the size of the dataset used to generate the model. This is the total size and does not separate the data used for training the model. The colors of the dots correspond to the expression datasets that were used in model building. Table S2 in the supplemental information online details citations, the name of the method, the predicted property, the dataset used for model generation, the expression system that the dataset used came from, and the dataset size.

of the most commonly used metrics for success in VEP models is the Spearman correlation coefficient between the model predictions and the experimental data, where 0.6 is interpreted as a good signal and 0.2 is a weak signal. This curated set of benchmarks reveals that the best supervised model for predicting expression data, ProteinNPT [22], achieves a Spearman correlation coefficient of 0.637. The best unsupervised model for predicting protein expression is ProSST [23], with a Spearman correlation coefficient of 0.53. By contrast, the lowest-performing supervised and unsupervised models have Spearman coefficients of 0.226 and 0.171, respectively. Notably, these models are built from scratch each time on each new dataset encompassed within ProteinGym; they cannot be represented appropriately in Figure 4. These benchmarks demonstrate the potential success of protein variant prediction models in predicting protein expression; however, there remains significant room for improvement to achieve stronger correlations with experimental results.

The ideal soluble protein expression dataset

Building a generalized model of soluble protein expression across organisms will require a next-generation dataset that is larger, more comprehensive, and more extensible than the data available today. In order for the required dataset to provide value, it would need to be large, diverse, unified, and openly available. Specifically, the dataset should contain expression data across diverse ORF families and include expression of the same proteins in several organisms to be compared side-by-side. The dataset should include expression experiments *in vivo* rather than *in vitro* to capture intrinsic and extrinsic factors for expression and should be set up to generate ML-ready data and metadata. For data to be considered ML-ready, it should be freely available and provided in a standardized and consistent experimental format(s), and data collection should be **reproducible**, **shareable**, and **scalable** and should overall comply with FAIR (findability, accessibility, interoperability, and reuse of digital assets) principles (https://www.go-fair.org/fair-principles/) to enable use by scientists across fields and sectors to expand the dataset.

Roadmap toward an expression dataset to inform predictive design

If the scientific community is going to invest in gathering data on soluble protein expression, how should it be done to maximize the data quality and impact? Here, we roadmap technical options for data collection strategies. We review options for the two most critical design choices: which organisms to use as expression hosts and which assays to use to collect data. We evaluate organisms for their technical risk and possible impact and assays for their scalability and potential for creating a robust, open dataset.

Expression hosts

There are numerous protein expression hosts that are used because of their diverse strengths and weaknesses. (For more details, please reference this recent review on the topic [15].) Table 1 summarizes some of the most commonly used microbial expression systems.

It is challenging to build a framework for prioritizing the order of data acquisition when data in every organism would be valuable. To begin, we suggest acquiring data in organisms with low technical risk and high impact even in isolation. Data acquisition for these organisms can create momentum for a much larger dataset collection effort. We suggest *E. coli* and *P. pastoris* as starting organisms for an expression dataset because these are commonly used, easy-to-scale microbes that would produce data useful to a wide audience, including both academia and industry. They provide a quick avenue to test the data collection platform at an economical price point.

Beyond the first two organisms, we suggest prioritizing the order of host expansion on the basis of several criteria, including promoting organisms that are widely used and deprioritizing

Table 1. Comparison of commonly used protein expression microbes[a]

Organism	Summary	Growth	Genetics	Post-translational modification	Expression efficiency	Refs
Escherichia coli	Most commonly used, used in both academia and industry, easy to use, many genetic tools, cheap, many strain options, not a good choice for complex proteins or those that require PTM or many disulfide bonds, easy to do HTP screening, only some strains are considered GRAS	Fast and high efficiency, simple media requirement	Well-defined, simple, and high efficiency, automation-friendly	Limited	High without efficient secretion	[15,73]
Pichia pastoris	More recent addition to the protein expression lineup, easy to use, although fewer tools than common bacterial hosts, used for complex proteins, cytokines, nanobodies, have eukaryotic PTMs, amenable to HTP screening, GRAS	High cell density, easy scale-up	Well-established, automation-friendly manipulation	Yes but hypermannosylation can be an issue	Moderate to high of secreted proteins	[15,78,79]
Bacillus subtilis	Commonly used, used in both academia and industry, easy to use, many genetic tools, cheap, many strain options, good option for secreted proteins, easy to do HTP screening, GRAS	Fast and high efficiency	Convenient for gene modification, automation-friendly	Limited	High yield with secretory expression and produces no lipopolysaccharide	[80]
Aspergillus niger	Used primarily in industry, can degrade biomass (CAZzyme expression for industry), strong survivability (good for producing antibiotics, enzymes), good for secreting organic acids, proteins, enzymes, and secondary metabolites. HTP screening is challenging. Strain generation and screening are not quick, GRAS	Fast and high efficiency, measurement and control of filamentous growth and macromorphology not trivial	Complex manipulation and lower transformation efficiency, automation is challenging to use	Typical eukaryotic post-translational modifications	High and efficient secretion	[81,82]

[a] Table adapted from Zhang and colleagues [76] and Ntana and colleagues [77], licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). Abbreviations: GRAS, generally regarded as safe; HTP, high throughput; PTM, post-translational modifications.

organisms that introduce new technical risks. As the dataset grows, key milestones will be expansion to other microbes; eukaryotes, including mammalian cell lines and filamentous fungi; and a broader exploration of expression hosts, including insect cells and cell-free methods. While expression of the same ORFs between hosts builds data toward predicting expression in different host systems, the addition of cell-free methods could bolster predictive power between these *in vitro* methods that are capable of rapidly screening many combinations of enzymes and the *in vivo* systems ultimately used for large-scale fermentation [24]. For each expression system, data collection requires identifying a specific strain genotype and expression cassette for initial data acquisition.

Data collection methods

In this section, we review experimental methods that were used in gathering existing datasets, describe alternative methods, and provide an assessment of these for the purpose of collecting a large-scale expression dataset (Table 2, Table 3).

The assays discussed are categorized as either **singleplex (SPX)** or **pooled expression assay** methods. SPX methods measure one sample per assay (arrayed), even when multiple samples

Table 2. Assay assessment for singleplex measurements[a]

Assay	Reproducible	Shareable	Scalable	Quick/easy to run and analyze	Affordable	Adaptability to other expression systems	+/–
Bradford (with Ni-NTA enrichment) Total = 14	✓✓ Common method, variability between proteins	✓✓✓	✓✓ Plate-based	✓✓✓ Common method	✓ Kit >$5/sample	✓✓✓ His tag + protein extraction	+ Adaptable, ‘gold standard’ method – Variability between proteins, standards must be run each time
BCA (bicinchoninic acid) (with Ni-NTA enrichment) Total = 14	✓✓✓ Common method	✓✓✓	✓✓ Plate-based	✓✓ Common method, multistep	✓ Kit >$5/sample	✓✓✓ His tag + protein extraction	+ Adaptable, ‘gold standard’ method – Interference, standards must be run each time
AlphaLISA Total = 14	✓✓✓ Kit-based method	✓✓✓	✓✓ Plate-based	✓✓ Derivative of a common multistep method	✓ Kit >$5/sample	✓✓✓ His tag + protein extraction	+ Adaptable, ‘gold standard’ adjacent method, can be automated – Requires two shared epitopes for all proteins
Coomassie SDS-PAGE (with Ni-NTA enrichment) Total = 10	✓✓ Lysis variability	✓✓✓	✗ Some steps not plate-based	✓ Many steps involved, quantitation can be variable	✓ Ni-NTA for enrichment, cost of people time	✓✓✓ His tag + protein extraction	+ Adaptable, common ‘gold standard’ method – Low throughput, quantitation variability
MS (with Ni-NTA enrichment) Total = 14	✓✓✓ Common method	✓✓✓	✓✓ Plate-based, instrument time	✓✓✓ Common method	✗ Ni-NTA for enrichment, instrument time	✓✓✓ His tag + protein extraction	+ Adaptable, common ‘gold standard’ method – Costly (instrument time)
HiBiT Total = 14	✓✓ Kit-based, lysis variability	✓✓✓	✓✓ Plate-based	✓✓ Growth, lysis, readout	✓✓ Kit <$1/sample	✓✓ HiBiT tag, lysis troubleshooting	+ Easy, commonly used method industrially, kit-based – Cell lysis and reaction timing may require troubleshooting, not a gold standard
Split-FP Total = 16	✓✓✓ Plate reader protocol	✓✓✓	✓✓ Plate-based	✓✓✓ Growth, readout	✓✓✓ Plate reader time	✗ FP expression, requires tuning	+ Easy, commonly used – FP expression tuning, not a gold standard

[a] Check marks denote positive attributes, ‘x’ marks denote suboptimal characteristics. The ‘?’ highlights areas where the characteristic is unknown. ‘Total’ refers to the number of check marks per row. See Table S3 in the supplemental information online for assay details and citations. Abbreviation: FP, fluorescent protein.

are processed in parallel or using automated systems (e.g., fluorescence readout from plates). By contrast, pooled methods combine many samples into a single measurement and require downstream deconvolution of data. Generally, pooled methods are more cost-effective and have higher throughput, while SPX methods remain simpler to implement and analyze.

Quantification of specific proteins can be done using a variety of assays, each with unique characteristics. Assays can measure protein expression in living cells, or they can quantify protein from cell lysate, either in the whole lysate or after enrichment of the target protein. Some commonly used assays (UV-Vis absorbance, Bradford, bicinchoninic acid) are typically used to quantify total protein in an SPX format. To be used for an expression dataset, ideally, the protein of interest would be isolated and enriched before quantification. Assays used to detect a specific protein in a complex sample include ELISA, western blots, size exclusion HPLC and mass spectrometry (MS). We focus on assays that have previously been used for high-throughput expression data collection. Reformatting the graph in Figure 3 to color and group points by assay used, one can see that MS, SDS-PAGE, and growth-based assays are the most common methods used in multiple large datasets (Figure 5).

We briefly describe the assays found in Figure 5, covered in order of most to least prevalent (from left to right of the graph in Figure 5). The assays are assessed in Table 2 and Table 3.

MS

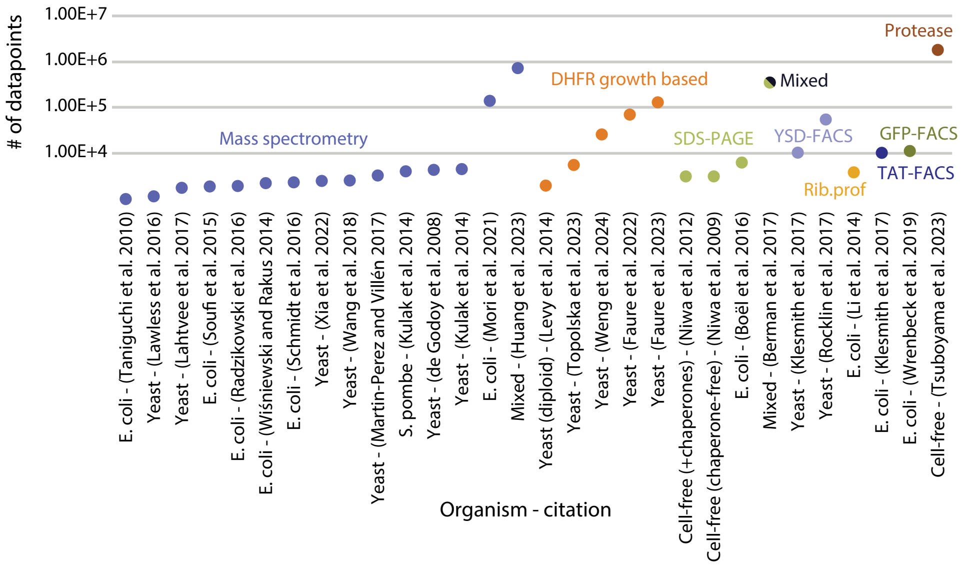
MS is a way to measure a molecule's mass-to-charge ratio. This information can be used to identify and quantify the molecule of interest (https://www.broadinstitute.org/technology-areas/what-mass-spectrometry). Most studies in Figure 5 that used MS were quantifying proteomes rather than targeted ORFs, as would be the suggested implementation for a protein expression dataset. Stable isotope labeling or label-free proteomics with ORF targeted data analysis could be implemented in a singleplex or pooled manner (https://www.protocols.io/view/label-free-quantification-lfq-proteomic-data-analy-5qpvobk7xl4o/v2).

Peptide barcodes ('flycodes')

There is a recently developed peptide barcoding strategy in which proteins are directly barcoded with short peptide barcodes (also called 'flycodes') [25]. These flycodes are cleaved from the final protein and then distinguished/quantified with MS. The number of available unique peptide flycodes that can be pooled and distinguised using MS is not limiting [25,26].

Growth-based assays

Growth-based assays use growth as a readout for the fitness of a target protein and can be performed on SPX or pooled samples. One popular version of a growth-based assay requires reconstitution of a protein that confers antibiotic resistance, such as dihydrofolate reductase (DHFR), which provides resistance to methotrexate [27]. Growth-based assays have commonly been



Trends in Biotechnology

Figure 5. Summary of assays used to generate existing expression datasets. Reformatted Figure 3 (summary of existing protein expression datasets). Here, data points are colored and grouped by the assay that was used. The assay text is the same color as the point showing the dataset that was collected using that assay. The *x*-axis shows the expression host followed by the first author and year of the dataset reference. Refer to Table S1 in the supplemental information online for full references.

Table 3. Assay assessment for pooled measurements[a]

Assay	Reproducible	Shareable	Scalable	Quick/easy to run and analyze	Affordable	Adaptability to other expression systems	+/–
Label-free proteomics Total = 12	✓✓ Newly adopted method	✓✓✓	✓✓ Pool size limitation, MS time	? Pool size limitation, requires protocol and analysis development	✓ Ni-NTA for pull-downs, MS time, pooled, depends on pool sizes	✓✓✓ His tag + protein extraction, need identity of native proteome	+ Adaptable, pooled, direct measurement – Uncommon use case requires development, pool size limited by ORF peptide diversity
Sort-seq Total = 13	✓✓ FACS method/analysis	✓✓✓	✓✓✓ FACS time	✓✓ Growth, readout, FACS + sequencing	✓✓ Pooled, FACS time	✓ FP expression, tuning, FACS methods at facilities may be limiting	+ Pooled, mostly adaptable – FP expression tuning; FACS method and analysis onboarding / development
Growth-based assay (DHFR) Total = 13	✓✓ Antibiotic selection	✓✓✓	✓✓✓ Automatable NGS methods	✓✓ Growth, antibiotic selection, sequencing	✓✓✓ Pooled, NGS	? Requires specific antibiotic sensitivity	+ Affordable, pooled – Indirect readout, adaptability unknown, uncommon assay for expression, large tag on ORF required
Peptide flycodes	? New method and analysis pipeline	✗ Method sharing prohibited	? Large pools possible, MS method time	✗ Pooled growth, new method and analysis pipeline	✓ Pooled, MS time, method dev'l cost	✓✓✓ Peptide tag on ORF and protein extractions	+ Adaptable, pooled – New method with unclear chance of success, nontrivial data analysis and design, not a shareable method
Cell-free protease assay	? An existing method moved to a cell-free system	✓✓✓	✓✓✓	✓✓✓	✓✓✓	✗ Cell free	+ Affordable and quick – Not commonly used, cell-free
YSD+FACS	✓✓ FACS method/ analysis	✓✓✓	✓✓ FACS time	✓✓ Growth, staining, readout, FACS + sequencing	✓✓ Pooled, FACS time, stain	✗ Yeast only	+ Easy, commonly used – Yeast-specific, requires protein export to cell surface
TAT+FACS	✓✓ Antibiotic selection	✓✓✓	✓✓✓ Automatable NGS methods	✓✓? Growth, antibiotic selection, sequencing	✓✓✓ Pooled, NGS	✗ *E. coli* only	+ Easy – *E. coli*-specific, requires protein export

[a] Check marks denote positive attributes, 'x' marks denote suboptimal characteristics. The '?' highlights areas where the characteristic is unknown. 'Total' refers to the number of check marks per row. See Table S4 in the supplemental information online for assay details and citations. Abbreviations: FP, fluorescent protein; NGS, next-generation sequencing.

used to interrogate protein–protein interactions where two proteins are each tagged with half of the antibiotic resistance marker, and, upon interacting, the reconstituted marker confers a fitness advantage in the face of antibiotic selection [28–30]. Split-DHFR has been used for protein quantification in yeast to generate large datasets (Figure 5, orange points). To measure the abundance (expression) of a target protein, the target is tagged with a portion of DHFR, and the complementing part is overexpressed in the cell [28,30]. The higher the expression of the tagged ORF, the more DHFR is reconstituted (since the other half of DHFR is expressed in excess over the tagged protein and is not limiting) and therefore the more this specific strain can grow in the presence of methotrexate. When used to interrogate a pool of strains, the population is sequenced before and after selection, and the abundance from deep sequencing can be used to calculate fitness (growth rate) and infer expression of each variant, with thousands of variants per pool [30].

SDS-PAGE

This technique is one of the traditionally used SPX methods for confirming protein expression for purified proteins. However, it is difficult to scale this to assay many samples. As was mentioned previously, this method was used by TargetTrack. One of the participating locations, the Northeast Structural Genomics Consortium (NESG) was responsible for expressing over 25 000 constructs in *E. coli* [BL21(DE3)] from a pET-based system where researchers used an SPX assay for expression. They quantified via Bradford assay and subsequently included additional SDS-PAGE to verify their Bradford results [31,32]. In more recent protocols, the protein purification step was removed, and soluble and total cell lysates were run on SDS-PAGE followed by staining the gels with Coomassie Blue and quantifying expression in the soluble and insoluble fractions on a scale from 0 to 5 on the basis of the intensity of the band observed [31,33,34]. Capillary electrophoresis is a modern alternative to SDS-PAGE but carries high consumable costs.

Yeast surface display + fluorescence-activated cell sorting

Yeast surface display (YSD) is a technique where target proteins are tagged for extracellular localization (N-terminal Aga2p domain) and with an epitope tag on the C-terminus. Successfully expressed cell surface proteins can be quantified using epitope specific antibodies. To use this in a pooled manner, fluorescently conjugated antibodies can be coupled to fluorescence-activated cell sorting (FACS). Bins of sorted populations can be sequenced to identify constituent ORFs [35].

TAT [twin-arginine translocation (Tat)-selective export of folded proteins into the bacterial periplasm]

Similarly to the YSD assay, proteins are tagged with a periplasmic export signal (Tat) and a beta-lactamase enzyme that confers resistance to ampicillin when extracellular. ORFs that are successfully expressed and exported confer resistance to ampicillin [35]. Similar to a growth-based assay, this method can be used on a pool of strains.

GFP + FACS

Wrenbeck and colleagues [36] fused GFP to thousands of protein variants (two proteins with thousands of mutations) to probe expression (dark green point in Figure 5). The cells were then sorted into bins by FACS, and sequencing on binned populations was used to map mutations to different amounts of fluorescence. This approach allows pools of cells to be interrogated at the same time. A variation on this method is based on a method called ‘bimolecular fluorescence complementation (BiFC) [37]. The ORF of interest is tagged with a portion of a fluorescent protein (e.g., a portion of GFP, GFP_{11}), and the complementing fragment ($\text{GFP}_{1\text{-}10}$) would be constitutively overexpressed in the same cell. Interaction of the two fragments produces a fluorescence signal, dependent on the expression and proper folding of the ORF of interest [37]. In an SPX setting, a plate reader can be used to gather fluorescence data per strain. Extending the method to a pooled approach requires FACS and sequencing, as was used by Wrenbeck and colleagues [36]. This method is also referred to as ‘Sort-Seq’ or ‘FlowSeq.’

Cell-free protease assay

In this method, test proteins are transcribed and translated with cell-free cDNA display in a pool. The proteins are N-terminally PA tagged, and after translation they contain their cDNA at the C-terminus. These tagged proteins are then challenged with proteases, pulled down and quantified via sequencing of the C-terminal cDNA tag [38]. Protease resistance is used as a metric for protein stability.

HiBiT

Protein abundance can be measured using a split NanoLuc luciferase system. Proteins of interest are N- or C-terminally tagged with the 11–amino acid HiBiT peptide tag (a portion of NanoLuc).

The HiBiT complementing LgBit polypeptide is added after cell lysis. When HiBiT and LgBiT interact, they reconstitute the luminescent enzyme, NanoLuc. Luminescence intensity is proportional to the abundance of the HiBiT tagged protein of interest (https://www.promega.com/resources/technologies/hibit-protein-tagging-system/).

Comparison of assays

An ideal assay for large-scale protein expression dataset collection should be reproducible, shareable, scalable, quick and easy, affordable, and extensible to different host organisms. Reviewing the assays used to collect previous datasets, we conclude that there is no clear 'winner' for large dataset collection. Rather, each assay has unique pros and cons (Table 2, Table 3). Some of the assays that were used to populate existing datasets (Figure 5) could be used to generate a large protein expression dataset (MS, HiBiT, Split-FP, growth-based assay), while others are difficult to scale with automation (SDS-PAGE), are not adaptable (cell-free protease assay, YSD, TAT), or use methods that cannot be made publicly available (peptide flycodes). We summarize our assessment of the assays in Table 2 and Table 3.

A pooled assay would allow large-scale dataset generation in the most efficient way possible. Pooled methods such as label-free proteomics and Sort-seq are attractive options because of their economy, adaptability, and shareability. However, because pooled assays are nontraditional for protein quantification, it is critical that they undergo validation using low-throughput (singleplex) methods of protein quantification. The SPX assays in Table 2 contain strong candidate assays for this orthogonal data collection toward pooled assay validation. It is important to note that these methods could also be scoped as large-scale dataset collection methods, pending their scalability.

Whichever assay is chosen for large-scale dataset collection, data collected using an alternative method for a small percentage of samples is recommended. These orthogonal data serve both as a continuous spot-check for the high-throughput assay during data generation and as additional training or testing data for models capable of handling labels generated in different ways.

ML analysis strategy

The purpose of this review is not to prescribe ML models for predicting protein expression but rather to showcase that a dataset enables such an approach. Data must be collected in a manner such that the ML community can easily use the dataset to develop a variety of models for predicting protein expression. Involving ML experts throughout data collection can enable a 'smarter not harder' approach that prioritizes informative, diverse new sequences rather than simply expanding dataset size with redundant or low-value datapoints.

Calculating quantitative protein abundance and storing metadata

Data would be composed of a protein DNA sequence and corresponding protein abundance. Protein abundance calculations would be assay-specific. Metadata would include expression host; plasmid information, including annotated sequence in .gb format; protein amino acid sequence with tags; UniProt and NCBI accession numbers for protein sequences; GO terms; source species of the DNA sequence; detailed expression protocol; automation protocols where applicable; and specific assay protocols [6].

Data storage

Data should be made available in a relational database accessible to the public through a REST API [e.g., Experiment Data Depot (https://public-edd.agilebiofoundry.org/) and/or other available options]. Snapshots of this database at certain time points (e.g., for the initial dataset or for large data increases) will be made available to the public in a static manner through Nature Scientific Data (https://www.nature.com/sdata/), Zenodo, or similar established third-party repositories.

Proposed ML models

A supervised ML approach could be used to predict protein concentration (response or label) from protein DNA sequence and metadata (input or features). Specifically, the prediction task takes two inputs – the DNA sequence and the desired host – and produces one output – predicted reconciled protein concentration from the SPX and pooled approaches (Figure 6, Key figure). Unsupervised methods have also shown a signal toward predicting protein expression [23,39]. Models trained on multiple sequence alignments of a reference protein or likelihoods extracted from a large pretrained large language model can both be used as proxies to predict protein expression. Semisupervised approaches could be designed that integrate both the labeled data, as described with the supervised approach, and the likelihoods calculated from the unsupervised approach. Finally, transfer learning could be applied to any of the above approaches to extrapolate information from a data-rich area to a data-poor area. For instance, protein expression data collected under specific growth conditions (e.g., culture media, temperature, etc.) could inform prediction on protein expression in untested growth conditions. Similarly, plentiful unsupervised data – many natural sequences or structures – available for a homologue of the protein of interest could be used for modeling the protein itself.

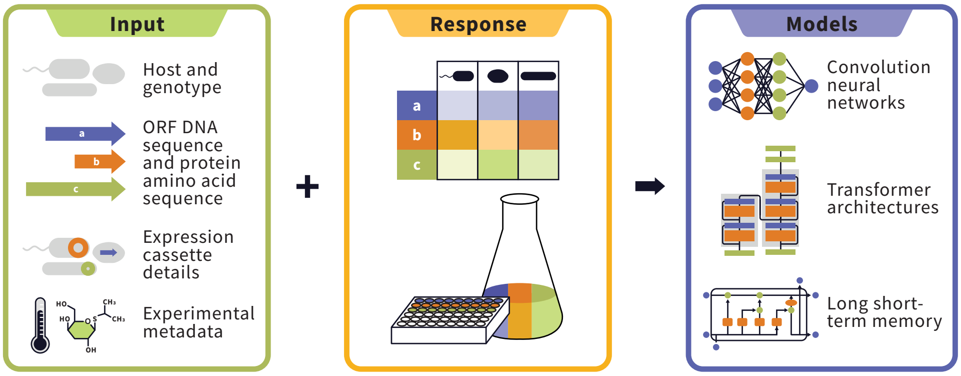
Proposed baseline model architecture

ML approaches for predicting the properties of proteins have seen significant advancements in recent years. Here, we identify promising modeling approaches by comparing the prediction of expression of soluble proteins with other biological prediction tasks.

One of the closest analogies is protein structure prediction from sequence, which was revolutionized by AlphaFold2 in 2021 [40]. Similar performance was later achieved using protein language

Key figure

Project inputs and data collected (response) provide foundation for model building


Input
Host and genotype
a
b
c
ORF DNA sequence and protein amino acid sequence
Expression cassette details
HO
HO
CH₃
CH₃
OH
Experimental metadata
+
Response
a
b
c
Models
Convolution neural networks
Transformer architectures
Long short-term memory

Trends in Biotechnology

Figure 6. An expression dataset for model building would use the ORF sequence and experimental metadata as input (green box), and the response would be the protein abundance measured in an arrayed or pooled format (orange box). Figure adapted from Radivojevic and colleagues [75], licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

models, such as ESM in 2023 [41]. These models are trained on large datasets of unlabeled protein sequences through unsupervised learning. Like natural language models, they tokenize proteins into short peptide snippets and challenge a model to predict the next token. This process enables the models to identify meaningful patterns in proteins that go beyond the raw sequence data, such as structural features. For instance, protein language models such as UniRep can correctly classify the domain of life where the protein originated, whether a given residue is in an alpha helix or beta sheet, and has inferred the chemical similarity of the 20 amino acids [42]. Today, scientists widely rely on structure-based and multiple sequence alignment–based models, such as AlphaFold, as well as protein language models, such as ESM-3, for accurate structure prediction.

Protein embeddings, derived from these language models, have been used in transfer learning to predict a variety of labeled data types, including fluorescence [43], stability [44], and other biochemical properties [45–47]. This technique involves extracting the learned embeddings from a large pretrained model and using them as input to train a model on a smaller experimental dataset [42]. Protein embeddings are valuable because they encapsulate relevant features in a compressed form, which can enhance predictive accuracy and reduce the amount of labeled data required for training [48].

Currently, predicting protein properties is limited by the availability of high-quality data. Existing ML methods can be applied to analyze new, large datasets. We suggest benchmarking new protein expression prediction models on the curated and diverse datasets included in ProteinGym and comparing the new models' performance against existing state-of-the-art models. Additionally, we suggest using the TAPE framework [47] as a baseline for benchmarking the performance of protein transfer learning on new expression datasets. This framework could be updated with embeddings from more recent models, including ESM [49], and use SPECTRA-based train–test splits [50]. After establishing baseline model performance, further research could explore new ML architectures that integrate additional inputs, such as codon optimization scores, Rosetta free energy predictions [51], and microbial genome embeddings [52]. These inputs could provide valuable information that is not captured by protein embeddings alone, potentially enhancing predictive accuracy.

Scaling laws

Empirical scaling laws in other ML domains have guided trade-offs between computer resources and data collection [53]. In protein ML, efforts to establish similar scaling laws are still nascent [54]. Further research is needed to determine how data quantity and diversity influence predictive performance. A large-scale protein expression dataset, as outlined in this review, would provide a crucial foundation for investigating these relationships and optimizing model design.

Concluding remarks and future perspectives

A predictive model for protein expression would be transformative for biotechnology, offering the potential to improve the efficiency of protein engineering and biomanufacturing. In this review, we have compiled the experimental and computational approaches that could make such a model possible. Although the 'ultimate' expression dataset would be a tremendously challenging undertaking to generate in one combined effort, we believe that this type of dataset can be built over time. The field needs a ML-ready dataset containing a large number of quantitative measurements of diverse proteins. Data must be collected in a manner such that the ML community can easily use the dataset to develop a variety of models for predicting protein expression. Furthermore, it must be freely available, comparable between organisms, and reproducible at different sites.

A broad and diverse set of scientists would greatly benefit from expanded protein expression data. However, each scientist has their favorite host, protein type, and assay, making it difficult to select a single starting point for data collection. We suggest choosing an initial, well-defined

Outstanding questions

What is the best assay(s) to generate a large-scale dataset?

How much expression data is necessary to build a predictive model that generalizes?

Can enough data be collected over the vast variable space to seed a predictive model?

How much of the experimental space does a dataset for model training need to cover versus what transfer learning can fill in?

What variables of protein expression experiments are most important to gather experimental data on?

How will models built on existing proteins perform using *de novo* designed proteins?

Does the same assay readout need to be used for all organisms tested with a specific set of ORFs?

How many ORFs need to be expressed in an organism to generate a predictive model of expression? How sequence diverse do the ORFs need to be?

With all the data generated toward 'solving' protein expression, what other basic science questions could be answered with non-ML methods?

Will a learned ML model outperform a mechanistic model on the basis of codon optimization, RNA secondary structure, chaperones or other specific phenomena known to impact expression?

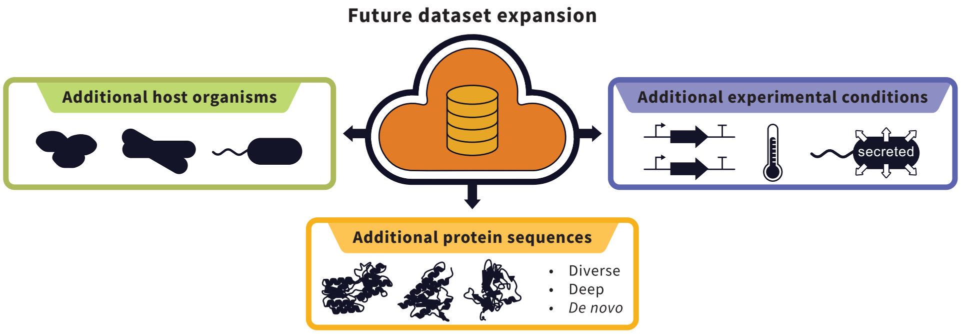
dataset that can be independently valuable to many scientists and also act as a rallying point to expand to and test a wider number of variables. We suggest *E. coli* and *P. pastoris* as starting organisms for this dataset because these are commonly used, easy-to-scale microbes that will produce data useful to a wide audience (including both academia and industry). They would provide a quick way to test the data collection platform at an economical price point and build momentum to expand to data collection in more hosts. SPX methods will likely be cost-prohibitive for large-scale data collection. Instead, these methods could be used as a validation assay to spot-check samples analyzed via pooled methods. Pooled methods are attractive options because of their economy, adaptability, and shareability. We believe that label-free proteomics and growth-based assays would be the first pooled assays to test since proteomics would be organism-agnostic and growth-based assays have produced large-scale expression data previously. Building biological datasets fit for training robust and generalizable ML models is a new endeavor, and therefore the amount and types of data required remain unknown (see Outstanding questions).

The number of variables involved in protein expression is large; however, these variables can be viewed as dimensions for future dataset growth. Examples of different variables that future expression datasets could explore include different host organisms (i.e., strains with more or less ability to post-translationally modify), new waves of protein sequences (diverse proteins, deep mutation scanning of proteins, different codon usage, *de novo* proteins), additional experimental conditions (temperature, media, growth vessel), expression cassette modifications (i.e., solubility tags), targeted localization of proteins, and coexpression of chaperones, and these are a few avenues in which the dataset can grow (Figure 7). Beyond soluble expression as a readout, future collection of different types of data that could help analyze failure modes of expression is another way that this dataset could be enriched in follow-up collection efforts and could even generate new mechanistic insights.

Acknowledgments

Thanks to the Align Team for their support and contributions to this work. Additionally, thanks to Ben Lehner, Christopher Reynolds, David Ross, Eli Bixby, Elise de Reus, Howard Salis, Swati Choudhary, and Gabe Rocklin for providing expert input on this review.

The Align Foundation: This work was supported in part by Griffin Catalyst. M.C.J. gratefully acknowledges the Army Research Office (W911NF-22-2-0210, W911NF-22-2-0246) and the Department of Energy (DE-SC0023278, DE-NA0003525). A.G. gratefully acknowledges the Slovenian Research and Innovation Agency (Research Core Funding Grant No. P1-0140). R.J.C. acknowledges funding from the Dr. Miriam and Sheldon Adelson Medical Research Foundation


Future dataset expansion
Additional host organisms
Additional experimental conditions
secreted
Additional protein sequences
Diverse
Deep
De novo

Trends in Biotechnology

Figure 7. Dataset expansion dimensions. There are many factors that can be modified in an expression experiment. We suggest beginning with one set of conditions for the first data collection effort and then expanding in various dimensions (e.g., testing a range of temperatures) for subsequent rounds.

(AMRF). H.G.M. and C.J.P. supported by the DOE Joint BioEnergy Institute (http://www.jbei.org), U.S. Department of Energy, Office of Science, Biological and Environmental Research Program, and was part of the DOE Agile BioFoundry (http://agilebiofoundry.org), supported by the U.S. Department of Energy, Energy Efficiency and Renewable Energy, Bioenergy Technologies Office, under Award Number DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U.S. Department of Energy.

Declaration of interests

M.C.J. is a cofounder and/or has financial interest in Stemloop, Inc., Pearl Bio, Gauntlet Bio, National Resilience Inc., and Synolo Therapeutics. These interests are reviewed and managed by Northwestern University and Stanford University in accordance with their conflict-of-interest policies. H.G.M. declares financial interests in XLSI Bio. E.-M.N. is the Head of AI at Myria Biosciences and declares no conflict of interest. D.S. is an adviser to NE47 Bio, Inc.

Supplemental information

Supplemental information associated with this article can be found online https://doi.org/10.1016/j.tibtech.2025.04.021.

References

1. Hon, J. *et al.* (2021) SoluProt: prediction of soluble protein expression in *Escherichia coli*. *Bioinformatics* 37, 23–28
2. Kramer, R.M. *et al.* (2012) Toward a molecular understanding of protein solubility: increased negative surface charge correlates with increased solubility. *Biophys. J.* 102, 1907–1915
3. Govindarajan, S. and Goldstein, R.A. (1997) The foldability landscape of model proteins. *Biopolymers* 42, 427–438
4. Nelson, D.L. and Cox, M. (2021) *Lehninger Principles of Biochemistry: International Edition*, Macmillan Learning
5. Schramm, F.D. *et al.* (2020) Protein aggregation in bacteria. *FEMS Microbiol. Rev.* 44, 54–72
6. de Marco, A. *et al.* (2021) Quality control of protein reagents for the improvement of research data reproducibility. *Nat. Commun.* 12, 2795
7. Gao, K. *et al.* (2020) Theory and applications of differential scanning fluorimetry in early-stage drug discovery. *Biophys. Rev.* 12, 85–104
8. Nikolados, E.-M. and Oyarzún, D.A. (2023) Deep learning for optimization of protein expression. *Curr. Opin. Biotechnol.* 81, 102941
9. Nikolados, E-M. *et al.* (2023) Publisher correction: accuracy and data efficiency in deep learning models of protein expression. *Nat. Commun.* 14, 2838
10. Fu, X. *et al.* (2025) A foundation model of transcription across human cell types. *Nature* 637, 965–973
11. Kryshtafovych, A. *et al.* (2019) Critical assessment of methods of protein structure prediction (CASP)-round XIII. *Proteins* 87, 1011–1020
12. Kryshtafovych, A. *et al.* (2021) Critical assessment of methods of protein structure prediction (CASP)-round XIV. *Proteins* 89, 1607–1617
13. Armer, C. *et al.* (2024) Results of the protein engineering tournament: an open science benchmark for protein modeling and design [preprint]. *bioRxiv*, Published online November 19, 2024. https://doi.org/10.1101/2024.08.12.606135
14. Ackloo, S. *et al.* (2022) CACHE (critical assessment of computational hit-finding experiments): a public-private partnership benchmarking initiative to enable the development of computational methods for hit-finding. *Nat. Rev. Chem.* 6, 287–295
15. Schütz, A. *et al.* (2023) A concise guide to choosing suitable gene expression systems for recombinant protein production. *STAR Protoc.* 4, 102572
16. Kouba, P. *et al.* (2023) Machine learning-guided protein engineering. *ACS Catal.* 13, 13863–13895
17. Ranaghan, M.J. *et al.* (2021) Assessing optimal: inequalities in codon optimization algorithms. *BMC Biol.* 19, 36
18. Niwa, T. *et al.* (2012) Global analysis of chaperone effects using a reconstituted cell-free translation system. *Proc. Natl. Acad. Sci. U. S. A.* 109, 8937–8942
19. Niwa, T. *et al.* (2009) Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of *Escherichia coli* proteins. *Proc. Natl. Acad. Sci. U. S. A.* 106, 4201–4206
20. Berman, H.M. *et al.* (2017) *Protein Structure Initiative - TargetTrack 2000-2017 - all data files*. Published online July 5, 2017. https://doi.org/10.5281/zenodo.821653
21. Notin, P. *et al.* (2023) ProteinGym: large-scale benchmarks for protein fitness prediction and design. *Adv. Neural Inf. Proces. Syst.* 36, 64331–64379
22. Notin, P. *et al.* (2023) ProteinNPT: Improving protein property prediction and design with non-parametric transformers. *Adv. Neural Inf. Proces. Syst.* 36, 33529–33563
23. Li, M. *et al.* (2024) ProSST: Protein language modeling with quantized structure and disentangled attention [preprint]. *bioRxiv*, Published online May 17, 2024. https://doi.org/10.1101/2024.04.15.589672
24. Karim, A.S. *et al.* (2020) In vitro prototyping and rapid optimization of biosynthetic enzymes for cell design. *Nat. Chem. Biol.* 16, 912–919
25. Egloff, P. *et al.* (2019) Engineered peptide barcodes for in-depth analyses of binding protein libraries. *Nat. Methods* 16, 421–428
26. Kim, D.E. *et al.* (2023) De novo design of small beta barrel proteins. *Proc. Natl. Acad. Sci. U. S. A.* 120, e2207974120
27. Remy, I. *et al.* (2007) Detection of protein-protein interactions using a simple survival protein-fragment complementation assay based on the enzyme dihydrofolate reductase. *Nat. Protoc.* 2, 2120–2125
28. Levy, E.D. *et al.* (2014) High-resolution mapping of protein concentration reveals principles of proteome architecture and adaptation. *Cell Rep.* 7, 1333–1340
29. Diss, G. and Lehner, B. (2018) The genetic landscape of a physical interaction. *eLife* 7, e32472
30. Faure, A.J. *et al.* (2022) Mapping the energetic and allosteric landscapes of protein binding domains. *Nature* 604, 175–183
31. Xiao, R. *et al.* (2010) The high-throughput protein sample production platform of the Northeast Structural Genomics Consortium. *J. Struct. Biol.* 172, 21–33
32. Acton, T.B. *et al.* (2005) Robotic cloning and protein production platform of the Northeast Structural Genomics Consortium. *Methods Enzymol.* 394, 210–243
33. Boël, G. *et al.* (2016) Codon influence on protein expression in *E. coli* correlates with mRNA levels. *Nature* 529, 358–363
34. Price, W.N., 2nd *et al.* (2011) Large-scale experimental studies show unexpected amino acid effects on protein expression and solubility in vivo in *E. coli*. *Microb. Inform. Exp.* 1, 6
35. Klesmith, J.R. *et al.* (2017) Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning. *Proc. Natl. Acad. Sci. U. S. A.* 114, 2265–2270
36. Wrenbeck, E.E. *et al.* (2019) An automated data-driven pipeline for improving heterologous enzyme expression. *ACS Synth. Biol.* 8, 474–481
37. Bischof, J. *et al.* (2018) Generation of a versatile BiFC ORFeome library for analyzing protein-protein interactions in live *Drosophila*. *eLife* 7, e38853

38. Tsuboyama, K. *et al.* (2023) Mega-scale experimental analysis of protein folding stability in biology and design. *Nature* 620, 434–444
39. Su, J. *et al.* (2024) SaProt: protein language modeling with structure-aware vocabulary [preprint]. *bioRxiv*, Published online April 19, 2024. https://doi.org/10.1101/2023.10.01.560349
40. Jumper, J. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589
41. Lin, Z. *et al.* (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 379, 1123–1130
42. Alley, E.C. *et al.* (2019) Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* 16, 1315–1322
43. Biswas, S. *et al.* (2021) Low-N protein engineering with data-efficient deep learning. *Nat. Methods* 18, 389–396
44. Dieckhaus, H. *et al.* (2024) Transfer learning to leverage larger datasets for improved prediction of protein stability changes. *Proc. Natl. Acad. Sci. U. S. A.* 121, e2314853121
45. Rives, A. *et al.* (2021) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U. S. A.* 118, e2016239118
46. Hsu, C. *et al.* (2022) Learning protein fitness models from evolutionary and assay-labeled data. *Nat. Biotechnol.* 40, 1114–1122
47. Rao, R. *et al.* (2019) Evaluating protein transfer learning with TAPE. *Adv. Neural Inf. Proces. Syst.* 32, 9689–9701
48. Yang, K.K. *et al.* (2018) Learned protein embeddings for machine learning. *Bioinformatics* 34, 4138
49. Hayes, T. *et al.* (2025) Simulating 500 million years of evolution with a language model. *Science* 387, 850–858
50. Ektefaie, Y. *et al.* (2024) Evaluating generalizability of artificial intelligence models for molecular datasets [preprint]. *bioRxiv*, Published online February 28, 2024. https://doi.org/10.1101/2024.02.25.581982
51. Alford, R.F. *et al.* (2017) The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theory Comput.* 13, 3031–3048
52. Tillquist, R.C. and Lladser, M.E. (2019) Low-dimensional representation of genomic sequences. *J. Math. Biol.* 79, 1–29
53. Kaplan, J. *et al.* (2020) Scaling laws for neural language models [preprint]. *arXiv*, Published online January 23, 2020. https://doi.org/10.48550/arXiv.2001.08361
54. Cheng, X. *et al.* (2024) Training compute-optimal protein language models [preprint]. *arXiv*, Published online November 4, 2024. https://doi.org/10.48550/arXiv.2411.02142
55. Sapranauskas, R. *et al.* (2011) The *Streptococcus thermophilus* CRISPR/Cas system provides immunity in *Escherichia coli*. *Nucleic Acids Res.* 39, 9275–9282
56. Gasiunas, G. *et al.* (2012) Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc. Natl. Acad. Sci. U. S. A.* 109, E2579–E2586
57. Jinek, M. *et al.* (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 337, 816–821
58. Cong, L. *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science* 339, 819–823
59. Mali, P. *et al.* (2013) RNA-guided human genome engineering via Cas9. *Science* 339, 823–826
60. Rocklin, G.J. *et al.* (2017) Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* 357, 168–175
61. Zheng, J. *et al.* (2020) Selection enhances protein evolvability by increasing mutational robustness and foldability. *Science* 370, eabb5962
62. Bloom, J.D. *et al.* (2006) Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. U. S. A.* 103, 5869–5874
63. Wang, T. *et al.* (2018) Continuous directed evolution of proteins with improved soluble expression. *Nat. Chem. Biol.* 14, 972–980
64. Pavlopoulos, G.A. *et al.* (2023) Unraveling the functional dark matter through global metagenomics. *Nature* 622, 594–602
65. Hogg, B.N. *et al.* (2024) The impact of metagenomics on biocatalysis. *Angew. Chem. Int. Ed. Eng.* 63, e202402316
66. Hou, Q. *et al.* (2022) Using metagenomic data to boost protein structure prediction and discovery. *Comput. Struct. Biotechnol. J.* 20, 434–442
67. Levin, D.B. and Budisa, N. (2023) Synthetic biology encompasses metagenomics, ecosystems, and biodiversity sustainability within its scope. *Front. Synth. Biol.* 1, 1255472
68. Ngara, T.R. and Zhang, H. (2018) Recent advances in function-based metagenomic screening. *Genom. Proteom. Bioinform.* 16, 405–415
69. Grossmann, L. and McClements, D.J. (2023) Current insights into protein solubility: a review of its importance for alternative proteins. *Food Hydrocoll.* 137, 108416
70. Miserez, A. *et al.* (2023) Protein-based biological materials: molecular design and artificial production. *Chem. Rev.* 123, 2049–2111
71. Sørensen, H.P. and Mortensen, K.K. (2005) Soluble expression of recombinant proteins in the cytoplasm of *Escherichia coli*. *Microb. Cell Factories* 4, 1
72. Jayakrishnan, A. *et al.* (2024) Evolving paradigms of recombinant protein production in pharmaceutical industry: a rigorous review. *Sci* 6, 9
73. Pouresmaeil, M. and Azizi-Dargahlou, S. (2023) Factors involved in heterologous expression of proteins in *E. coli* host. *Arch. Microbiol.* 205, 212
74. Bhatwa, A. *et al.* (2021) Challenges associated with the formation of recombinant protein inclusion bodies in *Escherichia coli* and strategies to address them for industrial applications. *Front. Bioeng. Biotechnol.* 9, 630551
75. Radivojević, T. *et al.* (2020) A machine learning automated recommendation tool for synthetic biology. *Nat. Commun.* 11, 4879
76. Zhang, T. *et al.* (2020) Regulating strategies for producing carbohydrate active enzymes by filamentous fungal cell factories. *Front. Bioeng. Biotechnol.* 8, 691
77. Ntana, F. *et al.* (2020) *Aspergillus*: a powerful protein production platform. *Catalysts* 10, 1064
78. Karbalaei, M. *et al.* (2020) *Pichia pastoris*: a highly successful expression system for optimal synthesis of heterologous proteins. *J. Cell. Physiol.* 235, 5867–5881
79. Barone, G.D. *et al.* (2023) Industrial production of proteins with *Pichia pastoris – Komagataella phaffii*. *Biomolecules* 13, 441
80. Souza, C.C. de *et al.* (2021) The multifunctionality of expression systems in *Bacillus subtilis*: emerging devices for the production of recombinant proteins. *Exp. Biol. Med.* 246, 2443–2453
81. Cairns, T.C. *et al.* (2021) Something old, something new: challenges and developments in *Aspergillus niger* biotechnology. *Essays Biochem.* 65, 213–224
82. Liu, D. *et al.* (2023) Heterologous protein production in filamentous fungi. *Appl. Microbiol. Biotechnol.* 107, 5019–5033