

REVIEW

Open Access



# Recombinant expression of insoluble enzymes in *Escherichia coli*: a systematic review of experimental design and its manufacturing implications

Suraj Mital<sup>1</sup>, Graham Christie<sup>1</sup> and Duygu Dikicioglu<sup>2\*</sup>

## Abstract

Recombinant enzyme expression in *Escherichia coli* is one of the most popular methods to produce bulk concentrations of protein product. However, this method is often limited by the inadvertent formation of inclusion bodies. Our analysis systematically reviews literature from 2010 to 2021 and details the methods and strategies researchers have utilized for expression of difficult to express (DtE), industrially relevant recombinant enzymes in *E. coli* expression strains. Our review identifies an absence of a coherent strategy with disparate practices being used to promote solubility. We discuss the potential to approach recombinant expression systematically, with the aid of modern bioinformatics, modelling, and 'omics' based systems-level analysis techniques to provide a structured, holistic approach. Our analysis also identifies potential gaps in the methods used to report metadata in publications and the impact on the reproducibility and growth of the research in this field.

**Keyword:** *Escherichia coli*, Recombinant industrial enzymes, Difficult to express enzymes, Inclusion bodies, Systems biology, Solubility

## Background

Enzymes serve a wide range of biocatalytic purposes across multiple key industrial sectors; our observation shows the food and beverage, pharmaceutical/health-care, chemical, starch and paper processing, detergent, bioremediation, textile, agriculture, biosensor, and waste management industries have the highest usage (Fig. 1A). The total biocatalysis market is a rapidly growing sector of industrial biotechnology with an estimated global market value projected to reach \$10 billion by 2024 [1]. A review of the literature demonstrates a growing amount of research dedicated to discovering, isolating, and

characterizing novel enzymes; this research is driven by a demand for enzymes that can replace current catalysts that show limited functional stability at specific operational conditions such as increased temperature or pH. Enzymes furthermore serve as tools to lessen the environmental impact of chemical processes traditionally driven by inorganic catalysts leading to 'greener' manufacturing [2].

Heterologous expression of a recombinant product is the preferred strategy when sufficient quantities of an enzyme of interest cannot be achieved in the native host organism. This method provides an efficient and economically favorable method to produce high quantities of recombinant protein in a relatively short amount of time. The industrial manufacturing sectors often adopt *E. coli* as a heterologous host for protein expression to facilitate rapid product. A common, challenging caveat to

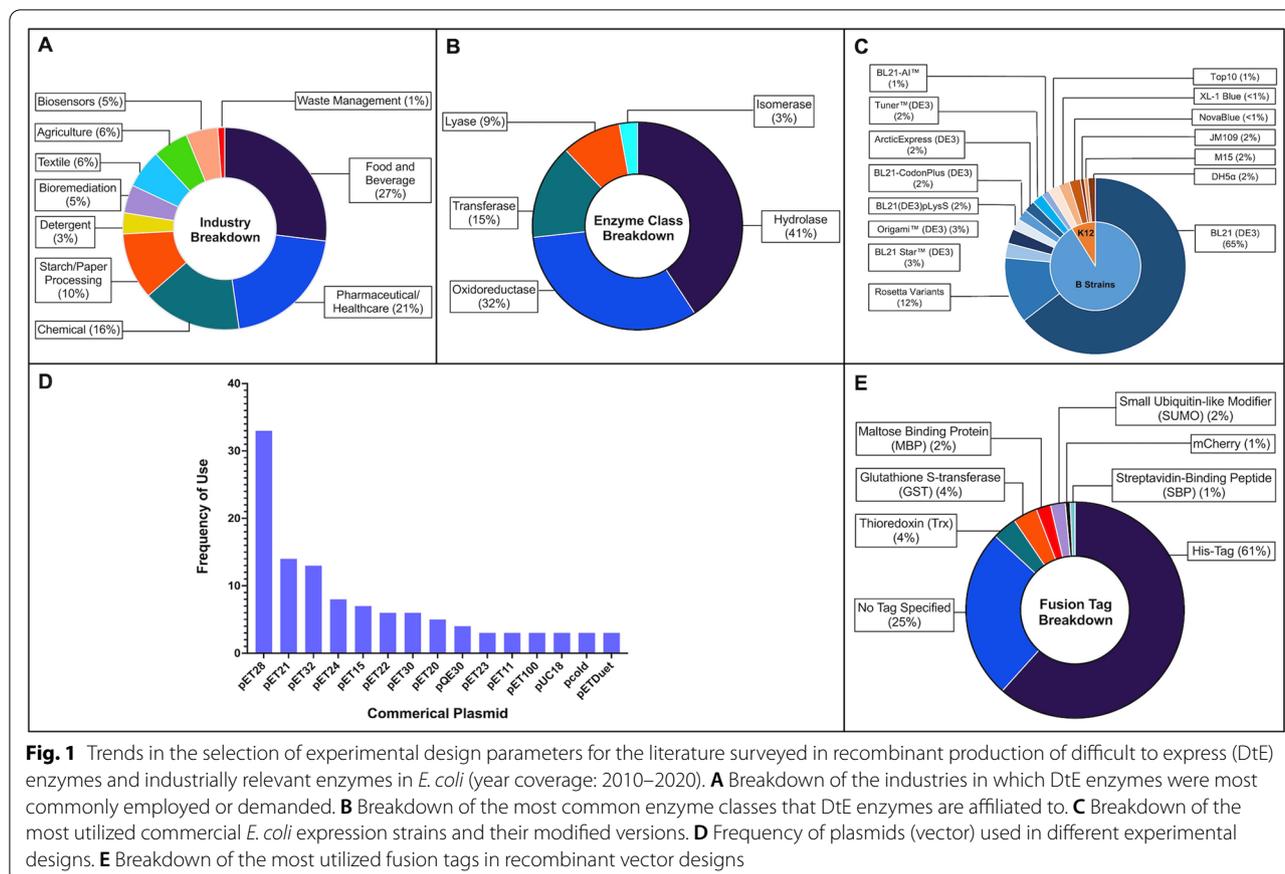
\*Correspondence: d.dikicioglu@ucl.ac.uk

<sup>2</sup> Department of Biochemical Engineering, University College London, London WC1E 6BT, UK

Full list of author information is available at the end of the article



© The Author(s) 2021. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.



this expression method is the high likelihood of generating inclusion bodies due to protein misfolding. Inclusion bodies are aggregated masses of misfolded or partially folded peptide chains that can result from a variety of factors including but not limited to: when the rate of protein synthesis in vivo surpasses the capabilities of the cell, lack of eukaryotic chaperones for specific proteins, reduced cytosol environment, and limited post-translational machinery [3]; this is often the case when overexpressing a protein product in a recombinant expression system [4]; this misfolded state can be inhibitory to the biocatalytic capability of the enzyme and as a result, solubility is a property highly valued in the manufacturing supply chain. Over the past few decades, protocols have been modified by introducing different experimental design strategies to instigate the production of soluble products. These strategies explore variations in regulatory sequences (promoters), plasmid backbones, strains of *E. coli*, fusion partners employed, incubation temperatures, medium components, chaperone proteins, and inducer concentrations to name a few common variables.

In the event that these strategies prove unsuccessful, an extra refolding/renaturation and purification step is often necessary to generate a soluble, functional enzyme

[5]. However, stepwise protocol of producing inclusion bodies and subsequent solubilization has proven to be a viable strategy to generate a higher volume of product. Recombinant protein within inclusion bodies has been found to occupy 30–40% of the total cellular proteins [6]; furthermore, inclusion bodies can be comprised of up to 90% pure recombinant protein [7].

A caveat to the solubilization methodology is that it is not a one size fit's all strategy and often requires case-by-case protocols to be developed that cannot be widely applied all protein types. For instance, multi-copper lacases from four distinct organisms: *Bacillus* sp. HR03 [8], *Geobacillus* sp. JS12 [9], strains of *Yersinia enterocolitica* [10], and *Bacillus subtilis* strain R5 [11], though similar, prove to have unique purification protocols in each respective study. Furthermore, a loss in the secondary structure after exposure to strong denaturants can lead to a reduction in the overall bioactivity of the nascent protein [12]. The case-by-case specific nature of solubilization does not guarantee a final protocol fit for industrial use. Solubilization was shown to result in a reduced final recovery, from 50% or less of bioactive product in some cases [13] to no biologically active product in other examples [14]. At bench-scale, this is a small price to pay

to generate protein that is appropriate for structural and or functional characterization. However, at industrial scales, the overarching costs associated with capacity underuse combined with operational costs including but not limited to consumables, utilities, and personnel time in addition to extensive protein quality control often render this option infeasible.

Current economic analysis associated with industrial scale manufacturing and downstream processing is limited in this field; there is a need for an updated techno-economic analysis of the processes discussed previously to reflect their current costs for the industrial biotechnology sector. Depending on the method utilized and the scale of operation; in 2011, the direct fixed costs and labor associated with this additional treatment was reported to add up to \$8.2 million to implement and operate inclusion body solubilization equipment for individual companies [15], however these costs may be higher today. As a result, it is highly desirable to produce an enzyme of interest in a soluble state from the start to achieve cost-effective production.

### Scope of this review

A survey of the literature in this field from the past decade has revealed no standardized method developed to promote solubility for enzymes expressed through recombinant technology. This review identifies trends in the experimental design for recombinant expression studies, in the industrial biotechnology sector, that ultimately generated inclusion bodies in *E. coli*. Our analysis identified which methods or tools, if any, were employed in designing the recombinant expression system and the impact on the mitigation of inclusion bodies. Our goal was to highlight the factors/strategies researchers tend to prioritize and provide a measure for their popularity implicated by the frequency of their use. Our analysis was focused on work published in the field of industrial biotechnology since 2010. Manufacturing of numerous recombinant protein products, including those of biopharmaceutical use such as growth factors, antibodies, or cytokines had historically been in the remit of recombinant protein production by *E. coli*. Several well-cited reviews exist on the subject, which address the challenges associated with using *E. coli* for these purposes [16–18]. Over the past decade, Chinese Hamster Ovary (CHO) cells have increasingly dominated the manufacturing process as hosts for expressing protein-based biologics, therapeutics, and other relevant eukaryotic proteins. CHO cell-based systems are currently used to manufacture up to 84% [19] of approved biopharmaceuticals as opposed to 30% in *E. coli* [16]. This was primarily following the efforts towards the sequencing the CHO genome and its subsequent publication in 2011 [20]. Nevertheless

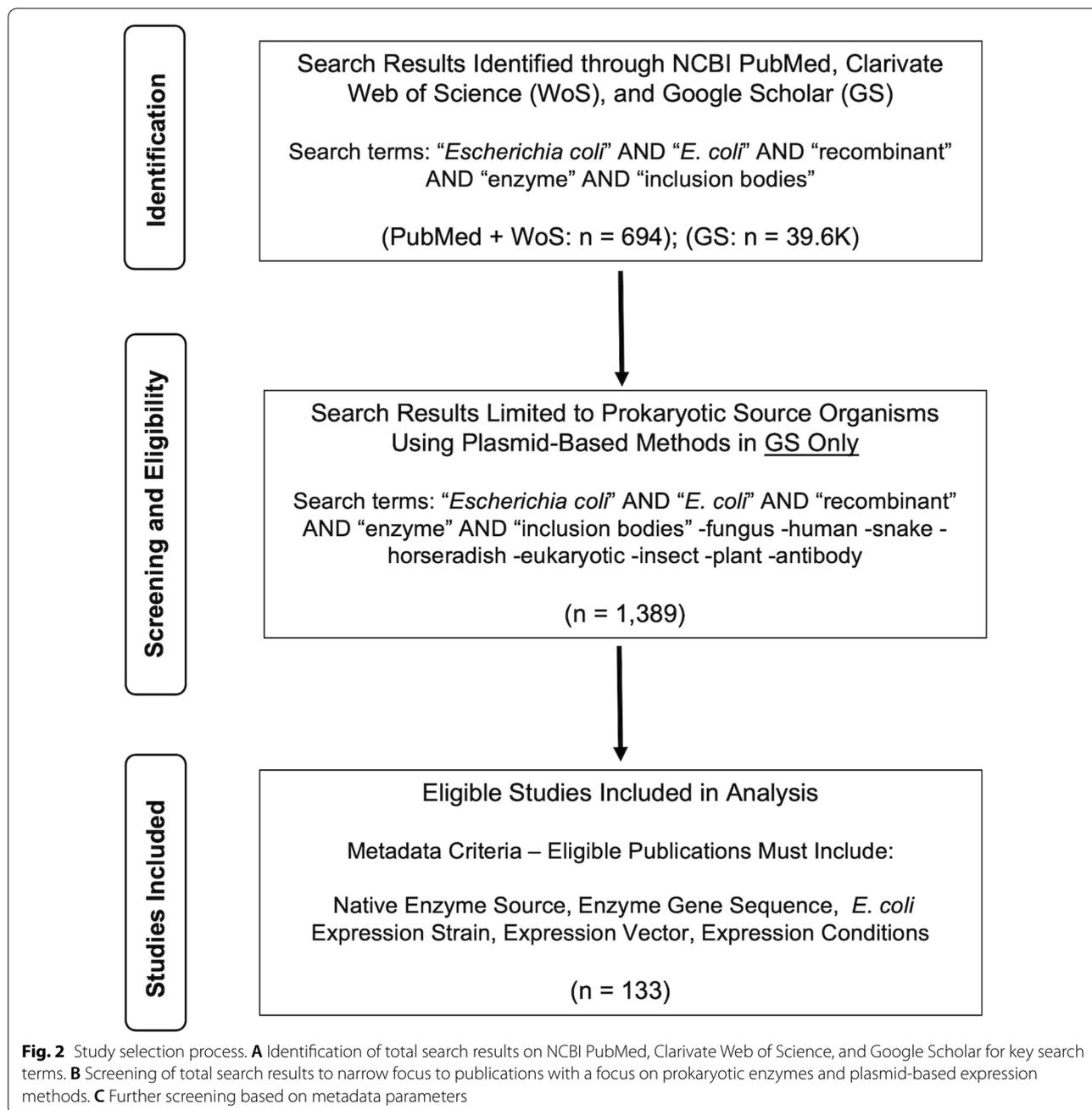
*E. coli* has remained the standard workhorse for industrial biotechnology applications. Our analysis, therefore, focused on enzymes of prokaryotic origin and their plasmid-based expression in *E. coli*. These enzymes are industrially valuable due to their sustained performance in non-conventional niche environments owing to the wide distribution prokaryotes in adverse or unique habitats [21]. This reallocation of dominant sector preference for specific hosts in the manufacturing of different types of proteins has necessitated that even more attention needed to be paid to the improvement of host strains and expression systems of *E. coli* tailored to specialized applications. However, we still fail to see any systematic effort to streamline the development of efficient expression systems that overcome the insolubility of enzymatic proteins. The analysis conducted in this systematic review can serve this purpose and act as a starting point for future experimentation in the field.

### Identification of studies and selection

Databases such as NCBI PubMed and Clarivate Web of Science (WoS) provide a vast number of examples of scholarly literature that demonstrate the widespread prevalence of inclusion body formation. When used together for both databases, the search terms ‘recombinant’, ‘enzyme’, ‘inclusion bodies’, ‘*E. coli*’ and ‘*Escherichia coli*’ yielded 1891 publications, of which 659 were from the past decade (access June 2021). The omission of ‘*Escherichia coli*’ from these key words adds only 678 additional results indicating that *E. coli* continues to serve as the standard production workhorse and is discussed in 64% of recent recombinant enzyme expression work in this field.

Publications for this review were sourced from NCBI PubMed and WoS databases, in addition to Google Scholar. Relevant articles were published between 1 January 2010 and 31 June 2021. Search terms included those for *Escherichia coli* (including the terms ‘*Escherichia coli*’ and ‘*E. coli*’), ‘recombinant’, ‘enzyme’, and ‘inclusion bodies’, generating a total of 501 results in PubMed, 193 result in WoS and 39.6 K search results in Google Scholar (Fig. 2). The difference in the number of search results generated between PubMed/WoS versus Google Scholar is due to the algorithm employed. A large proportion of the search results on Google Scholar have been found to be ‘grey literature’—a term that encompasses books, book chapters, patents, theses, non-peer reviewed research, and/or ambiguous citations that do not fall within a specific category, in addition to duplicate search results [22]. This volume of grey literature accounts for the discrepancy of search results generated.

These papers were manually curated within the search results for PubMed and WoS. Logical operators were



used to remove irrelevant search results from the large number of results returned in Google Scholar specifically leaving 730 results; this curation process is highlighted in Fig. 2. The metadata reported by a given study is required for future reproducibility. Therefore, our literature search for this review was limited to those studies that explicitly reported: (i) the native source of the enzyme of interest, (ii) gene sequence, (iii) the *E. coli* expression strain, (iv) the expression vector backbone, and (v) expression conditions (i.e. expression temperature, inducer

concentration, cell density). These details are deemed essential for reproducibility in this field. 21% of the publications, which were initially thought to be relevant to the scope of this review, were omitted from further investigation due to the lack of at least one of the aforementioned details (Fig. 2). In total 133 publications that expressed 140 enzymes were selected between 2010 and 2021; some publications highlighted multiple enzymes leading to a higher number of enzymes compared to publications. A full list of this literature can be found in Additional file 1.

Interestingly, all enzymes were identified to be relevant within industry or for direct commercial sale, although this aspect was not explicitly mentioned in their respective publications.

We observe that the food and beverage, pharmaceutical, and chemical industries showed the largest demand for bulk biocatalyst manufacturing and had the greatest variety of enzymes employed in their processes during the period of investigation (Fig. 1A). Hydrolases (41%) and oxidoreductases (32%) were the most prevalent classes of enzymes studied in the publications investigated within the scope of this review, followed by transferases (15%), lyases (9%), and isomerases (3%); none of the enzymes recombinantly expressed in these studies belonged to the class of ligases (Fig. 1B). This breakdown suggests that hydrolases and oxidoreductases are of substantial industrial interest, however it is difficult to interpret from this data whether these classes have a higher propensity to form inclusion bodies in comparison to other enzyme classes.

#### Data reporting and limitations

The narrative of published research targets a selective audience in its cognate field of work, for whom each detail regarding the experimental design may not necessarily be of interest. Our analysis identified an absence of methodical organization, or a system to report experimental design details. The aforementioned details were reported in all 133 studies, however the curation process for this review required a comprehensive search of the introduction, methods, and supplementary text of each publication. This method of reporting metadata is inefficient, limiting reproducibility by the scientific community post-publication. Research narratives are centered around their principal objective; this objective guides the details emphasized and their priority within the text. For example, in characterization studies, where the primary goal was to report on enzyme structure, less emphasis may be placed on the expression conditions used to produce of the enzyme of interest. This bias was observed in the past in The Protein Data Bank (PDB), the major repository for protein structural datasets. Zhou et al. reported that expression host information for recombinant studies were omitted from 62 (12%) of the PDB study examples they selected for their analysis [23]. In another case, Magnan et al. identified that the PDB and the SwissProt databases were observed to report on the solubility of recombinant proteins but neglect to include the experimental conditions that were used; this was critiqued as the information provided by the retrieved datasets retained inconsistencies rendering it ineffective for modeling purposes [24]. The omission of such details can often hamper follow-up work by scientists in other fields.

One other aspect of this work worth noting was only one out of the 133 studies shortlisted was conducted through a collaborative partnership with industry [25]. The remaining studies were led by either academic institutions or government agencies. This underrepresentation of industrial contribution does not truly reflect the economic challenges inclusion body formation poses on commercial manufacturing [26]. However, this contributes a bias to the information currently available in the public domain. The issues highlighted above demonstrate that a degree of implicit bias currently limits our understanding to facilitate reproducible research. The evaluation discussed below should be considered in light of these limitations.

#### Enzyme discovery and in vitro expression

Recently, the process of enzyme discovery has been led by metagenomics and genome mining from environmental microbiome samples [27–29]. Screening a range of habitats with varying physical environmental conditions have led researchers to uncover organisms with evolutionarily adapted tolerance to these conditions. Higher temperatures, for example, in the case of *Thermus thermophilus* HB9 [30], cold temperatures, in the case of *Shewanella arctica* [31], or elevated salt concentrations in the case of *Halobacterium salinarum* [32] are examples of adapted bacteria. Researchers have observed that enzymes isolated from these unique prokaryotic hosts share a similar tolerance to environmental conditions in vitro and therefore have use in industrial manufacturing conditions that are typically harsh in comparison to the growth environment of many organisms. In other instances, in vitro directed evolution through mutagenesis or computationally driven protein engineering can serve as alternative methods to develop stable or functionally novel enzymes within a laboratory setting [33]. Directed evolution, for example, was used to expand the substrate range of *Rhodococcus* phenylalanine dehydrogenase for the highly enantioselective reductive amination of ketones to amines [34].

While the enzymes from extremophiles are of industrial interest, extremophilic bacteria, for example, are often difficult to culture in a laboratory setting due to challenges associated with providing culture conditions that adequately mimic their native environment [35–43]. These additional considerations create a non-standard set of culture conditions. Moreover, enzymes isolated from organisms found in complex environmental samples pose additional challenges as optimal culture conditions for many organisms are typically not well-understood. Such a challenge was reported for a strain of *Aeromonas hydrophila* isolated from sludge samples collected from textile wastewater treatment plants; the complex

chemical composition of this sludge rendered it difficult to determine a suitable, replicable medium formulation to support growth of the organism in a laboratory setting [44]. Therefore, the process of generating suitable high quantities of active enzyme from a novel isolated prokaryotic source is not always a straightforward task.

In other cases, certain enzymes have been identified as being toxic to host cells when overexpressed at high levels. Microbial transglutaminase (MTGase) monomers from *Streptoverticillium mobaraensis*, for example, were shown to have a high tendency to cross-link and oligomerize during intracellular expression [25]. MTGase was also observed to be expressed at low abundance in *S. mobaraensis*—at concentrations inadequate for subsequent experimental work—which is an issue frequently observed in similar studies [45–49]. Additionally, enzyme purification methods from native host sources often require a combination of methods such as salt precipitation and/or a series of chromatographic separations to deliver a product with sufficient purity. However, these methods are often difficult and economically challenging at scale, particularly when purifying enzymes expressed in limited quantities [50, 51].

### Current experimental design practices

#### Expression strains—K12 and B strains

A majority of the available *E. coli* expression hosts in use for recombinant protein expression typically fall under two categories: K12 or B derivative strains. BL21 (DE3) is one of the most common B strains preferred for recombinant expression; often, it is preferentially selected over its K12 counterparts as an ideal recombinant expression host due to several key advantages. BL21 (DE3) is deficient in Lon and OmpT proteases, thus providing a layer of protection to misfolded proteins that would normally be targets for degradation [52]. A short doubling time of approximately 20 min coupled with rapid protein synthesis via the T7 expression system generates a stable protein product at high titers [53]. K12's longer growth times and predisposition to produce acetate generates lower biomass in comparison to BL21 (DE3) [54].

Our analysis observed B strains were the most widely employed for enzyme expression in 88% of the total cases. Among those, BL21 (DE3) was selected as the primary expression host in 65% of our reference cases (Fig. 1C). The remaining 12% utilized K12 derivatives. Commercially available K12 strains used include JM109 [25, 51, 55], DH5 $\alpha$  [30, 56], NovaBlue [57], XL1 Blue [58], M15 [31, 59, 60], and Top10 [61, 62]; alternatively non-commercial K12 strains W3110 [63] and MG1655 [62, 64] were used in a limited number of cases. K12 strains serve as useful tools when plasmid instability is encountered resulting in plasmid loss from the host [54]; this may explain its use

in limited cases for expression. Beyond protein expression, K12 strains are also mentioned as tools plasmid propagation and cloning. For example, Vadala et al. [65] propagated their pET21a vector in DH5 $\alpha$  however their eventual expression host was BL21 (DE3). A breakdown of commercial strains used, and their applications can be found in Table 1; a full list of expression strains used can be found in Additional file 1: Table S1.

#### Specialized *E. coli* strains—emerging tools for protein expression

The challenging nature of DTE enzymes has led to the adoption of alternative hosts to improve performance, with several companies developing strains that researchers are opting for in favor of traditional strains. These BL21 (DE3) variants were the strains of choice in 30% of the selected expression studies. Industrial manufacturers have evolved the efficiency of their strains by producing variants capable of reducing the common causes of aggregation that manifest during overexpression (Table 1). These strains are marketed to have superior performance that are well-suited to mitigating issues concerning:

- i. High disulfide bond formation—BL21 Origami B [66]
- ii. Codon Bias BL21 CodonPlus [67]; BL21 Rosetta
- iii. Temperature instability—BL21 ArcticExpress [14, 31]
- iv. Toxic proteins—BL21 AI [68, 69], BL21 Tuner [70–72].
- v. Low expression yield—BL21 pLysS [14, 73] and BL21 Star [74, 75].

The reason behind the selection of a specific strain is not overtly mentioned in the text in a majority of the cases investigated, but rather it may be inferred from the broader context of the study. The researchers' selection could also be led by previous experience or through justification using structural bioinformatics to gain insight as to how a protein of interest may behave in vivo. Many options exist for specific expression issues, and therefore these strains are adopted on a case-by-case basis. BL21 Origami (DE3), for example, was identified to be an ideal starting strain for the expression of maltooligosyl trehalose trehalohydrolase (MTHase) as the protein's structure and, in turn, enzyme function was impacted by disulfide linkages [66]. Likewise, pLysS strains of BL21 showed promise to produce soluble levels of toxic proteins such as certain metalloproteins [76]. In other instances, the use of strains such as ArcticExpress are motivated by promoting solubility through facilitating expression at

**Table 1** Commercial *E. coli* expression strains employed for the production of DtE enzymes (2010–2021)

Expressed industrial enzyme example	<i>E. coli</i> expression strain	Commercial supplier	Benefit	References
Bis-γ-glutamylcystine	ArcticExpress (DE3) (B)	Agilent Technologies, Inc	Expression at low temperatures with active molecular chaperones. Promotes ideal folding under low temperature conditions, increasing solubility	[183]
Malto-oligosyltrehalose trehalohydrolase	Origami™ B (DE3) (B)	Merck KGaA	Expression of proteins rich in in disulphide bonds. Promotes cytoplasmic disulphide bond formation	[66]
Phenol hydroxylase component 2	BL21-CodonPlus (DE3) (B)	Agilent Technologies, Inc	Expression of proteins rich in AGA/AGG, AUA, and CUA. Promote correct synthesis and folding for proteins with rare codons in <i>E. coli</i>	[184]
Flavin reductase	BL21(DE3)pLysS (B)	Various Suppliers (Thermo Fisher Scientific, Promega Corporation, Merck KGaA)	Lower background expression, useful for toxic proteins. Allows greater control over expression	[73]
Arylamine <i>N</i> -acetyltransferase				[14]
Beta-glucosidase	One Shot™ BL21 Star™ (DE3) (B)	Thermo Fisher Scientific	High expression of non-toxic proteins. Promotes high expression of protein product	[74]
<i>N</i> -Acyl-D-glucosamine 2-epimerase	Tuner™(DE3) (B)	Merck KGaA	Allows slower protein synthesis to promote solubility using adjustable inducer concentrations	[71]
Sphingomyelinase	Rosetta/Rosetta 2 (DE3) (B)	Merck KGaA	Promote correct synthesis and folding for proteins (i.e. eukaryotic proteins) with rare codons (AGA, AGG, AUA, CUA, GGA, CCC, and CGG) in <i>E. coli</i>	[79]
Lysine 6-Dehydrogenase	Rosetta-gami™ (DE3) (B)	Merck KGaA	Alleviates codon bias and enhances disulphide bond formation	[83]
2-hydroxyethyl-phosphonate methyltransferase	Rosetta™ 2 (DE3)pLysS (B)	Merck KGaA	Alleviates codon bias and lowers background expression, useful for toxic proteins	[84]
Tryptophan-2-C-methyltransferase	RosettaBlue™(DE3) pLysS (B)	Merck KGaA	Alleviates codon bias and lowers background expression, useful for toxic proteins. High transformation efficiency	[78]
Fructose-1,6-bisphosphate aldolase	DH5α (K12)	Various Suppliers (Thermo Fisher Scientific, New England Biolabs, Gold Biotechnology Ltd.)	Generally, a strain used for cloning and blue/white screening. (recA) mutation allows for better insert stability	[30]
Chitobiase	M15 (K12)	Qiagen	Generally used in conjunction with plasmid (pQE) found within a expression kit from Qiagen	[59]
Cellobiose phosphorylase	JM109 (K12)	Promega Corporation	Generally, a strain used for cloning and blue/white screening	[55]
Quinoprotein glucose dehydrogenase B	NovaBlue (DE3) (K12)	Merck KGaA	Generally, a strain used for cloning and blue/white screening	[57]
Fucose-1-phosphate aldolase	XL-1 Blue (K12)	Promega Corporation	Generally, a strain used for cloning and blue/white screening	[58]
Trehalose transferase	TOP10 (K12)	Thermo Fisher Scientific	Generally, a strain used for cloning and plasmid propagation	[61]

low temperatures, which is a widely accepted strategy that is implemented to control the rate of synthesis.

In addition to the strains mentioned above, Rosetta, which has many variants available (Table 1), has a growing popularity as an *E. coli* host for DtE bacterial

enzymes, engineered to have an increased tRNA supply for such codons as AUA, AGG, AGA, CUA, CCC, GGA, which are less abundant in *E. coli* [77–80]. Originally this strain was utilized for the expression of complex eukaryotic proteins [81]. However, it is often thought that

tRNA limitations can factor into the formation of inclusion bodies [82]. Other studies utilized modified strains of Rosetta such as Rosetta-gami-pLysS [83] and Rosetta pLysS [84], which bring additional benefits of the Origami and pLysS systems together with the enhanced tRNA supply of Rosetta (Table 1). Other enhanced expression strains, include BL21 Lemo21 [85] for proteins potentially displaying toxicity effects such as membrane proteins, currently exist in the market but are not included in Table 1 since these tools were not used as expression hosts for the studies selected within the scope of this review.

Specialized BL21 (DE3) *E. coli* strains were utilized in only 30% of the studied surveyed here and in a large majority of cases, these specialized strains had little effect in the context of their respective studies. It often appears that the success of such strains varies on a case-by-case basis based on the properties of the protein and the additional expression parameters used. Soluble expression of MTHase from *Sulfolobus acidocaldarius* was improved by 40% when this protein was expressed in *E. coli* Origami (DE3) as opposed to BL21 (DE3); this enzyme's structure is dependent on disulfide linkages for proper folding. This study noticed the combination of *E. coli* Origami (DE3) in addition to using thioredoxin (Trx) as a fusion partner on pET32a improved the folding propensity of MTHase [66]. Likewise, the use of maltose binding protein (MBP) coupled with rare tRNAs found in *E. coli* Rosetta (DE3) improved the expression of prenyltransferase (NovQ) by up to 50% [86].

In each of the cases described above, solubility was improved in combination with another experimental factor. It is therefore difficult to discern what the true impact of the specialized strains on the final recombinant product. It is possible that a combination of such strain and experimental design parameter combinations impact the final end-product. BL21 CodonPlus (DE3)-RIL on its own demonstrated no improvement for *Bacillus subtilis* strain R5 laccase, however a 30% improvement was observed when the expression temperature was lowered from 37 to 17 °C [11]. We require larger amounts of aggregated metadata to draw such accurate conclusions.

Despite their success in previous studies, topoisomerase I from *Mycobacterium tuberculosis* when expressed in BL21 Arctic Express [87] or DAHP synthase from *M. tuberculosis* when expressed in BL21 RosettaGami [88], specialized strains remain an underutilized potential solution to address difficulties around the formation of inclusion bodies during heterologous expression of Dte enzymes (Fig. 1C). We observe that no specific specialized strain was preferred over the others; the variants of Rosetta, when treated as group, were the second most prevalent strain after BL21 (DE3) accounting for 13%.

It is possible that specialized strains have found limited use to date as they are a relatively recent development in comparison to the first reports of BL21 (DE3); Studier and Moffatt introduced BL21 (DE3) in 1986 [89], whereas the specialized strains were developed years later; BL21 Origami in 2001 [90], BL21 Star in 2002 [91], and Rosetta-gami B in 2005 [92], offering a possible explanation for their current underutilization. One other possible reason could be that although modified strains have been designed to alleviate the impact of specific challenges persistently faced in the expression of Dte enzymes, they may have limited application areas outside the scope of their tailored use. This coupled with the additional costs to purchase each individual strain make this process economically unviable for many laboratories. For these reasons, it is likely that long-established research environments prefer to opt for the tools (i.e. expression strains) that are readily available within their existing practice.

In 58% of the papers we reviewed, the research narrative did not overtly mention experimental factors taken into consideration to promote solubility and rather focused on their downstream solubilization methodology. From our observation, only the approaches that led to the final protein of interest were detailed in each paper, while other unsuccessful attempts may have been omitted.

#### Plasmid considerations for Dte enzyme expression

Whereas the previous section focused on different strains available for expression experiments there are actually far more variants of plasmid vectors employed in the past decade for the expression of difficult and problematic enzymes in *E. coli*. We observed a large demographic of vectors utilized throughout. pET vectors were by far the most commonly employed (in 66% of cases), and they were primarily utilized in conjunction with DE3 strains since the T7 RNA polymerase gene of DE3 is required for efficient synthesis of sequences downstream of T7/Lac hybrid promoters present in many pET plasmids [93]. pET28 was the most popular choice of vector, featuring in approximately 31% of the studies that utilized pET derivatives, followed by pET21 [8, 29, 36, 94–96], pET32 [97, 98], pET24 [99, 100], and pET15 [48, 77] (Fig. 1D).

The remaining 34% of plasmids did not fall into a singular group with high use and were uniquely employed in the study reporting them in varying frequencies. We observe that certain plasmids were used for specific purposes; a similar observation to that of the specialized strains discussed previously. The pCold(I–III) plasmid set was used for expression at cold temperatures i.e., whereby expression is induced by the cold-shock response [83, 101, 102]; this plasmid was used in conjunction with

ArcticExpress (DE3) for suitable expression at temperatures below 13 °C. pACYC-Duet-1, with its two multiple cloning site locations, was utilized for the co-expression of native chaperone proteins from *Pyrococcus furiosus* as a strategy to promote proper folding of an  $\alpha$ -amylase [103]. A full list of expression vectors used can be found in Additional file 1: Table S1.

Tight basal expression control of proteins appeared to be a key factor that was prevalent among the literature. In most instances, this control was managed through lower concentrations of inducer compounds such as 0.1–0.5 mM IPTG for inducible promoters such as T7/lac [6, 59, 65, 72, 104–106]. Soluble expression of *Bacillus acidopullulyticus* pullulanase was highly dependent on the control of basal gene expression that was only achieved in pET22b/pET28a harboring an inducible T7/lac promoter as opposed to pET20b that contained a constitutive T7 promoter region [104]. Tighter regulatory control was observed when a PHsh vector was used to moderately enhance the solubility of *Thermus thermophilus* HB27 pullulanase. PHsh contains a synthetic heat-shock promoter (Hsp); proteins synthesized on this vector system are regulated by the heat-shock transcription factor  $\sigma^{32}$ . It was observed that *E. coli* JM10 reached a higher cell density and displayed a lower stress response in comparison to expression using T7/lac in BL21 (DE3). To contrast, however, solubility of *Mesorhizobium loti* carbonic anhydrase was enhanced using the J23100 constitutive promoter in combination with a N-terminal TrxA fusion on a pSUM backbone and co-expression of GroEL/ES; this improvement was in comparison with a similar experimental design with pET28a and pET32a using an inducible promoter system [107].

These examples represent a small fraction of the options available for expression plasmids, with variants available from different manufacturers. The vast number of vector-expression strain combinations allow for a range of different possibilities for researchers to customize their experimental designs. Customarily, commercial suppliers offer plasmids with different combinations of promoters, selection markers, multiple cloning sites, and fusion tags adding further to the myriad of combinations a researcher can ultimately choose to utilize in their designs. Discussion of factors contributing to the final design were extremely limited; this further contributes to the challenge of facilitating rational decision making for designing protein expression studies.

#### Use of fusion tags in construct designs

77% of publications discussed the use of at least one fusion tag within their design; polyhistidine tags (His-tag) comprised 83% of all tags used. Fusion tags are essential tools for protein recovery as well as for the

analytical quantification of products. Commercial plasmids often contain peptide sequences encoding tags that can be used for purification, act as reporter genes, or promote solubility. pSUMO/Champion™ pET SUMO Expression System are examples of plasmids engineered to contain a native SUMO tag to enhance the solubility of fused proteins [78, 108, 109]. A majority of pET vectors such as pET28, pET15, and pET21 contain incorporated His-tags, adjacent to multiple cloning sites, that can be used for downstream purification using immobilized metal affinity chromatography columns (see manufacturer's manual, Novagen, accessed June 2021). This could explain why we observe a larger use of pET expression vectors and His-tags in these enzyme characterization studies. However, it is important to note that His-tags can be incorporated independently of features encoded on plasmids using specific primers encoding the His-tag sequence.

Apart from facilitating affinity purification [110–112], fusion tags can serve essential roles to enhance and promote the solubility of difficult to express enzyme constructs. In a small fraction of the studies reviewed here (16%) peptide tags such as thioredoxin (Trx) [66, 98, 113], glutathione S-transferase (GST) [114], small ubiquitin-related modifier (SUMO) [108], and maltose binding protein (MBP) [70] were used to promote solubility (Fig. 1E). These values were calculated based on whether the text mentioned the tag in their narrative. As mentioned previously pET32a, for example, contains a TrxA tag, however the number publications mentioning both pET32a and TrxA is not equally proportional. Therefore, the frequency of use for these fusion partner proteins may be higher in reality; it is difficult to interpret based on current metadata reporting practices.

The widespread use of the His-tag system suggested that bench-scale enzyme characterization studies prioritized producing an enzyme product regardless of its physicochemical state; we observe these publications mention protocols for refolding inclusion bodies back to their native state following downstream purification. However, the preventative steps considered to avoid the formation of inclusion bodies in the first place were not explicitly mentioned. Furthermore, in some cases production of inclusion bodies was a preferred strategy for ease of purification or as the only means to achieve large amounts of protein for characterization. However, as mentioned previously, this approach is not economically feasible for industrial scale manufacturing.

Suppliers such as Novagen provide tags within a majority of their constructs for fusion at the N-terminal (see manufacturer's manual, Novagen, accessed June 2021). We observed a twofold stronger preference for N-terminal allocation of the His-tag than for C-terminal

allocation in the reviewed literature [58, 104, 114–116]. C-terminal fusions were found in a limited number of cases [35, 40, 42, 51, 57, 117–119]. However, the discussion on the factors contributing to this predilection was limited; our analysis of the reviewed literature did not indicate a strong benefit received from either choice in terms of the physical condition of the final product. The choice of tag location is typically guided by an enzyme's structure, as to not interfere with the active site during catalysis. Fusion partners attach additional peptide residues to the construct. This could, in some cases, increase the possibility of misfolding due to the increased size of the construct, although this was not explicitly acknowledged or studied in detail. Primarily, studies using His-tags did not mention downstream removal of the tag from the protein of interest; generally larger fusion partners such as MBP [86], TrxA [120], or SUMO [109] were removed via protease cleavage sites following purification.

Our analysis suggests that the adoption of fusion partners is mainly for purification and is not a widely utilized technique to promote solubility of DtE enzymes despite past successes [121, 122]. The few instances, mentioned previously, showed MBP and TrxA promoting solubility in conjunction with other experimental parameters like temperature, inducer concentration, and/or expression strain. However, it is difficult to discern whether these fusion partners have merit for a wider range of proteins. A full list of fusion tags used can be found in Additional file 1: Table S1.

### **Role of systems biology in addressing the challenges around the formation of inclusion bodies**

The physiological effects of expressing a recombinant enzyme in *E. coli* was infrequently considered in the literature that we have discussed until now. Heterologous protein expression in prokaryotic recombinant systems is not always a straightforward task following a clear and well-defined recipe. Biological processes are optimised for supporting the organism's survival; overexpression of foreign enzyme products causes system-level responses in the transcriptome, metabolome, and proteome of *E. coli* as observed by dynamic changes introduced to gene expression [123]. Recombinant protein expression and the overproduction of a heterologous protein was reported to potentially create a large burden on the cell, consequently leading to stress [4]. Although *E. coli* cells are agile, they can only adapt to stress conditions, such as increased protein synthesis, to a certain limit. In such instances, metabolic resources normally dedicated to cell propagation would then need to be committed to the synthesis of a non-endogenous protein product.

Additionally, the production of misfolded aggregates would lead to the accumulation of low-quality products that the cell would not be able to breakdown or fully refold back to their native state [123].

Within our analysis, the discussion of changes at the metabolomic or the proteomic level in response to over-expression was very limited. Understandably, this was not the intent nor the purpose of the research narrative in the evaluated studies. Research appears to take advantage of inclusion bodies; these act as sources of relatively pure, stable, and large protein deposits that can be easily isolated for refolding, as a means to end to generate the enzyme of interest [124]. However, it does beg the question whether this experimental design can be improved with a holistic, systems perspective to optimize protein synthesis, promote solubility and in turn reduce the need for additional downstream processing steps. Within this field of research, we find a growing number of alternative outlooks, led by omics-based techniques and bioinformatics, which could potentially be used to modify or evaluate experimental parameters to improve the expression of DtE enzymes. We will discuss these approaches in addition to the use of computational tools within our selected review papers in the subsequent sections.

### **Role of bioinformatics and modelling within our literature survey**

Among the selected literature, the use of bioinformatics and modelling tools for structural and functional characterization was discussed in 19% of cases. Primarily, computational methods we found to be used in the context of identifying uncharacterized gene clusters encoding enzymes from environmental samples such as a cold-active esterase from *Rhodococcus* sp. AW25M09 found in arctic ocean water [125] or a halotolerant lipase from *Marinobacter lipolyticus* found in the hypersaline regions of southern Spain [42]. Adaptation traits provide molecular biologists and industrial manufacturers additional tools capable of withstanding adverse conditions of temperature, pH, chemicals (i.e. co-solvents), or salinity for example [126]. In other instances, genome mining of sequenced organisms or related species highlighted uncharacterized variants of well-known enzymes such as sarcosine oxidase [127], or  $\beta$ -agarase [108].

Homologous sequence alignment tools, such as BLAST [128] were used to assess and compare the similarity of novel enzymes to known sequences found within the GenBank online database [129]. The conserved regions for a given enzyme were compared across species using alignment programs such as EMBL-EBI's Clustal Omega [130–133] or other algorithms such as Needleman–Wunsch Global Align Nucleotide Sequence [68, 134]; these conserved domains, such as active site residues, assisted

researchers to interpret and predict an enzyme's function [135]. Beyond these methods, we observed the use of SWISS-MODEL [127, 135–137] and PyMOL [42, 138] for comparative 3D modelling of the evolutionary relationship between target proteins and the prediction of substrate-enzyme docking interactions. It is expected that this methodology for predictive structural modelling of new DtE enzymes will be highly influenced by innovations incorporating novel neural network architectures such as AlphaFold [139]. In a limited number of cases (5%) SignalP prediction server [140] was used to detect putative signal sequence motifs in the gene sequence of the novel enzyme and extrapolate the native subcellular localization the enzyme when expressed [108, 132, 141, 142].

#### **Peptide sequence-driven computational methods to predict solubility of DtE enzymes**

Sequence-based analyses can highlight the folding patterns of proteins such as how surface residue patches can interact with their surrounding environment. Protein engineering research has observed that larger patches of positively charged residues and hydrophobic surface residues impact aggregation within proteins. A mutation in a single residue can tremendously impact the charge distribution of recombinant proteins and in turn increase solubility [143]. Furthermore, restricting the exposure of hydrophobic patches on a protein's surface has been shown to increase the likelihood of producing a soluble protein in aqueous environments, depending on the ratio of hydrophobic to polar amino acids on its surface [144].

Sequence-based modelling tools, derived from the statistical solubility model of Wilkinson and Harrison [145, 146], such as PROSO II and SOLpro can help predict the solubility of a protein by its amino acid composition. These tools take sequence-specific factors including folding propensity, residue charge, cysteine fractions, and hydrophilicity, into consideration in their algorithms [24, 147]; often, these programs can be a preliminary resource to predict the folding patterns of a protein of interest. In the past, similar sequence-based prediction methods were used to evaluate plasmid design by ranking the choice of expression constructs with fusion carrier proteins such as NusA, GrpE, and thioredoxin bound to insoluble protein human interleukin-3 (hIL-3) in *E. coli* [148]. It was found that this method was successful in predicting the effectiveness of a given tag in promoting solubility when fused to hIL-3. In another study, Chan et al. applied a model-based approach to assess the cloning regions of six vector designs for the effect of varying the location of solubility fusion tags (Trx, MBP, NusA) and affinity tags such as the His-tag on the solubility of their product; their methodology presented a

model to evaluate the design of plasmids for recombinant expression—validated by machine-learning based prediction tools [149]. Often, despite their potential, such modelling-based tools are still criticized for disregarding sequence-independent features such as growth temperatures, media conditions, inducer concentration, etc. that also play a role in the formation of inclusion bodies [150]. Furthermore, there is a need to validate such models through experimental methods. The sequence-based protein design algorithm—PROSS has already been validated by community-motivated efforts against a range of DtE proteins; it was found that 9 out of 14 target proteins showed improvement in heterologous expression under the experimental conditions designed by the prediction tool [151].

However, recent efforts on the recombinant expression of DtE enzymes in *E. coli* did not indicate bioinformatics were involved with experimental or amino acid sequence evaluation—despite the open-access to such tools. For example, our analysis did not observe a systematic consideration when selecting fusion partners in the design of an experiment, but the process was rather ad hoc, with decisions likely being made based on prior experience. Contrary to existing practices, computational sequence-based modelling tools may be useful to predict how a protein may be expressed based on the design of an expression vector in addition to guiding protein engineering. Design modifications can be made based on these predictive models on the road to promoting the solubility of a DtE enzyme. The use of these advanced technologies can expand our capabilities to systematically investigate aspects of protein biology and streamline our decisions for future experimentation.

#### **Codon bias and peptide sequence as modulators of correct folding**

Recombinant expression was reported as arguably one of the most metabolically taxing activities that an organism could undergo [4]. It requires an abundance of resources in the form of energy and raw materials, and therefore there is a limit to the extent of resources each organism could allocate to such a task. When the resource demand surpassed an organism's capacity, a stress response was observed, accompanied by a decrease in biomass production and growth rate due to the rewiring of metabolic fluxes in the cell [152]. Beyond energy and metabolite shortages, this stress response could also manifest itself in the form of cellular component shortages through changes in global gene expression [153].

Beyond folding patterns, the amino acid sequences of a protein can drastically impact the metabolic stress that *E. coli* may undergo during overexpression of exogenous proteins. The change of even a single amino acid residue

was reported to impact the metabolic burden of *E. coli* during recombinant expression; these minor changes were shown to negatively impact cellular respiration activity and heterologous protein production levels [154]. Studies also revealed that silent exchanges in specific synonymous codons could impact growth, protein production levels, and respiration activity—demonstrating the growth sensitivity of *E. coli* to amino acid sequences [155]. Understanding codon biases and optimizing peptide sequences in accordance with the genetic makeup of the expression host was reported to be elemental in achieving a high-performing expression system [156].

Codon optimization was considered in only 16% of the work addressing issues around improving the recombinant protein expression performance of DTE enzymes in *E. coli*. It is possible that a conscious choice has been made not to codon optimize, as an expression strategy. The placement of rare codons with an mRNA region can promote stability in addition to ribosomal stalling allowing extra time for folding of problematic peptide regions [157]. These studies purchased synthetic genes from commercial manufacturers including GenScript USA Inc. [132, 158, 159], Invitrogen [96, 160], Sloning BioTechnology GmbH, and Synbio Technologies [47] that carry out codon optimisation on their products as a default service. The Graphical Codon Usage Analyser [41] and the Genescript Rare Codon Analysis Tool [72] were used for in-house codon analysis [6, 40]. This, however, does not rule out the possibility that codon optimization was carried out more extensively, but was not explicitly acknowledged in each respective publication. This ambiguity obscures the evidence about whether inclusion bodies were formed despite codon optimization or not and may limit the reproducibility of these experiments in the future. In select cases, DTE enzymes were expressed in commercial *E. coli* strains such as Rosetta, which were specifically recommended for alleviating codon bias in *E. coli* [27, 77–79] in addition to CodonPlus [67, 82, 161]; this may be a potential initial strategy to express a non-codon optimized gene.

#### **Cellular quality control mechanisms and the role of molecular chaperones**

Molecular chaperones play an essential role in facilitating the recovery of misfolded protein aggregates. This *in vivo* quality control naturally exists within *E. coli* as its natural metabolism relies on cellular proteins that depend on appropriate folding patterns for proper function [162]. Recent research has found that chaperone systems such as GroEL/ES interact with a specific, smaller subset of the total proteome; this suggested that individual proteins in *E. coli*'s proteome were predetermined to be under the direct quality control management of a specific

chaperone system such as GroEL/ES, Trigger factor (TF), or DnaKJE rather than this process being a random event [162, 163]. The challenge has been in determining which proteins would be assigned as substrates for specific molecular chaperones, and what attributes determine this distinction. Microarray studies showed that inclusion body formation led to the upregulation of genes associated with protein refolding and the heat shock genes, in addition to those associated with proteolysis [123]. Furthermore, molecular chaperones were reported to directly interact with aggregated recombinant protein products [164, 165].

Understanding the underlying factors that determine protein-chaperone interactions would be useful in ensuring that the correct chaperone would be favourably upregulated during protein expression. Chaperone-substrate interaction models were explored using metabolic network analysis techniques to understand the distribution of chaperone substrate enzymes in the metabolic network of *E. coli*; Takemoto et al. observed that metabolic enzymes that act as chaperone substrates became extensively distributed in the metabolic network as the chaperone requirements increased [163]. Although only limited amount of work has been reported in this field, detailed bioinformatics and metabolic network analyses could improve our understanding of the interaction patterns of molecular chaperones with recombinant proteins.

Analysis of sequence homology may prove useful to gain detailed insight into chaperone-substrate interaction patterns. Raineri et al. found closely related proteins from the *E. coli* and *Salmonella typhimurium* proteome, which were likely to show similar behavior and interact with the same or related chaperones in the GroEL system [166]. Therefore, sequence homology affects a recombinant enzyme's interactions with molecular chaperones and has an indirect effect on the amount of product recovered from a misfolded state, and consequently on product titre. Mutations or changes introduced to the amino acid sequence of the peptide to be folded was reported to hinder the correct operation of the chaperone-mediated folding pathway [165]. This would be a risk to consider even in the case of beneficial mutations such as those introduced by site-directed mutagenesis to improve the catalytic activity of enzymes [158, 167–169].

Understanding these chaperone-substrate interaction patterns could be useful to selectively target and upregulate specific chaperone genes compatible with the DTE enzymes of interest to assist in folding. Co-expression of specific chaperones such as DnaK, DnaJ, GrpE, GroEL, GroES, or tig using commercial molecular chaperone plasmid sets from Takara Bio was observed in 12% of cases surveyed. This strategy proved useful

in every case, with a varying degree of success depending on the study. However, this method is not as simple as expressing all chaperones at once. Soluble expression of *Psychrobacter* sp. lipase (Lip-498) 15 °C using pColdI plasmid was hindered by the individual co-expression of tig (pTf16) and GroEL/ES (pGro7) with the enzyme comprising 0.9% of total soluble proteins [170]. Lip-498 comprised 11.8% soluble protein when tig and GroEL/ES were co-expressed simultaneously. This value increased to 19.8% of total soluble protein when DnaK/DnaJ/GrpE and GroEL/ES (pG-KJE8) were co-expressed simultaneously. pGro7 and pG-KJE8 had the highest frequency of use among all commercial chaperone sets. Alternatively, a study by Peng et al. [103] showcased that native chaperones (Hsp60 and small heat-shock protein) of *Pyrococcus furiosus* can improve the soluble expression of its  $\alpha$ -amylase in *E. coli* BL21 (DE3). It is unclear whether the co-expression of chaperones will always provide benefit, however, there is scope to investigate this hypothesis further.

#### 'Omics'-based investigation to improve experimental design and host genetic background

In the past, recombinant protein expression and its associated stress responses have been investigated at the 'omic'-level in *E. coli* with the aim of improving heterologous expression performance. *E. coli* cells have demonstrated transcriptional changes at the global level in response inclusion bodies within the cytoplasm [123]. Genes taking an active role in protein folding (i.e. molecular chaperone genes), protein synthesis (i.e. aminoacyl-tRNA synthetases and ribosomal genes), and genes responsible for energy metabolism (e.g. ATP synthase) were observed to have a dynamic upregulation in response to the formation of inclusion bodies of recombinant protein fusions tagged with green fluorescent protein (GFP) [123]. Sharma et al. [171] provided a comparative analysis of how metabolic networks in *E. coli* BL21 (DE3) were reorganized in response to the physical state of the end protein product being soluble or being confined to inclusion bodies. Their study employed fed-batch cultures, mimicking industrial conditions, to over-express rhIFN-b, xylanase and GFP; the transcription of amino acid biosynthesis and uptake genes was reported to be upregulated during inclusion body formation whereas the expression of these genes was downregulated during soluble expression, indicating that the solubility of the recombinantly expressed protein had a global impact on the transcription of the metabolic genes in *E. coli* [171].

The endometabolome of a cell is often thought to provide a physiological snapshot of a cell at a specific point in time. Chae et al. used two-dimensional NMR

spectroscopy to evaluate the effect of stressors on the endometabolome of *E. coli*. They assessed the effects of elevated NaCl concentration as a stressor for *E. coli* expressing recombinant proteins; at high NaCl concentrations, the cells accumulated maltose and 2-hydroxy-3-methylbutanoic acid, which, in turn, promoted the solubility of two of the eleven aggregation prone proteins that were investigated in the study [172]. The names of these proteins were not explicitly mentioned in the text. Recombinant protein expression is a source of cellular stress on its own, therefore these fingerprinting studies can assist with pinpointing differences in the metabolite profiles of expression systems based on changes in the experimental conditions that the cells are exposed to during recombinant expression. The information gained at the metabolomic level could assist designing an experiment to redirect metabolic flux towards the intracellular accumulation of specific metabolites to overcome or alleviate inclusion body formation. Furthermore, inferences from this type of analysis can guide the choice of media. It was observed that factors including maintaining a pH 6 medium [45], the addition of betaine as an osmolyte [106], and the addition L-arginine [173] could improve the solubility of the final product in three studies, however ad hoc methods led to this discovery.

Supporting the design of expression studies with 'omics'-based analyses could prove to be a useful strategy to improve solubility in addition to improving cellular biomass. The transcriptomic, metabolomic, and proteomic profiles of the *E. coli* expression hosts were under consideration in only one study within the scope of this review. This study discussed the use of metabolic engineering for the active production of xanthine dehydrogenase; their work demonstrated that the combinatorial overexpression of three global regulator and chaperone/helper proteins could improve the specific activity and solubility of their enzyme by up to 129% [174]. We proposed that 'omic'-level information in combination with sequence-based modelling, codon optimization, or molecular chaperone studies could help us better understand *E. coli* as a production organism. Heterologous overexpression of proteins by *E. coli* can be considered to mimic the operation of a cell factory; 'omics'-based technologies provide a level of process management over the operation by identifying the underlying bottlenecks in the manufacturing process in order to improve efficiency.

#### Outlook: need for a systematic roadmap to address the demands of an expanding field

Our review of literature since 2010 showed that although the field of recombinant protein expression is associated with a plethora of knowledge, a systematic roadmap to help guide researchers to express problematic

enzymes does not yet exist. We observe a variety of disparate practices and approaches adopted in the interest of promoting solubility, and the process is often led by ad hoc decisions. There is no standardized guideline for how enzyme expression is approached. Through experience researchers choose to adopt at least one method to preemptively reduce the possibility of their expression system to form inclusion bodies. Strategies could include inducing protein synthesis at low temperatures to reduce the rate of protein synthesis, and consequently promote correct folding, and to provide sufficient time for intracellular molecular chaperones to act [94, 156, 160]. This was a successful strategy in 14% of examples, with temperatures ranging from 10 to 25 °C as opposed to 37 °C, however even this strategy does not always lead to success solubility [175]. Reducing the inducer concentration, using plasmids with low copy numbers, or alternative promoters [176] are additional strategies employed.

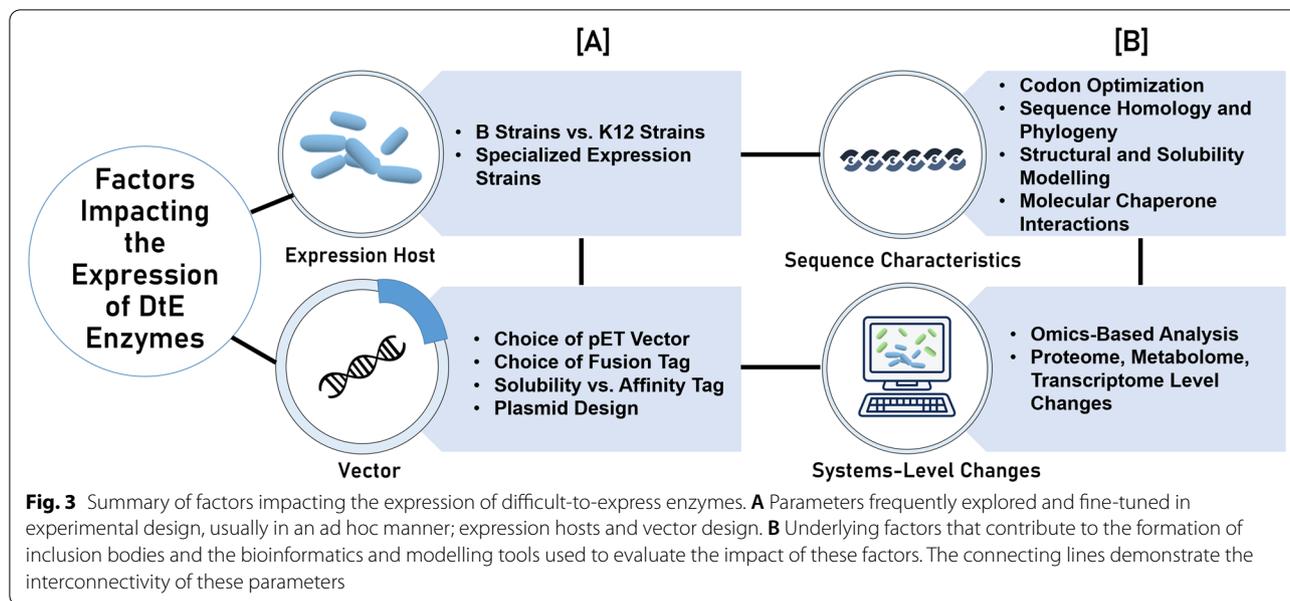
The bulk of literature surveyed in this review focused on functional and structural characterization of enzymes. It appears that if the initial design led to soluble product, the follow-up experiments were conducted as planned. However, in the event that the heterologous protein product was not soluble, a trial-and-error approach was employed to achieve the correct combination of parameter settings to promote solubility. This approach was observed to be successful on individual cases but does not provide any guarantee of success. Each study employed quality control parameters in the form of catalytic activity assays to assess the final product; each quality control measure varied based on the specific enzyme evaluated. Expression system design may be dictated by previous experience or limited by the availability of the tools and materials in the individual lab in which the experiments were carried out. This may explain the limited use of specialized expression strains or alternative plasmid backbones, for example, that would provide an additional cost without a guarantee of success.

Our analysis has found no consensus in reporting basic aspects of the experimental protocol as indicated by these studies. The details on the gene of interest (i.e. its native host, amino acid sequence), how it was modified prior to expression (i.e. codon optimization carried out or not), and in certain cases the rationale behind how the vector was designed, or the choice of host, can all play an important role in shaping future work related to that expression system, as well as providing guidance for future studies. A lack of these details can lead to low reproducibility for research in this field. We believe that this necessitates the development of a minimum information standard scheme to systematize work in this field as a community effort, similar to existing efforts such as MIAPE or MIPFE for proteomics studies to standardize

the reporting practices of experimental metadata relevant to structural and/or functional quality attributes of recombinant protein expression experiments [177, 178]. This information is required to facilitate reproducibility to build upon. A systematic collection of this standard metadata in database repositories in standalone format or accompanying any relevant experimental data is required. This metadata can also serve as foundation for modelling and bioinformatics in the field.

Amalgamated metadata detailing attempts to improve recombinant product solubility can potentially lead to the rapid discovery of broadly applicable rules for soluble enzyme expression in *E. coli*. While this review itself does not attempt to derive such rules, nor is there sufficient information available to derive these rules at this time, it is important to highlight this relevance. One effort that could assist the development of such a roadmap for recombinant enzymes would be to investigate the literature in which solubility of the enzyme product was successfully achieved from the start. Certain strategies used successfully for non-enzymatic protein including the use of inteins [179], reduced genome *E. coli* strains [180], or chromosomal integration [181, 182], though not discussed in this review, can provide further insights to support the efforts discussed above; however, the transferability of such techniques between protein types needs to be further understood. Furthermore, this volume of aggregated metadata highlighting all successful, partially successful, or unsuccessful experimental expression conditions can assist us to develop a strategic, evidence-based workflow for soluble recombinant enzyme production.

The availability of a wide range of non-dominating options for the strains, vectors, and the design tools indicates a strong drive among researchers to have increased control over their experimental design to overcome the challenges that are associated with DTE enzymes. However, the current literature reveals a different landscape where these techniques were often underutilized or overlooked. This presents an opportunity to approach the challenge systematically. An understanding of the most frequently utilized tools—the expression strains, vectors, and the experimental conditions, can serve as a baseline for researchers to optimize their expression models from the start. In cases where this proves ineffective, the use of an integrated systems biology approach based on bioinformatics, modelling, and/or 'omics' technologies can highlight problematic pitfalls in the experimental design and provide additional information on the system of interest (Fig. 3). In other instances, these approaches can be adopted as a preliminary investigation before laboratory applications are made. The combination of all



these approaches will assist in determining successful experimental conditions to recombinantly express a challenging candidate enzyme—promoting solubility.

#### Abbreviations

DtE: Difficult to express; SUMO: Small Ubiquitin-like Modifier; Trx: Thioredoxin; MBP: Maltose binding protein; His-tag: Polyhistidine-Tag; GST: Glutathione S-transferase.

#### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12934-021-01698-w>.

**Additional file 1: Table S1.** Experimental breakdown of all included publications.

#### Acknowledgements

The authors thank Dr. Annette Alcasabas for her insightful discussions on heterologous expression of problematic enzymes, and gratefully acknowledge the funding from Johnson Matthey Cambridge.

#### Authors' contributions

SM gathered and interpreted review literature for systematic analysis and wrote the manuscript, DD and GC conceived the review, and revised the manuscript. All authors read and approved the final manuscript.

#### Availability of data and materials

All data generated or analysed during this study are included in this published article (and its additional information files).

#### Declarations

##### Ethics approval and consent to participate

Not applicable.

##### Consent for publication

Not applicable.

#### Competing interests

The authors declare no competing interests.

#### Author details

<sup>1</sup>Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge CB3 0AS, UK. <sup>2</sup>Department of Biochemical Engineering, University College London, London WC1E 6BT, UK.

Received: 19 August 2021 Accepted: 22 October 2021

Published online: 30 October 2021

#### References

- Abdelraheem EMM, Busch H, Hanefeld U, Tonin F. Biocatalysis explained: from pharmaceutical to bulk chemical production. *React Chem Eng.* 2019;4:1878–94.
- Chapman J, Ismail A, Dinu C. Industrial applications of enzymes: recent advances, techniques, and outlooks. *Catalysts.* 2018;8:238. <http://www.mdpi.com/2073-4344/8/6/238>.
- Singh A, Upadhyay V, Upadhyay AK, Singh SM, Panda AK. Protein recovery from inclusion bodies of *Escherichia coli* using mild solubilization process. *Microb Cell Fact.* 2015;14:41. <https://doi.org/10.1186/s12934-015-0222-8>.
- Carneiro S, Ferreira EC, Rocha I. Metabolic responses to recombinant bioprocesses in *Escherichia coli*. *J Biotechnol.* 2013;164:396–408.
- Wingfield PT. Overview of the purification of recombinant proteins. *Curr Protoc Protein Sci.* 2015;80:6.1.1-6.1.35.
- Kovalenko GA, Beklemishev AB, Perminova LV, Mamaev AL, Rudina NA, Moseenkov SI, et al. Immobilization of recombinant *E. coli* thermostable lipase by entrapment inside silica xerogel and nanocarbon-in-silica composites. *J Mol Catal B Enzym.* 2013;98:78–86. <https://doi.org/10.1016/j.molcatb.2013.09.022>.
- García-Fruitós E, Vázquez E, Díez-Gil C, Corchero JL, Seras-Franzoso J, Ratera I, et al. Bacterial inclusion bodies: making gold from waste. *Trends Biotechnol.* 2012;30:65–70.
- Mollania N, Khajeh K, Ranjbar B, Rashno F, Akbari N, Fathi-Roudsari M. An efficient in vitro refolding of recombinant bacterial laccase in *Escherichia coli*. *Enzyme Microb Technol.* 2013;52:325–30. <https://doi.org/10.1016/j.enzmictec.2013.03.006>.
- Jeon SJ, Park JH. Refolding, characterization, and dye decolorization ability of a highly thermostable laccase from *Geobacillus sp* JS12.

- Protein Expr Purif. 2020;173:105646. <https://doi.org/10.1016/j.pep.2020.105646>.
10. Ahlawat S, Singh D, Virdi JS, Sharma KK. Molecular modeling and MD-simulation studies: fast and reliable tool to study the role of low-redox bacterial laccases in the decolorization of various commercial dyes. *Environ Pollut.* 2019;253:1056–65. <https://doi.org/10.1016/j.envpol.2019.07.083>.
  11. Basheer S, Rashid N, Akram MS, Akhtar M. A highly stable laccase from *Bacillus subtilis* strain R5: gene cloning and characterization. *Biosci Biotechnol Biochem.* 2019;83:436–45. <https://doi.org/10.1080/09168451.2018.1530097>.
  12. Qi X, Sun Y, Xiong S. A single freeze-thawing cycle for highly efficient solubilization of inclusion body proteins and its refolding into bioactive form. *Microb Cell Fact.* 2015;14:24.
  13. Upadhyay AK, Singh A, Mukherjee KJ, Panda AK. Refolding and purification of recombinant L-asparaginase from inclusion bodies of *E. coli* into active tetrameric protein. *Front Microbiol.* 2014;5:1–10.
  14. Abuhammad A, Lack N, Schweichler J, Staunton D, Sim RB, Sim E. Improvement of the expression and purification of *Mycobacterium tuberculosis* arylamine N-acetyltransferase (TBNAT) a potential target for novel anti-tubercular agents. *Protein Expr Purif.* 2011;80:246–52. <https://doi.org/10.1016/j.pep.2011.06.021>.
  15. Freydel EJ, van der Wielen LAM, Eppink MHM, Ottens M. Techno-economic evaluation of an inclusion body solubilization and recombinant protein refolding process. *Biotechnol Prog.* 2011;27:1315–28.
  16. Huang CJ, Lin H, Yang X. Industrial production of recombinant therapeutics in *Escherichia coli* and its recent advancements. *J Ind Microbiol Biotechnol.* 2012;39:383–99.
  17. Chen R. Bacterial expression systems for recombinant protein production: *E. coli* and beyond. *Biotechnol Adv.* 2012;30:1102–7.
  18. Ferrer-Miralles N, Domingo-Espín J, Corchero J, Vázquez E, Villaverde A. Microbial factories for recombinant pharmaceuticals. *Microb Cell Fact.* 2009. <https://doi.org/10.1186/1475-2859-8-17>.
  19. Tripathi NK, Shrivastava A. Recent developments in bioprocessing of recombinant proteins: expression hosts and process development. *Front Bioeng Biotechnol.* 2019. <https://doi.org/10.3389/fbioe.2019.00420>.
  20. Xu X, Nagarajan H, Lewis NE, Pan S, Cai Z, Liu X, et al. The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. *Nat Biotechnol.* 2011;29:735–41.
  21. Rampelotto P. Extremophiles and extreme environments. *Life.* 2013;3:482–5.
  22. Haddaway NR, Collins AM, Coughlin D, Kirk S. The role of Google Scholar in evidence reviews and its applicability to grey literature searching. *PLoS ONE.* 2015;10:e0138237. <https://doi.org/10.1371/journal.pone.0138237>.
  23. Zhou R-B, Lu H-M, Liu J, Shi J-Y, Zhu J, Lu Q-Q, et al. A systematic analysis of the structures of heterologously expressed proteins and those from their native hosts in the RCSB PDB archive. *PLoS ONE.* 2016;11:e0161254.
  24. Magnan CN, Randall A, Baldi P. SOLpro: accurate sequence-based prediction of protein solubility. *Bioinformatics.* 2009;25:2200–7.
  25. Salis B, Spinetti G, Scaramuzza S, Bossi M, Sacconi Jotti G, Tonon G, et al. High-level expression of a recombinant active microbial transglutaminase in *Escherichia coli*. *BMC Biotechnol.* 2015;15:84.
  26. Fakruddin Md, Mohammad Mazumdar R, Bin Mannan KS, Chowdhury A, Hossain MN. Critical factors affecting the success of cloning, expression, and mass production of enzymes by recombinant *E. coli*. *ISRN Biotechnol.* 2013;2013:1–7.
  27. Garg R, Srivastava R, Brahma V, Verma L, Karthikeyan S, Sahni G. Biochemical and structural characterization of a novel halotolerant cellulase from soil metagenome. *Sci Rep.* 2016;6:39634.
  28. Sun H, Gao W, Wang H, Wei D. Expression, characterization of a novel nitrilase Ppl19 from *Pseudomonas psychrotolerans* with S-selectivity toward mandelonitrile present in active inclusion bodies. *Biotechnol Lett.* 2016;38:455–61.
  29. Wang X-C, Liu J, Zhao J, Ni X-M, Zheng P, Guo X, et al. Efficient production of trans-4-hydroxy-L-proline from glucose using a new trans-proline 4-hydroxylase in *Escherichia coli*. *J Biosci Bioeng.* 2018;126:470–7.
  30. Ninh PH, Honda K, Sakai T, Okano K, Ohtake H. Assembly and multiple gene expression of thermophilic enzymes in *Escherichia coli* for in vitro metabolic engineering. *Biotechnol Bioeng.* 2015;112:189–96.
  31. Elleuche S, Qoura FM, Lorenz U, Rehn T, Brück T, Antranikian G. Cloning, expression and characterization of the recombinant cold-active type-I pullulanase from *Shewanella arctica*. *J Mol Catal B Enzym.* 2015;116:70–7.
  32. Munawar N, Engel PC. Overexpression in a non-native halophilic host and biotechnological potential of NAD<sup>+</sup>-dependent glutamate dehydrogenase from *Halobacterium salinarum* strain NRC-36014. *Extremophiles.* 2012;16:463–76. <https://doi.org/10.1007/s00792-012-0446-z>.
  33. Rigoldi F, Donini S, Redaelli A, Parisini E, Gaudieri A. Review: Engineering of thermostable enzymes for industrial applications. *APL Bioeng.* 2018;2:011501.
  34. Ye LJ, Toh HH, Yang Y, Adams JP, Snajdrova R, Li Z. Engineering of amine dehydrogenase for asymmetric reductive amination of ketone by evolving *Rhodococcus phenylalanine* dehydrogenase. *ACS Catal.* 2015;5:1119–22.
  35. Wu H, Yu X, Chen L, Wu G. Cloning, overexpression and characterization of a thermostable pullulanase from *Thermus thermophilus* HB27. *Protein Expr Purif.* 2014;95:22–7. <https://doi.org/10.1016/j.pep.2013.11.010>.
  36. Tayyab M, Rashid N, Akhtar M. Isolation and identification of lipase producing thermophilic *Geobacillus* sp. SBS-45: cloning and characterization of the lipase. *J Biosci Bioeng.* 2011;11:272–8.
  37. Kim S, Park H, Choi J. Cloning and characterization of cold-adapted  $\alpha$ -amylase from antarctic *Arthrobacter agilis*. *Appl Biochem Biotechnol.* 2017;181:1048–59. <https://doi.org/10.1007/s12010-016-2267-5>.
  38. Cheng H, Luo Z, Lu M, Gao S, Wang S. The hyperthermophilic  $\alpha$ -amylase from *Thermococcus* sp. HJ21 does not require exogenous calcium for thermostability because of high-binding affinity to calcium. *J Microbiol.* 2017;55:379–87. <https://doi.org/10.1007/s12275-017-6416-5>.
  39. Chang J, Lee Y-S, Fang S-J, Park I-H, Choi Y-L. Recombinant expression and characterization of an organic-solvent-tolerant  $\alpha$ -amylase from *Exiguobacterium* sp. DAU5. *Appl Biochem Biotechnol.* 2013;169:1870–83. <https://doi.org/10.1007/s12010-013-0101-x>.
  40. Kuschel B, Claaßen W, Mu W, Jiang B, Stressler T, Fischer L. Reaction investigation of lactulose-producing cellobiose 2-epimerases under operational relevant conditions. *J Mol Catal B Enzym.* 2016;133:S80–7. <https://doi.org/10.1016/j.molcatb.2016.11.022>.
  41. Tong Y, Feng S, Xin Y, Yang H, Zhang L, Wang W, et al. Enhancement of soluble expression of codon-optimized *Thermomicrobium roseum* sarcosine oxidase in *Escherichia coli* via chaperone co-expression. *J Biotechnol.* 2016;218:75–84.
  42. Pérez D, Kovačić F, Wilhelm S, Jaeger K-E, García MT, Ventosa A, et al. Identification of amino acids involved in the hydrolytic activity of lipase LipBL from *Marinobacter lipolyticus*. *Microbiology.* 2012;158:2192–203. <https://doi.org/10.1099/mic.0.058792-0>.
  43. Uehara R, Ueda Y, You D, Koga Y, Kanaya S. Accelerated maturation of Tk-subtilisin by a Leu $\rightarrow$ Pro romutation at the C-terminus of the propeptide, which reduces the binding of the propeptide to Tk-subtilisin. *FEBS J.* 2013;280:994–1006. <https://doi.org/10.1111/febs.12091>.
  44. Wu J, Kim K-S, Lee J-H, Lee Y-C. Cloning, expression in *Escherichia coli*, and enzymatic properties of laccase from *Aeromonas hydrophila* WL-11. *J Environ Sci.* 2010;22:635–40. [https://doi.org/10.1016/S1001-0742\(09\)60156-X](https://doi.org/10.1016/S1001-0742(09)60156-X).
  45. Zhang J, Lu J, Su E. Soluble recombinant pyruvate oxidase production in *Escherichia coli* can be enhanced and inclusion bodies minimized by avoiding pH stress. *J Chem Technol Biotechnol.* 2019;94:2661–70. <https://doi.org/10.1002/jctb.6075>.
  46. Ni H, Guo P-C, Jiang W-L, Fan X-M, Luo X-Y, Li H-H. Expression of nat-kinase in *Escherichia coli* and renaturation of its inclusion body. *J Biotechnol.* 2016;231:65–71.
  47. Yao D, Fan J, Han R, Xiao J, Li Q, Xu G, et al. Enhancing soluble expression of sucrose phosphorylase in *Escherichia coli* by molecular chaperones. *Protein Expr Purif.* 2020;169:105571.
  48. Li Y, Zhou Z, Chen Z. High-level production of ChSase ABC I by co-expressing molecular chaperones in *Escherichia coli*. *Int J Biol Macromol.* 2018;119:779–84.
  49. Li S, Pang H, Lin K, Xu J, Zhao J, Fan L. Refolding, purification and characterization of an organic solvent-tolerant lipase from *Serratia marcescens*

- ECU1010. *J Mol Catal B Enzym*. 2011;71:171–6. <https://doi.org/10.1016/j.molcatb.2011.04.016>.
50. Mohammadi HS, Omidinia E. Process integration for the recovery and purification of recombinant *Pseudomonas fluorescens* proline dehydrogenase using aqueous two-phase systems. *J Chromatogr B*. 2013;929:11–7. <https://doi.org/10.1016/j.jchromb.2013.03.024>.
  51. Curiel JA, de las Rivas B, Mancheño JM, Muñoz R. The pURI family of expression vectors: a versatile set of ligation independent cloning plasmids for producing recombinant His-fusion proteins. *Protein Expr Purif*. 2011;76:44–53.
  52. Rosano GL, Ceccarelli EA. Recombinant protein expression in *Escherichia coli*: advances and challenges. *Front Microbiol*. 2014;5:1–17.
  53. Rosano GL, Morales ES, Ceccarelli EA. New tools for recombinant protein production in *Escherichia coli*: a 5-year update. *Protein Sci*. 2019;28:1412–22.
  54. Waegeman H, De Lausnay S, Beauprez J, Maertens J, De Mey M, Soetaert W. Increasing recombinant protein production in *Escherichia coli* K12 through metabolic engineering. *New Biotechnol*. 2013;30:255–61.
  55. Singh SP, Purohit MK, Aoyagi C, Kitaoka M, Hayashi K. Effect of growth temperature, induction, and molecular chaperones on the solubilization of over-expressed cellobiose phosphorylase from *Cellvibrio Gilvus* under in vivo conditions. *Biotechnol Bioprocess Eng*. 2010;15:273–6.
  56. Li X, Wang L, Bai L, Yao C, Zhang Y, Zhang R, et al. Cloning and characterization of a glucosyltransferase and a rhamnosyltransferase from *Streptomyces* sp. 139. *J Appl Microbiol*. 2010;108:1544–51. <https://doi.org/10.1111/j.1365-2672.2009.04550.x>.
  57. Hofer M, Bönsch K, Greiner-Stöfle T, Ballschmiter M. Characterization and engineering of a novel pyrroloquinoline quinone dependent glucose dehydrogenase from *Sorangium cellulosum* So ce56. *Mol Biotechnol*. 2011;47:253–61. <https://doi.org/10.1007/s12033-010-9339-5>.
  58. Sans C, García-Fruitós E, Ferraz RM, González-Montalbán N, Rinas U, López-Santín J, et al. Inclusion bodies of fuculose-1-phosphate aldolase as stable and reusable biocatalysts. *Biotechnol Prog*. 2012;28:421–7.
  59. Kumar S, Sharma R, Tewari R. Production of *N*-acetylglucosamine using recombinant chitinolytic enzymes. *Indian J Microbiol*. 2011;51:319–25. <https://doi.org/10.1007/s12088-011-0157-7>.
  60. Singh G, Arya S, Narang D, Jadeja D, Singh G, Gupta UD, et al. Characterization of an acid inducible lipase Rv3203 from *Mycobacterium tuberculosis* H37Rv. *Mol Biol Rep*. 2014;41:285–96. <https://doi.org/10.1007/s11033-013-2861-3>.
  61. Mestrom L, Marsden SR, Dieters M, Achterberg P, Stolk L, Bento I, et al. Artificial fusion of mCherry enhances trehalose transferase solubility and stability. *Appl Environ Microbiol*. 2019;85:1–15.
  62. Van Der Henst C, Charlier C, Deghelt M, Wouters J, Matroule JY, Letesson JJ, et al. Overproduced *Brucella abortus* PdhS-mCherry forms soluble aggregates in *Escherichia coli*, partially associating with mobile foci of lbpA-YFP. *BMC Microbiol*. 2010. <https://doi.org/10.1186/1471-2180-10-248>.
  63. de Almeida TP, van Schie MMCH, Ma A, Tieves F, Younes SHH, Fernández-Fueyo E, et al. Efficient aerobic oxidation of trans-2-Hexen-1-ol using the aryl alcohol oxidase from *Pleurotus eryngii*. *Adv Synth Catal*. 2019;361:2668–72. <https://doi.org/10.1002/adsc.201801312>.
  64. Lin S, Hanson RE, Cronan JE. Biotin synthesis begins by hijacking the fatty acid synthetic pathway. *Nat Chem Biol*. 2010;6:682–8. <http://www.nature.com/articles/nchembio420>.
  65. Vadala BS, Deshpande S, Apte-Deshpande A. Soluble expression of recombinant active cellulase in *E. coli* using *B. subtilis* (natto strain) cellulase gene. *J Genet Eng Biotechnol*. 2021;19:7.
  66. Su L, Wu S, Feng J, Wu J. High-efficiency expression of *Sulfolobus acidocaldarius* maltooligosyl trehalose trehalohydrolase in *Escherichia coli* through host strain and induction strategy optimization. *Bioprocess Biosyst Eng*. 2019;42:345–54. <https://doi.org/10.1007/s00449-018-2039-4>.
  67. Gröning JAD, Kaschabek SR, Schlömann M, Tischler D. A mechanistic study on SMOB-ADP1: an NADH:flavin oxidoreductase of the two-component styrene monooxygenase of *Acinetobacter baylyi* ADP1. *Arch Microbiol*. 2014;196:829–45. <https://doi.org/10.1007/s00203-014-1022-y>.
  68. Vikram S, Pandey J, Bhalla N, Pandey G, Ghosh A, Khan F, et al. Branching of the *p*-nitrophenol (PNP) degradation pathway in *Burkholderia* sp. strain SJ98: evidences from genetic characterization of PNP gene cluster. *AMB Express*. 2012;2:30. <https://doi.org/10.1186/2191-0855-2-30>.
  69. Kufner K, Lipps G. Construction of a chimeric thermoacidophilic beta-endoglucanase. *BMC Biochem*. 2013;14:11.
  70. Hartinger D, Heini S, Schwartz H, Grabherr R, Schatzmayr G, Haltrich D, et al. Enhancement of solubility in *Escherichia coli* and purification of an aminotransferase from *Sphingopyxis* sp. MTA144 for deamination of hydrolyzed fumonisins B1. *Microb Cell Fact*. 2010;9:62. <https://doi.org/10.1186/1475-2859-9-62>.
  71. Klermund L, Riederer A, Groher A, Castiglione K. High-level soluble expression of a bacterial *N*-acyl-d-glucosamine 2-epimerase in recombinant *Escherichia coli*. *Protein Expr Purif*. 2015;111:36–41.
  72. García-Fraga B, da Silva AF, López-Seijas J, Sieiro C. Optimized expression conditions for enhancing production of two recombinant chitinolytic enzymes from different prokaryote domains. *Bioprocess Biosyst Eng*. 2015;38:2477–86.
  73. Chan GF, Rashid NAA, Yusoff ARM. Expression, purification and characterization of flavin reductase from *Citrobacter freundii* A1. *Ann Microbiol*. 2013;63:343–51. <https://doi.org/10.1007/s13213-012-0480-1>.
  74. Schröder C, Blank S, Antranikian G. First glycoside hydrolase family 2 enzymes from *Thermus antranikianii* and *Thermus brockianus* with  $\beta$ -glucosidase activity. *Front Bioeng Biotechnol*. 2015;3:1–10. <https://doi.org/10.3389/fbioe.2015.00076/abstract>.
  75. Alexander AK, Biedermann D, Fink MJ, Mihovilovic MD, Mattes TE. Enantioselective oxidation by a cyclohexanone monooxygenase from the xenobiotic-degrading *Polaromonas* sp. strain JS666. *J Mol Catal B Enzym*. 2012;78:105–10. <https://doi.org/10.1016/j.molcatb.2012.03.002>.
  76. Hoffman BJ, Broadwater JA, Johnson P, Harper J, Fox BG, Kenealy WR. Lactose fed-batch overexpression of recombinant metalloproteins in *Escherichia coli* BL21(DE3): process control yielding high levels of metal-incorporated, soluble protein. *Protein Expr Purif*. 1995;6:646–54.
  77. Chacón-Verdú MD, Campillo-Brocal JC, Lucas-Elió P, Davidson VL, Sánchez-Amat A. Characterization of recombinant biosynthetic precursors of the cysteine tryptophylquinone cofactors of l-lysine- $\epsilon$ -lysine-oxidase and glycine oxidase from *Marinomonas mediterranea*. *Biochim Biophys Acta BBA Proteins Proteom*. 2015;1854:1123–31. <https://doi.org/10.1016/j.bbapap.2014.12.018>.
  78. Blaszczyk AJ, Silakov A, Zhang B, Maiocco SJ, Lanz ND, Kelly WL, et al. Spectroscopic and electrochemical characterization of the iron-sulfur and cobalamin cofactors of TsrM, an unusual radical S-adenosylmethionine methylase. *J Am Chem Soc*. 2016;138:3416–26. <https://doi.org/10.1021/jacs.5b12592>.
  79. Dang G, Cao J, Cui Y, Song N, Chen L, Pang H, et al. Characterization of Rv0888, a novel extracellular nuclease from *Mycobacterium tuberculosis*. *Sci Rep*. 2016;6:19033. <http://www.nature.com/articles/srep19033>.
  80. Guidi B, Planchestainer M, Contente ML, Laurenzi T, Eberini I, Gourlay LJ, et al. Strategic single point mutation yields a solvent- and salt-tolerant transaminase from *Virgibacillus* sp. in soluble form. *Sci Rep*. 2018;8:16441.
  81. Nguyen V, Hatahet F, Salo KEH, Enlund E, Zhang C, Ruddock LW. Pre-expression of a sulfhydryl oxidase significantly increases the yields of eukaryotic disulfide bond containing proteins expressed in the cytoplasm of *E. coli*. *Microb Cell Fact*. 2011;10:1.
  82. Rosano GL, Ceccarelli EA. 2 content affects the solubility of recombinant proteins in a codon bias-adjusted *Escherichia coli* strain. *Microb Cell Fact*. 2009;8:41.
  83. Yoneda K, Fukuda J, Sakuraba H, Ohshima T. First crystal structure of l-lysine 6-dehydrogenase as an NAD-dependent amine dehydrogenase. *J Biol Chem*. 2010;285:8444–53.
  84. Allen KD, Wang SC. Initial characterization of Fom3 from *Streptomyces wedmorensis*: the methyltransferase in fosfomicin biosynthesis. *Arch Biochem Biophys*. 2014;543:67–73. <https://doi.org/10.1016/j.abb.2013.12.004>.
  85. Schlegel S, Löfblom J, Lee C, Hjelm A, Klepsch M, Strous M, et al. Optimizing Membrane Protein Overexpression in the *Escherichia coli* strain Lemo21(DE3). *J Mol Biol*. 2012;423:648–59.
  86. Ni W, Liu H, Wang P, Wang L, Sun X, Wang H, et al. Evaluation of multiple fused partners on enhancing soluble level of prenyltransferase NovQ in *Escherichia coli*. *Bioprocess Biosyst Eng*. 2019;42:465–74. <https://doi.org/10.1007/s00449-018-2050-9>.

87. Annamalai T, Dani N, Cheng B, Tse-Dinh Y-C. Analysis of DNA relaxation and cleavage activities of recombinant *Mycobacterium tuberculosis* DNA topoisomerase I from a new expression and purification protocol. *BMC Biochem*. 2009;10:18.
88. Rizzi C, Frazzon J, Ely F, Weber PG, da Fonseca IO, Gallas M, et al. DAHP synthase from *Mycobacterium tuberculosis* H37Rv: cloning, expression, and purification of functional enzyme. *Protein Expr Purif*. 2005;40:23–30.
89. Studier FW, Moffatt BA. Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *J Mol Biol*. 1986;189:113–30.
90. Kodan A, Kuroda H, Sakai F. A stilbene synthase from Japanese red pine (*Pinus densiflora*): implications for phytoalexin accumulation and down-regulation of flavonoid biosynthesis. *Proc Natl Acad Sci*. 2002;99:3335–9.
91. Ratzka A, Vogel H, Kliebenstein DJ, Mitchell-Olds T, Kroymann J. Disarming the mustard oil bomb. *Proc Natl Acad Sci*. 2002;99:11223–8.
92. Chen J, Avci FY, Muñoz EM, McDowell LM, Chen M, Pedersen LC, et al. Enzymatic redesigning of biologically active heparan sulfate. *J Biol Chem*. 2005;280:42817–25.
93. Studier FW. Use of bacteriophage T7 lysozyme to improve an inducible T7 expression system. *J Mol Biol*. 1991;219:37–44.
94. Kim HT, Chung JH, Wang D, Lee J, Woo HC, Choi I-G, et al. Depolymerization of alginate into a monomeric sugar acid using Alg17C, an exo-oligoalginate lyase cloned from *Saccharophagus degradans* 2–40. *Appl Microbiol Biotechnol*. 2012;93:2233–9. <https://doi.org/10.1007/s00253-012-3882-x>.
95. Shavandi M, Sohelli M, Zareian S, Akbari N, Khajeh K. The gene cloning, overexpression, purification, and characterization of dibenzothioephene monooxygenase and desulfinate from *Gordonia alkanivorans* RIP190A. *J Pet Sci Technol*. 2013;3:57–64. [http://jpst.ripi.ir/?\\_action=articleInfo&article=306](http://jpst.ripi.ir/?_action=articleInfo&article=306).
96. Shah S, Sunder AV, Singh P, Wangikar P. Characterization and application of a robust glucose dehydrogenase from *Paenibacillus pini* for cofactor regeneration in biocatalysis. *Indian J Microbiol*. 2020;60:87–95. <https://doi.org/10.1007/s12088-019-00834-w>.
97. Bakholdina SI, Sidorin EV, Khomenko VA, Isaeva MP, Bystritskaya EP, et al. The effect of conditions of the expression of the recombinant outer membrane phospholipase A1 from *Yersinia pseudotuberculosis* on the structure and properties of inclusion bodies. *Russ J Bioorg Chem*. 2018;44:178–87.
98. Zhao M, Cai M, Wu F, Zhang Y, Xiong Z, Xu P. Recombinant expression, refolding, purification and characterization of *Pseudomonas aeruginosa* protease IV in *Escherichia coli*. *Protein Expr Purif*. 2016;126:69–76. <https://doi.org/10.1016/j.pep.2016.05.019>.
99. Korovashkina AS, Rymko AN, Kvach SV, Zinchenko AI. Enzymatic synthesis of c-di-GMP using inclusion bodies of *Thermotoga maritima* full-length diguanylate cyclase. *J Biotechnol*. 2012;164:276–80. <https://doi.org/10.1016/j.jbiotec.2012.12.006>.
100. Ge Y, Guo S, Liu T, Zhao C, Li D, Liu Y, et al. Optimizing a production strategy for a nonspecific nuclease from *Yersinia enterocolitica* subsp. palearctica in genetically engineered *Escherichia coli*. *FEMS Microbiol Lett*. 2020;366:1–7.
101. Yang L, Liu X, Zhou N, Tian Y. Characteristics of refold acid urease immobilized covalently by graphene oxide-chitosan composite beads. *J Biosci Bioeng*. 2019;127:16–22.
102. Hayashi K, Kojima C. pCold-GST vector: a novel cold-shock vector containing GST tag for soluble protein production. *Protein Expr Purif*. 2008;62:120–7.
103. Peng S, Chu Z, Lu J, Li D, Wang Y, Yang S, et al. Co-expression of chaperones from *P. furiosus* enhanced the soluble expression of the recombinant hyperthermophilic  $\alpha$ -amylase in *E. coli*. *Cell Stress Chaperones*. 2016;21:477–84.
104. Chen A, Li Y, Liu X, Long Q, Yang Y, Bai Z. Soluble expression of pullulanase from *Bacillus acidopullulyticus* in *Escherichia coli* by tightly controlling basal expression. *J Ind Microbiol Biotechnol*. 2014;41:1803–10. <https://doi.org/10.1007/s10295-014-1523-3>.
105. Saeed H, Ali H, Soudan H, Embaby A, El-Sharkawy A, Farag A, et al. Molecular cloning, structural modeling and production of recombinant *Aspergillus terreus* L. asparaginase in *Escherichia coli*. *Int J Biol Macromol*. 2018;106:1041–51. <https://doi.org/10.1016/j.jbiomac.2017.08.110>.
106. Shao M, Chen Y, Zhang X, Rao Z, Xu M, Yang T, et al. Enhanced intracellular soluble production of 3-ketosteroid- $\Delta$ 1-dehydrogenase from *Mycobacterium neoaurum* in *Escherichia coli* and its application in the androst-1,4-diene-3,17-dione production. *J Chem Technol Biotechnol*. 2017;92:350–7.
107. Effendi SSW, Tan SI, Ting WW, Ng IS. Genetic design of co-expressed *Mesorhizobium loti* carbonic anhydrase and chaperone GroELs to enhancing carbon dioxide sequestration. *Int J Biol Macromol*. 2021;167:326–34.
108. Dong Q, Ruan L, Shi H. A  $\beta$ -agarase with high pH stability from *Flammovirga* sp. SJP92. *Carbohydr Res*. 2016;432:1–8. <https://doi.org/10.1016/j.carres.2016.05.002>.
109. Hrmova M, Stone BA, Fincher GB. High-yield production, refolding and a molecular modelling of the catalytic module of (1,3)- $\beta$ -D-glucan (curdlan) synthase from *Agrobacterium* sp. *Glycoconj J*. 2010;27:461–76.
110. Yamamoto T, Ugai H, Nakayama-Imaoishi H, Tada A, Elahi M, Houchi H, et al. Characterization of a recombinant *Bacteroides fragilis* sialidase expressed in *Escherichia coli*. *Anaerobe*. 2018;50:69–75. <https://doi.org/10.1016/j.anaerobe.2018.02.003>.
111. Cai X, Ma J, Wei D, Lin J, Wei W. Functional expression of a novel alkaline-adapted lipase of *Bacillus amyloliquefaciens* from stinky tofu brine and development of immobilized enzyme for biodiesel production. *Anaerobe* Van Leeuwenhoek. 2014;106:1049–60. <https://doi.org/10.1007/s10482-014-0274-5>.
112. Li L, Wang P, Tang Y. C-glycosylation of anhydrotetracycline scaffold with SsfS6 from the SF2575 biosynthetic pathway. *J Antibiot*. 2014;67:65–70. <http://www.nature.com/articles/ja201388>.
113. Su E, Xu J, You P. Functional expression of *Serratia marcescens* H30 lipase in *Escherichia coli* for efficient kinetic resolution of racemic alcohols in organic solvents. *J Mol Catal B Enzym*. 2014;106:11–6. <https://doi.org/10.1016/j.molcatb.2014.04.012>.
114. Chen K, Wu S, Zhu L, Zhang C, Xiang W, Deng Z, et al. Substitution of a single amino acid reverses the regioselectivity of the Baeyer–Villiger monooxygenase PntE in the biosynthesis of the antibiotic pentalenolactone. *Biochemistry*. 2016;55:6696–704. <https://doi.org/10.1021/acs.biochem.6b01040>.
115. Guo F-M, Wu J-P, Yang L-R, Xu G. Soluble and functional expression of a recombinant enantioselective amidase from *Klebsiella oxytoca* KCTC 1686 in *Escherichia coli* and its biochemical characterization. *Process Biochem*. 2015;50:1264–71.
116. Yeo KJ, Park J-W, Kim E-H, Jeon YH, Hwang KY, Cheong H-K. Characterization of the sensor domain of QseE histidine kinase from *Escherichia coli*. *Protein Expr Purif*. 2016;126:122–6.
117. Toda H, Imae R, Komio T, Itoh N. Expression and characterization of styrene monooxygenases of *Rhodococcus* sp. ST-5 and ST-10 for synthesizing enantiopure (S)-epoxides. *Appl Microbiol Biotechnol*. 2012;96:407–18. <https://doi.org/10.1007/s00253-011-3849-3>.
118. Acero EH, Ribitsch D, Rodriguez RD, Dellacher A, Zitzenbacher S, Marold A, et al. Two-step enzymatic functionalisation of polyamide with phenolics. *J Mol Catal B Enzym*. 2012;79:54–60.
119. Ogino H, Inoue S, Yasuda M, Doukyu N. Hyper-activation of foldase-dependent lipase with lipase-specific foldase. *J Biotechnol*. 2013;166:20–4.
120. Jia D, Yang Y, Peng Z, Zhang D, Li J, Liu L, et al. High efficiency preparation and characterization of intact poly(vinyl alcohol) dehydrogenase from *Sphingopyxis* sp. 113P3 in *Escherichia coli* by inclusion bodies renaturation. *Appl Biochem Biotechnol*. 2014;172:2540–51. <https://doi.org/10.1007/s12010-013-0703-3>.
121. Costa S, Almeida A, Castro A, Domingues L. Fusion tags for protein solubility, purification and immunogenicity in *Escherichia coli*: the novel Fh8 system. *Front Microbiol*. 2014;5:1–20.
122. Kosobokova EN, Skrypnik KA, Kosorukov VS. Overview of fusion tags for recombinant proteins. *Biochem Mosc*. 2016;81:187–200.
123. Baig F, Fernando LP, Salazar MA, Powell RR, Bruce TF, Harcum SW. Dynamic transcriptional response of *Escherichia coli* to inclusion body formation. *Biotechnol Bioeng*. 2014;111:980–99.
124. Singh SM, Panda AK. Solubilization and refolding of bacterial inclusion body proteins. *J Biosci Bioeng*. 2005;99:303–10.
125. De Santi C, Tedesco P, Ambrosino L, Altermark B, Willassen N-P, de Pascale D. A new alkaliphilic cold-active esterase from the psychrophilic marine bacterium *Rhodococcus* sp.: functional and structural

- studies and biotechnological potential. *Appl Biochem Biotechnol*. 2014;172:3054–68.
126. Karan R, Capes MD, DasSarma S. Function and biotechnology of extremophilic enzymes in low water activity. *Aquat Biosyst*. 2012;8:4.
  127. Xin Y, Zheng M, Wang Q, Lu L, Zhang L, Tong Y, et al. Structural and catalytic alteration of sarcosine oxidase through reconstruction with coenzyme-like ligands. *J Mol Catal B Enzym*. 2016;133:S250–8. <https://doi.org/10.1016/j.molcatb.2017.01.011>.
  128. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215:403–10.
  129. Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. *Nucleic Acids Res*. 2016;44:D67–72.
  130. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res*. 2019;47:W636–41.
  131. Rajnish KN, Asraf SAKS, Manju N, Gunasekaran P. Functional characterization of a putative  $\beta$ -lactamase gene in the genome of *Zymomonas mobilis*. *Biotech Lett*. 2011;33:2425–30. <https://doi.org/10.1007/s10529-011-0704-7>.
  132. Kim H, Yeon YJ, Choi YR, Song W, Pack SP, Choi YS. A cold-adapted tyrosinase with an abnormally high monophenolase/diphenolase activity ratio originating from the marine archaeon *Candidatus Nitrosopumilus koreensis*. *Biotech Lett*. 2016;38:1535–42. <https://doi.org/10.1007/s10529-016-2125-0>.
  133. Wang L, Li S, Yu W, Gongs Q. Cloning, overexpression and characterization of a new oligoalginate lyase from a marine bacterium *Shewanella* sp. *Biotechnol Lett*. 2015;37:665–71.
  134. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*. 1970;48:443–53.
  135. Zhang G, An Y, Parvez A, Zabed HM, Yun J, Qi X. Exploring a highly  $D$ -galactose specific  $L$ -arabinose isomerase from bifidobacterium adollescens for  $D$ -tagatose production. *Front Bioeng Biotechnol*. 2020;8:1–10. <https://doi.org/10.3389/fbioe.2020.00377/full>.
  136. Schwede T. SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res*. 2003;31:3381–5.
  137. Luo Y, Zhao Q, Liu Q, Feng Y. An artificial biosynthetic pathway for 2-amino-1,3-propanediol production using metabolically engineered *Escherichia coli*. *ACS Synth Biol*. 2019;8:548–56. <https://doi.org/10.1021/acssynbio.8b00466>.
  138. DeLano WL. Pymol: an open-source molecular graphics tool. CCP4 Newsletter on protein crystallography. 2002.
  139. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596:583–9.
  140. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat Biotechnol*. 2019;37:420–3.
  141. Salwan R, Sharma V, Pal M, Kasana RC, Yadav SK, Gulati A. Heterologous expression and structure-function relationship of low-temperature and alkaline active protease from *Acinetobacter* sp. IHB B 5011(MN12). *Int J Biol Macromol*. 2018;107:567–74.
  142. Emanuelsson O, Brunak S, von Heijne G, Nielsen H. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc*. 2007;2:953–71.
  143. Carballo-Amador MA, McKenzie EA, Dickson AJ, Warwicker J. Surface patches on recombinant erythropoietin predict protein solubility: engineering proteins to minimise aggregation. *BMC Biotechnol*. 2019;19:26.
  144. Jacak R, Leaver-Fay A, Kuhlman B. Computational protein design with explicit consideration of surface hydrophobic patches. *Proteins Struct Funct Bioinform*. 2012;80:825–38.
  145. Wilkins DL, Harrison RG. Predicting the solubility of recombinant proteins in *Escherichia coli*. *Bio/Technology*. 1991;9:443–8.
  146. Diaz AA, Tomba E, Lennarson R, Richard R, Bagajewicz MJ, Harrison RG. Prediction of protein solubility in *Escherichia coli* using logistic regression. *Biotechnol Bioeng*. 2010;105:374–83.
  147. Smailowski P, Doose G, Torkler P, Kaufmann S, Frishman D. PROSO II—a new method for protein solubility prediction. *FEBS J*. 2012;279:2192–200.
  148. Davis GD, Elisee C, Mewham DM, Harrison RG. New fusion protein systems designed to give soluble expression in *Escherichia coli*. *Biotechnol Bioeng*. 1999;65:382–8.
  149. Chan W-C, Liang P-H, Shih Y-P, Yang U-C, Lin W, Hsu C-N. Learning to predict expression efficacy of vectors in recombinant protein production. *BMC Bioinform*. 2010;11:S21.
  150. Chang CCH, Song J, Tey BT, Ramanan RN. Bioinformatics approaches for improved recombinant protein production in *Escherichia coli*: protein solubility prediction. *Brief Bioinform*. 2014;15:953–62.
  151. Peleg Y, Vincentelli R, Collins BM, Chen K-E, Livingstone EK, Weeraratunga S, et al. Community-wide experimental evaluation of the PROSS stability-design method. *J Mol Biol*. 2021;433:166964.
  152. Heyland J, Blank LM, Schmid A. Quantification of metabolic limitations during recombinant protein production in *Escherichia coli*. *J Biotechnol*. 2011;155:178–84.
  153. Durfee T, Hansen A-M, Zhi H, Blattner FR, Jin DJ. Transcription profiling of the stringent response in *Escherichia coli*. *J Bacteriol*. 2008;190:1084–96.
  154. Rahman N, Fulton A, Ihling N, Magni M, Jaeger K, Büchs J. Exchange of single amino acids at different positions of a recombinant protein affects metabolic burden in *Escherichia coli*. *Microb Cell Fact*. 2015;14:10.
  155. Rahman N, Schlupp CD, Mitsunaga H, Fulton A, Aryani T, Esch L, et al. A particular silent codon exchange in a recombinant gene greatly influences host cell metabolic activity. *Microb Cell Fact*. 2015;14:156.
  156. Francis DM, Page R. Strategies to optimize protein expression in *E. coli*. *Curr Protoc Protein Sci*. 2010;61:1–29.
  157. Parret AH, Besir H, Meijers R. Critical reflections on synthetic gene design for recombinant protein expression. *Curr Opin Struct Biol*. 2016;38:155–62.
  158. Zhang Y, Wang H, Wang X, Hu B, Zhang C, Jin W, et al. Identification of the key amino acid sites of the carbendazim hydrolase (MheI) from a novel carbendazim-degrading strain *Mycobacterium* sp. SD-4. *J Hazard Mater*. 2017;331:55–62.
  159. Jaroensuk J, Intasian P, Kiattisewee C, Munkajohnpon P, Chunthaboon P, Buttranan S, et al. Addition of formate dehydrogenase increases the production of renewable alkane from an engineered metabolic pathway. *J Biol Chem*. 2019;294:11536–48. <https://doi.org/10.1074/jbc.RA119.008246>.
  160. Kim KH, Jia X, Jia B, Jeon CO. Identification and characterization of  $L$ -malate dehydrogenases and the  $L$ -lactate-biosynthetic pathway in *Leuconostoc mesenteroides* ATCC 8293. *J Agric Food Chem*. 2018;66:8086–93. <https://doi.org/10.1021/acs.jafc.8b02649>.
  161. Kim S-Y, Seo D-H, Kim S-H, Hong Y-S, Lee J-H, Kim Y-J, et al. Comparative study on four amylosucrases from *Bifidobacterium* species. *Int J Biol Macromol*. 2020;155:535–42.
  162. Azia A, Unger R, Horovitz A. What distinguishes GroEL substrates from other *Escherichia coli* proteins? *FEBS J*. 2012;279:543–50.
  163. Takemoto K, Niwa T, Taguchi H. Difference in the distribution pattern of substrate enzymes in the metabolic network of *Escherichia coli*, according to chaperonin requirement. *BMC Syst Biol*. 2011;5:98.
  164. Carrió MM, Villaverde A. Localization of chaperones DnaK and GroEL in bacterial inclusion bodies. *J Bacteriol*. 2005;187:3599–601.
  165. Carrió MM, Villaverde A. Role of molecular chaperones in inclusion body formation. *FEBS Lett*. 2003;537:215–21.
  166. Raineri E, Ribeca P, Serrano L, Maier T. A more precise characterization of chaperonin substrates. *Bioinformatics*. 2010;26:1685–9.
  167. Ainciart N, Zylberman V, Craig PO, Nygaard D, Bonomi HR, Cauerhff AA, et al. Sensing the dissociation of a polymeric enzyme by means of an engineered intrinsic probe. *Proteins Struct Funct Bioinform*. 2011;79:1079–88.
  168. Goh PH, Ilias RM, Goh KM. Rational mutagenesis of cyclodextrin glucanotransferase at the calcium binding regions for enhancement of thermostability. *Int J Mol Sci*. 2012;13:5307–23.
  169. Kozai M, Sasamori E, Fujihara M, Yamashita T, Taira H, Harasawa R. Growth inhibition of human melanoma cells by a recombinant arginine deiminase expressed in *Escherichia coli*. *J Vet Med Sci*. 2009;71:1343–7.
  170. Shuo-Shuo C, Xue-Zheng L, Ji-Hong S. Effects of co-expression of molecular chaperones on heterologous soluble expression of the cold-active lipase Lip-948. *Protein Expr Purif*. 2011;77:166–72. <https://doi.org/10.1016/j.pep.2011.01.009>.

171. Sharma AK, Mahalik S, Ghosh C, Singh AB, Mukherjee KJ. Comparative transcriptomic profile analysis of fed-batch cultures expressing different recombinant proteins in *Escherichia coli*. *AMB Express*. 2011;1:33.
172. Chae YK, Kim SH, Um Y. Relationship between protein expression pattern and host metabolome perturbation as monitored by two-dimensional NMR spectroscopy. *Bull Korean Chem Soc*. 2019;40:634–41.
173. Wang Y, Li Y-Z. Cultivation to improve in vivo solubility of overexpressed arginine deiminases in *Escherichia coli* and the enzyme characteristics. *BMC Biotechnol*. 2014;14:53. <https://doi.org/10.1186/1472-6750-14-53>.
174. Wang CH, Zhang C, Xing XH. Metabolic engineering of *Escherichia coli* cell factory for highly active xanthine dehydrogenase production. *Bioresour Technol*. 2017;245:1782–9.
175. Li Z, Wei P, Cheng H, He P, Wang Q, Jiang N. Functional role of  $\beta$  domain in the *Thermoanaerobacter tengcongensis* glucoamylase. *Appl Microbiol Biotechnol*. 2014;98:2091–9.
176. Strandberg L, Enfors SO. Factors influencing inclusion body formation in the production of a fused protein in *Escherichia coli*. *Appl Environ Microbiol*. 1991;57:1669–74.
177. de Marco A. Minimal information: an urgent need to assess the functional reliability of recombinant proteins used in biological experiments. *Microb Cell Fact*. 2008;7:20.
178. Taylor CF, Paton NW, Lilley KS, Binz P-A, Julian RK, Jones AR, et al. The minimum information about a proteomics experiment (MIAPE). *Nat Biotechnol*. 2007;25:887–93.
179. Kwong KWY, Ng AKL, Wong WKR. Engineering versatile protein expression systems mediated by inteins in *Escherichia coli*. *Appl Microbiol Biotechnol*. 2016;100:255–62.
180. Sharma SS, Campbell JW, Frisch D, Blattner FR, Harcum SW. Expression of two recombinant chloramphenicol acetyltransferase variants in highly reduced genome *Escherichia coli* strains. *Biotechnol Bioeng*. 2007;98:1056–70. <https://doi.org/10.1002/bit.21491>.
181. Yeom S-J, Kim YJ, Lee J, Kwon KK, Han GH, Kim H, et al. Long-term stable and tightly controlled expression of recombinant proteins in antibiotics-free conditions. *PLoS ONE*. 2016;11:e0166890.
182. Gu P, Yang F, Su T, Wang Q, Liang Q, Qi Q. A rapid and reliable strategy for chromosomal integration of gene(s) with multiple copies. *Sci Rep*. 2015;5:9684.
183. Kim J, Copley SD. The orphan protein bis- $\gamma$ -glutamylcystine reductase joins the pyridine nucleotide disulfide reductase family. *Biochemistry*. 2013;52:2905–13. <https://doi.org/10.1021/bi4003343>.
184. Gröning JAD, Eulberg D, Tischler D, Kaschabek SR, Schlömann M. Gene redundancy of two-component (chloro)phenol hydroxylases in *Rhodococcus opacus* 1CP. *FEMS Microbiol Lett*. 2014;361:68–75. <https://doi.org/10.1111/1574-6968.12616>.

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

