






# Heterologous expression of naturally evolved putative *de novo* proteins with chaperones

Lars A. Eicholt<sup>1</sup>  | Margaux Aubel<sup>1</sup>  | Katrin Berk<sup>1</sup>  |  
Erich Bornberg-Bauer<sup>1,2</sup>  | Andreas Lange<sup>1</sup> 

<sup>1</sup>Institute for Evolution and Biodiversity,  
University of Muenster, Münster,  
Germany

<sup>2</sup>Max Planck-Institute for Biology  
Tuebingen, Tübingen, Germany

## Correspondence

Andreas Lange, Institute for Evolution  
and Biodiversity, University of Muenster,  
Huefferstraße 1, 48149 Muenster,  
Germany.

Email: [andreas.lange@wwu.de](mailto:andreas.lange@wwu.de)

## Funding information

Deutsche Forschungsgemeinschaft, Grant/  
Award Number: 281125614/GRK2220;  
Volkswagen Foundation, Grant/Award  
Number: 98183

**Review Editor:** John Kuriyan

## Abstract

Over the past decade, evidence has accumulated that new protein-coding genes can emerge *de novo* from previously non-coding DNA. Most studies have focused on large scale computational predictions of *de novo* protein-coding genes across a wide range of organisms. In contrast, experimental data concerning the folding and function of *de novo* proteins are scarce. This might be due to difficulties in handling *de novo* proteins *in vitro*, as most are short and predicted to be disordered. Here, we propose a guideline for the effective expression of eukaryotic *de novo* proteins in *Escherichia coli*. We used 11 sequences from *Drosophila melanogaster* and 10 from *Homo sapiens*, that are predicted *de novo* proteins from former studies, for heterologous expression. The candidate *de novo* proteins have varying secondary structure and disorder content. Using multiple combinations of purification tags, *E. coli* expression strains, and chaperone systems, we were able to increase the number of solubly expressed putative *de novo* proteins from 30% to 62%. Our findings indicate that the best combination for expressing putative *de novo* proteins in *E. coli* is a GST-tag with T7 Express cells and co-expressed chaperones. We found that, overall, proteins with higher predicted disorder were easier to express.

**Statement:** Today, we know that proteins do not only evolve by duplication and divergence of existing proteins but also arise from previously non-coding DNA. These proteins are called *de novo* proteins. Their properties are still poorly understood and their experimental analysis faces major obstacles. Here, we aim to present a starting point for soluble expression of *de novo* proteins with the help of chaperones and thereby enable further characterization.

## KEYWORDS

chaperones, *de novo* protein, disorder and secondary structure prediction, *Drosophila melanogaster*, *Homo sapiens*, Western blot

Lars A. Eicholt and Margaux Aubel have contributed equally.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *Protein Science* published by Wiley Periodicals LLC on behalf of The Protein Society.

## 1 | INTRODUCTION

*De novo* genes originate from intergenic or non-coding DNA regions<sup>1–7</sup> in contrast to genes that emerge by duplication<sup>8,9</sup> or rearrangement from existing gene fragments.<sup>10</sup> Therefore, recent, true *de novo* genes have no precursor by definition and have not been subjected to selection for particular structures or functions for long, if at all. Due to their recent emergence, *de novo* genes tend to be shorter, evolve more rapidly, and have lower expression than established genes.<sup>3,4</sup> Short-length and accelerated evolution make it difficult to reliably detect (or reject) homologs of orphan genes and thereby identify true *de novo* genes. By combining homology and synteny based approaches for *de novo* gene identification, the origin of *de novo* genes can be detected more accurately.<sup>11</sup>

Several *de novo* protein-coding genes have been identified and confirmed across a wide range of eukaryotes.<sup>12–22</sup> These *de novo* genes were mainly analyzed with comparative genomics and transcriptomics. A recent study by Grandchamp et al.<sup>23</sup> showed that proto-genes, an intermediate step in *de novo* gene emergence,<sup>24</sup> contain regulatory sequences similar to established genes. Depending on the genomic position of the recently emerged proto-gene, introns may already be present in the proto-gene, making it harder to distinguish from established genes. However, without experimental evidence on structure and function, our evolutionary understanding of how *de novo* proteins emerge, remains incomplete.

Difficulties in handling *de novo* proteins, together with the novelty of the research area, might be the reason for the lack of experimental studies on *de novo* proteins. So far only two *de novo* proteins were expressed and characterized experimentally, Goddard (Gdrd)<sup>25</sup> and Bsc4.<sup>26</sup> In both cases, the expressed *de novo* protein was difficult to analyze due to unstable or incorrect folding (Bsc4) or unusual behavior in SDS-PAGE (Gdrd). Compared to well-studied proteins with expression and purification data available, *de novo* proteins tend to behave differently when using standard protocols.

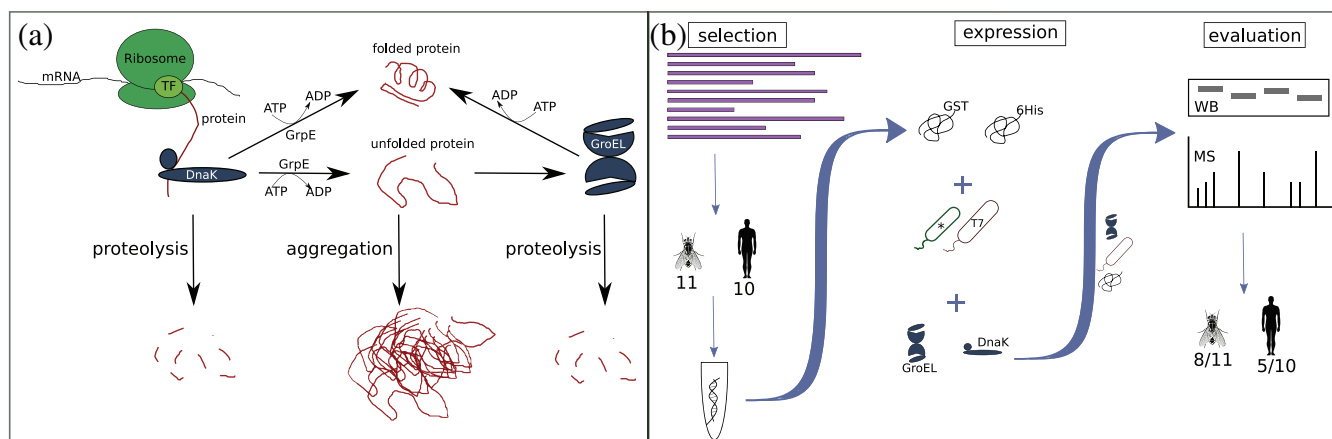
Several studies, foremost some from the laboratory of Dan Tawfik,<sup>27–30</sup> inspired us to apply co-expression with chaperones to achieve soluble expression of *de novo* proteins. Since *de novo* proteins evolve rapidly by becoming coding from scratch, they probably lack a stable structural configuration and contain high amounts of disorder.<sup>3,4</sup> Those properties determine the levels of soluble and insoluble fractions of a protein during *in vitro* experiments and could explain the obstacles faced during their expression.<sup>31,32</sup> On the other hand, it is not yet clear if *de novo* proteins undergo a similar hindrance in their native

organism or only in the expression hosts.<sup>33</sup> While Tawfik and colleagues used chaperones to explore the sequence space of enzymes and enable soluble expression of mutants,<sup>27–29</sup> we hypothesized that *de novo* protein expression might also profit from chaperones. With their “emergence from dark genomic matter” in the DNA<sup>34</sup> and predicted lack of stability and high disorder, *de novo* proteins are prospective targets for chaperones because their solubility can be increased.<sup>27,28</sup> Increased solubility can be relevant for protein purification and any follow-up experiments.

The chaperonin GroEL and its co-chaperone GroES are found throughout the bacterial domain, while their homologs, HSP60 and HSP10, respectively, are found in eukaryotes.<sup>35</sup> GroEL/GroES play a pivotal role in the translocation, dis-aggregation, function, and folding of newly synthesized peptides after translation.<sup>27,35–37</sup>

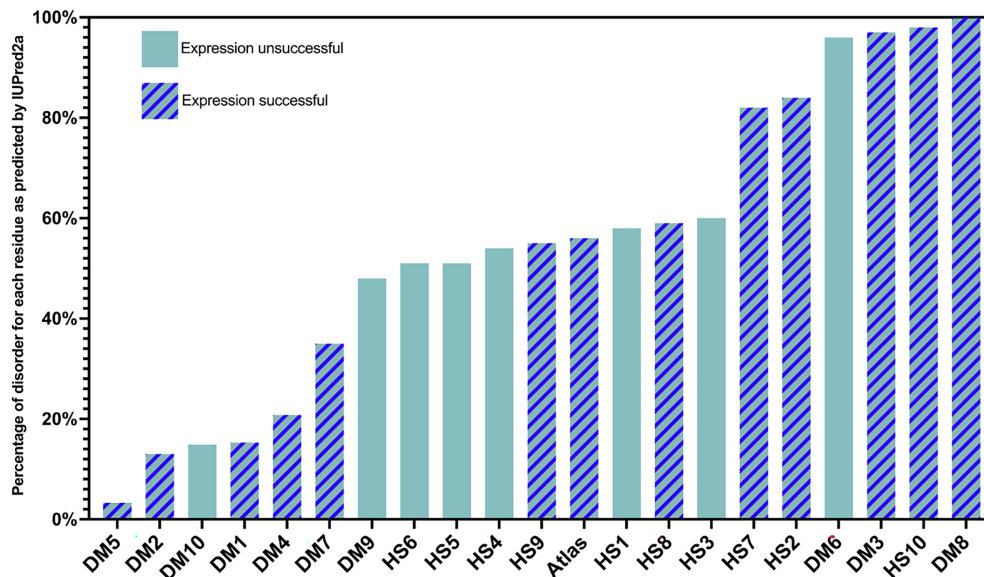
The other chaperone system used here is DnaK, DnaJ, and GrpE (homologous to HSP70 and HSP40 in eukaryotes). For simplicity we will refer to the chaperone system GroEL/ GroES as only GroEL and to DnaK, DnaJ, and GrpE as DnaK only. While the GroEL system targets misfolded and unfolded proteins, DnaK can refold an already aggregated protein to its native state using ATP (see Figure 1a).<sup>39–42</sup> The two different chaperone systems can be exploited for challenging heterologous expression of proteins which are foreign to the host and thus prevent misfolding and aggregation which is often associated with heterologous expression.<sup>27–29,35,41,43</sup>

For this study, we used 21 putative *de novo* proteins, 11 from *Drosophila melanogaster* (termed here as DM1-10 and Atlas) and 10 from *Homo sapiens* (termed here as HS1-10) as shown in Figure 1b). The sequences of all used putative *de novo* proteins can be found on Zenodo (<https://doi.org/10.5281/zenodo.6512224>), for genomic location and official gene names see Table S1. These *de novo* proteins have been recently published by Heames et al.<sup>21</sup> and Dowling et al., respectively.<sup>22</sup> Additionally, we tested our method on a recently published and better characterized putative *de novo* protein from *D. melanogaster*, called Atlas. Atlas appears to function as a DNA binding protein that facilitates the packaging of chromatin in developing *D. melanogaster* sperm.<sup>44</sup> Since experimental work with *de novo* proteins is still under-represented (compared with computational studies) and challenging, we want to propose a guideline for successful expression of putative *de novo* proteins in *E. coli*. We combined different chaperone systems (GroEL and DnaK) with different combinations of *E. coli* strains (BL21 Star™ [DE3] and T7 Express) in order to express putative *de novo* proteins solubly. To verify successful



**FIGURE 1** (a) Mechanism of chaperone assisted protein folding after Thomas et al.<sup>38</sup> The nascent protein is bound by the DnaK/J complex and release is triggered by GrpE under ATP hydrolysis. After release, the protein is either correctly folded, degraded (proteolysis), or remains unfolded. The unfolded protein can either aggregate or bind to the GroEL/ES complex. GroEL/ES either releases the folded protein by ATP hydrolysis or the protein is degraded. (b) Overview of the workflow on *de novo* protein expression: We first selected candidate proteins from *Drosophila melanogaster* (11, including Atlas) and 10 from *Homo sapiens* from a pool of putative *de novo* genes for expression. The 21 sequences were codon optimized for *E. coli* and ordered from Twist. For expression, different tags (GST and His), different *E. coli* expression cells (star, T7), and different chaperones (GroEL and DnaK systems) were tested. The success of protein expression was verified by Western blot (WB) and mass spectrometry (MS)

**FIGURE 2** Percentage of disorder as calculated with IUPred2a. All candidate *de novo* proteins used for expression experiments ordered by their disorder level from the left to right. Unicolor bars belong to the unsuccessfully expressed proteins, striped bars to the successfully expressed ones



expression of target proteins, Western blots were performed and samples sent for tryptic digest followed by mass spectrometry. We identified the best combination for expression of putative *de novo* proteins in *E. coli*. After first expressions with His-tag alone resulted in soluble expression for only 1/21 proteins, we increased the total number of solubly expressed putative *de novo* proteins to 13/21 with GST-tag and chaperones. The different chaperone systems increased or enabled soluble expression in four cases, while DnaK only helped in two, GroEL in all of those four.

## 2 | RESULTS

### 2.1 | Structural content of the putative *de novo* proteins

#### 2.1.1 | Disorder predictions

We performed disorder predictions with IUPred2a<sup>45,46</sup> on all candidate *de novo* proteins. For this we calculated the percentage of residues predicted to be disordered (Figure 2), as opposed to the overall average disorder

score (Figure S1). This allows direct comparison to secondary structure predictions (Figure 3). Our first objective here was to choose candidate *de novo* proteins with different levels of intrinsic disorder to observe any difference in their ability to express. If any trend in predicted disorder and soluble expression or susceptibility to chaperones was observed, this could help choosing promising candidates for characterization in future experiments. The predicted disorder ranged from around 3%–100% as shown in Figure 2. *DM5* was predicted to have least disorder content, while *DM6*, *DM3*, *HS10*, and *DM8* appear to be entirely disordered. The putative *de novo* protein Atlas has predicted disorder of 60%.

### 2.1.2 | Secondary structure predictions

Predictions of secondary structure elements were performed using Porter 5.0<sup>47,48</sup> for all candidate proteins and are shown in Figure 3. The predicted random coils should be equivalent to the disordered regions predicted by IUPred2a.<sup>49</sup> While the results indicate a high amount of random coils for most candidates, they do not completely follow the trend of the disorder predictions by IUPred2a (compare Figure 2). *DM3*, for example, is predicted to be ~100% disordered by IUPred2a, while, on the other hand, it is predicted to have over 20%  $\beta$ -sheet and ~70% random coils by Porter 5.0.

Our goal was to choose a cohort of *de novo* proteins that consist of a diverse range in composition of structural elements. We assumed that a protein containing more secondary structure elements should be better

accessible for soluble expression with chaperones.<sup>50</sup> Notably, *DM1* (~70%  $\alpha$ -helical), *DM2* (~70%  $\alpha$ -helical), *DM4* (~55%  $\alpha$ -helical, ~10%  $\beta$ -sheets), *DM5* (~60%  $\alpha$ -helical, ~10%  $\beta$ -sheets), and *DM10* (~70%  $\alpha$ -helical, ~10%  $\beta$ -sheets) are predicted to have secondary structure contents of 50% or more, with  $\alpha$ -helices to be more frequent than  $\beta$ -sheets and less than 50% random coils. *HS1* (~50% random coils, ~45%  $\alpha$ -helical, ~5%  $\beta$ -sheets), *HS3* (~70% random coils, ~25%  $\alpha$ -helical, ~5%  $\beta$ -sheets), *HS4* (~65% random coils, ~30%  $\alpha$ -helical, ~5%  $\beta$ -sheets), *HS5* (~55% random coils, ~5%  $\alpha$ -helical, ~40%  $\beta$ -sheets), *HS6* (~60% random coils, ~5%  $\alpha$ -helical, ~35%  $\beta$ -sheets), *HS7* (~70% random coils, ~5%  $\alpha$ -helical, ~25%  $\beta$ -sheets), *DM3* (~70% random coils, ~30% sheets), *DM7* (~65% random coils, ~10%  $\alpha$ -helical, ~25%  $\beta$ -sheets), and *DM9* (~60% random coils, ~10%  $\alpha$ -helical, ~30%  $\beta$ -sheets), on the other hand, are predicted to be mostly random coils (disordered) with otherwise higher amounts of  $\beta$ -sheets predicted. *DM6* (~90% random coils), *DM8* (~100% random coils), *HS2* (~85% random coils), *HS9* (~95% random coils), and *HS10* (~100% random coils) are predicted to contain more or less only random coils.

## 2.2 | Expression of putative *de novo* proteins

### 2.2.1 | Candidate proteins of *Drosophila melanogaster*

Our initial approach was similar to the successful expression of characterized putative *de novo* protein Gdrd.<sup>25</sup>

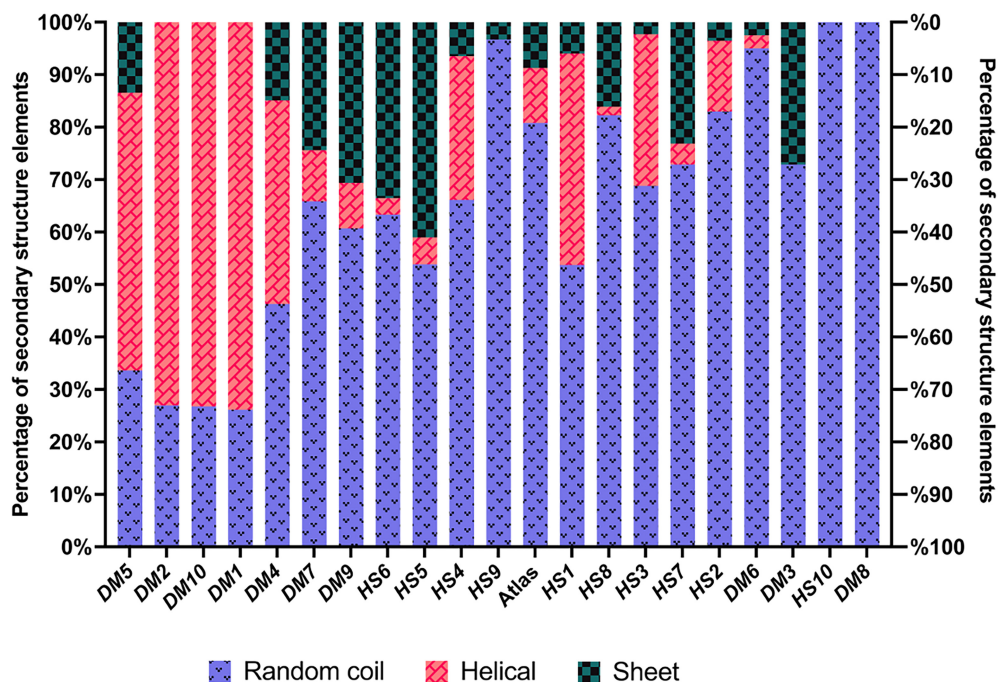


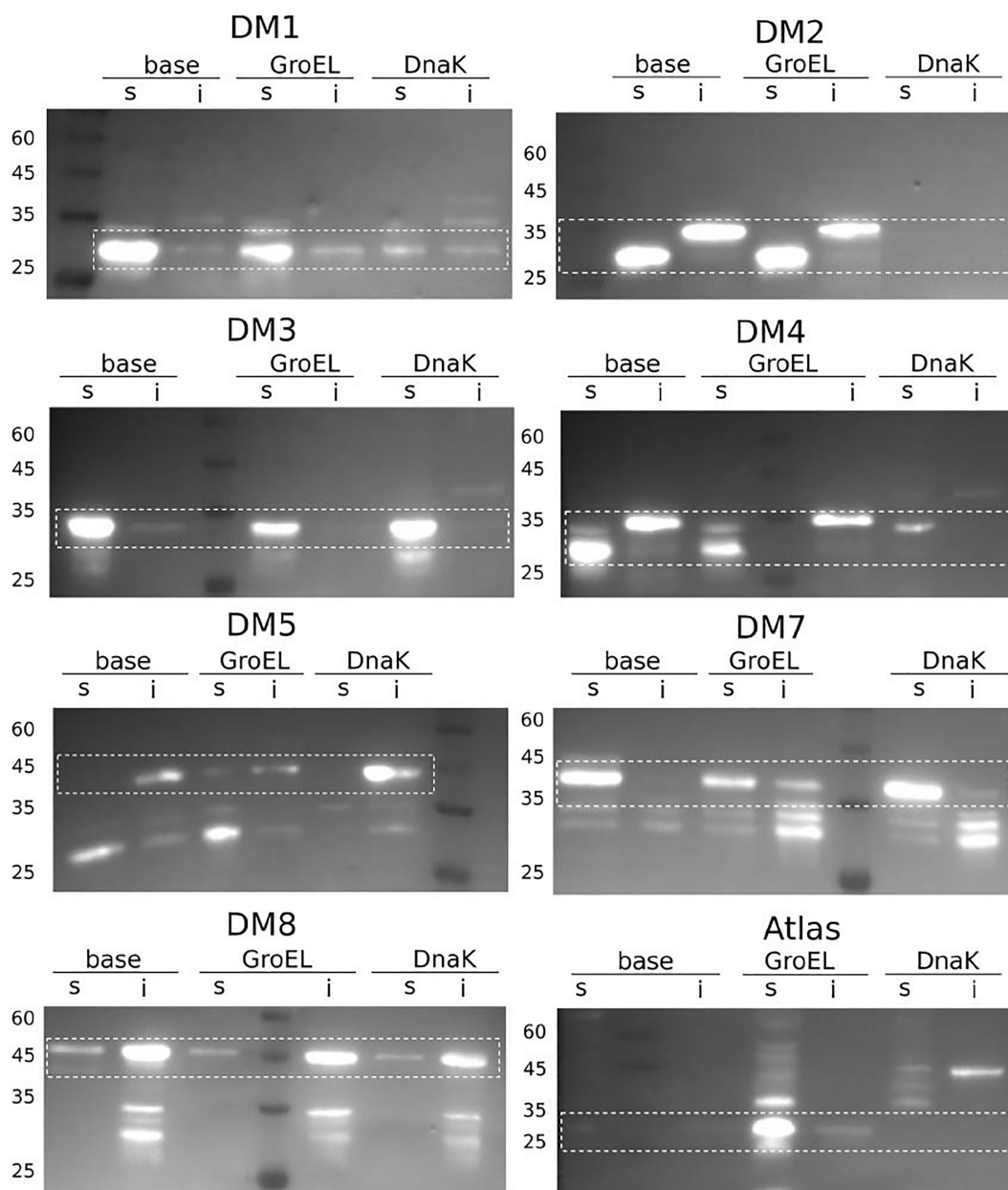
FIGURE 3 Percentage of random coils,  $\alpha$ -helices, and  $\beta$ -sheets predicted by Porter 5.0 for each *de novo* protein candidate. Left to right following increasing disorder level based on Figure 2



Therefore, we aimed to express our **11** putative *de novo* protein candidates with an N-terminal 6xHis-tag in *E. coli* BL21 Star™ (DE3) cells, and verify expression via SDS-PAGE and mass spectrometry. However, for our candidates, the expression level was either very low or not detectable, as can be seen in Figure S3. We switched to different *E. coli* cells (T7 Express), but expression remained unsuccessful. Shifting from an N-terminal 6xHis-Tag to a C-terminal 6xHis-tag showed similar negative results. Considering the size and levels of disorder,

we switched to a larger tag for increased solubility and stability, choosing an N-terminal GST-tag. In this way, we were able to observe a higher success rate in soluble expression of our target proteins. But not all proteins could be expressed at satisfying levels, especially solubility needed to be increased for some (Figure S3).

Inspired by successful work carried out by Tawfik et al.,<sup>27–29</sup> we hypothesized that chaperones could improve thermodynamic stability of these evolutionarily young proteins thus enabling their soluble expression.



**FIGURE 4** Western blots with anti-His antibody. Boxes indicate the height of the target protein band: **DM1** (34 kDa): highest solubility without chaperones, then GroEL, then DnaK; highly soluble. **DM2** (36 kDa): only insoluble, even with chaperones. **DM3** (33 kDa): DnaK highest solubility, then base, then GroEL; very soluble. **DM4** (34 kDa): DnaK highest solubility, then GroEL, then base; very insoluble. **DM5** (39 kDa): GroEL only one with soluble fraction, runs a bit high. **DM7** (36 kDa): DnaK highest solubility, then base, then GroEL very soluble. **DM8** (37 kDa): all similar, different expression levels, first base, then GroEL, then DnaK; more insoluble. **Atlas** (20 kDa): GroEL highest solubility, nothing in base and DnaK

We repeated our experiments with the addition of the two chaperone systems (i) GroEL and (ii) DnaK. We were able to increase the number of solubly expressed *de novo* candidate proteins of *D. melanogaster* using the combination of either GroEL or DnaK and N-terminal GST-tag (see Figure 4). However, for the candidate proteins DM6, DM9, and DM10 no soluble expression was achievable, despite the use of different tags, strains, or chaperones. Only in the case of Atlas, the combination of N-terminal 6xHis-tag and GroEL worked best. We tested all combinations in BL21 Star™ (DE3) and T7 Express *E. coli* cells. Six candidate proteins were expressed in T7, two were expressed in BL21 Star™ (DE3) cells. Three proteins were not expressible in either strain. In summary, with the combination of chaperones and switching to N-terminal GST-tag, we were able to express 8/11 of the

*D. melanogaster* putative *de novo* protein candidates (see Table 1).

### 2.2.2 | Comparison of different chaperone conditions for *D. melanogaster* proteins

Western blots were used for comparison of the soluble expression levels with and without chaperones, in order to test our hypothesis that chaperones would increase soluble expression of the target proteins. The optimal conditions identified by SDS-PAGEs were repeated under three settings: (i) without chaperones (base), (ii) with GroEL, and (iii) with DnaK. Surprisingly, we did not observe increased solubility for most putative *de novo* proteins when adding chaperones (see Figure 4 and Table 1).

**TABLE 1** Expression conditions and results of *D. melanogaster de novo* proteins. Base = no chaperones, GroEL = GroEL/ES, DnaK = DnaK/J/GrpE. Plus signs mean visible expression, two plus signs strong expression, 0 means no visible expression. Arrows indicate the change in expression with added chaperones in comparison with base

Protein	Cell/tag	Base	GroEL	DnaK	GroEL effect	DnaK effect	Disorder (%)
DM1	T7/GST	++	++	+	–	↓	15
DM2	Star/GST	++	++	0	–	↓	13
DM3	T7/GST	++	+	++	↓	–	97
DM4	T7/GST	++	+	+	↓	↓	21
DM5	T7/GST	0	+	0	↑	–	3
DM6	–/–	0	0	0	–	–	97
DM7	T7/GST	++	+	++	↓	–	35
DM8	T7/GST	+	+	+	–	–	100
DM9	–/–	0	0	0	–	–	49
DM10	–/–	0	0	0	–	–	15
Atlas	Star/6xHis	0	++	0	↑	–	56

**TABLE 2** Expression conditions and results of *H. sapiens de novo* proteins. Base = no chaperones, GroEL = GroEL/ES, DnaK = DnaK/J/GrpE. Plus signs mean visible expression, two plus signs strong expression, 0 means no visible expression. Arrows indicate the change in expression with added chaperones in comparison with base

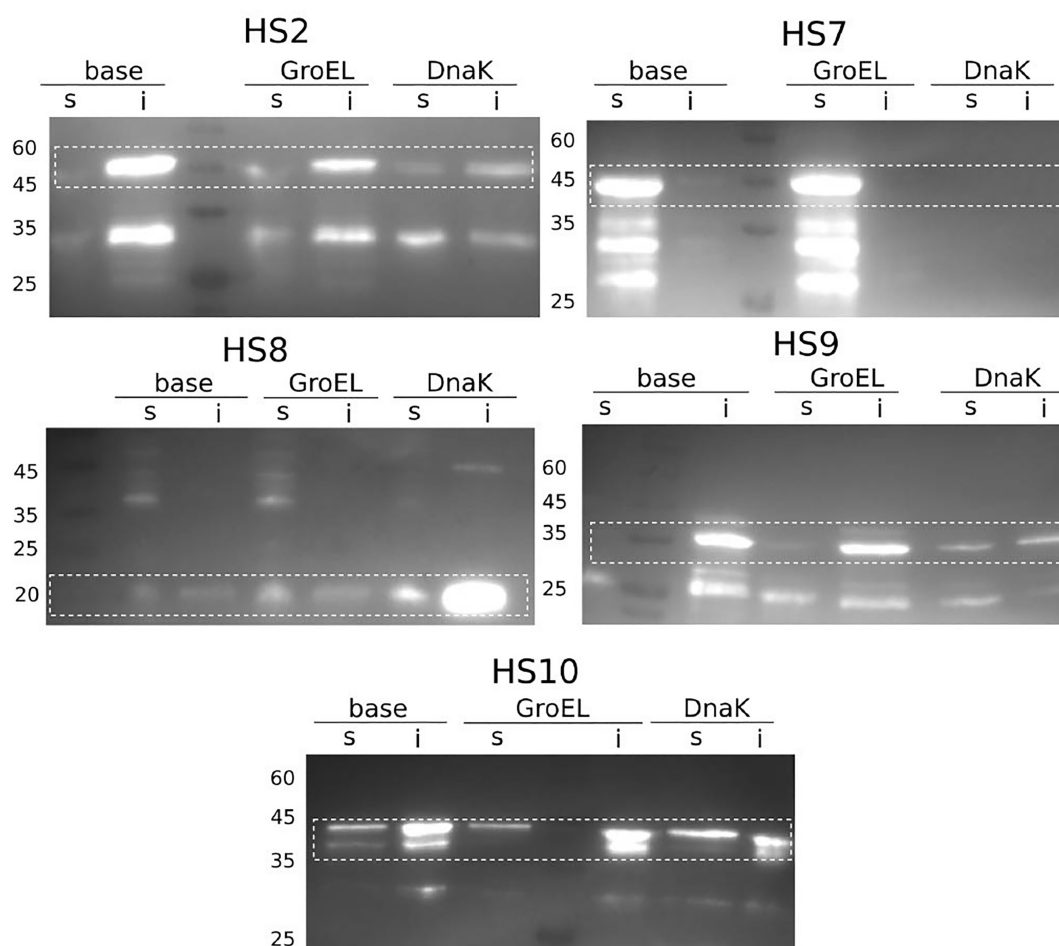
Protein	Cell/tag	Base	GroEL	DnaK	GroEL effect	DnaK effect	Disorder (%)
HS1	–/–	0	0	0	–	–	59
HS2	T7/GST	0	+	+	↑	↑	84
HS3	–/–	0	0	0	–	–	60
HS4	–/–	0	0	0	–	–	54
HS5	–/–	0	0	0	–	–	51
HS6	–/–	0	0	0	–	–	51
HS7	T7/GST	++	++	0	–	↓	82
HS8	T7/6xHis	+	+	+	–	–	60
HS9	T7/GST	0	+	+	↑	↑	56
HS10	T7/GST	+	+	+	–	–	99

In contrast, we observed soluble expression for most proteins without chaperones, for example, *DM1*, *DM2*, *DM3*, *DM4*, and *DM7*. In combination with GroEL, the intensity of the bands in the soluble fraction and therefore amount of soluble protein, even decreased for *DM3*, *DM4*, and *DM7*. For *DM2* and *DM5*, the amount of soluble protein increased when co-expressed with GroEL. When DnaK was co-expressed, protein solubility either appeared to decrease (*DM1*, *DM2*, and *DM4*) or was similar to the base (*DM3* and *DM7*). *DM8* showed similar soluble expression for all three conditions with most of the protein being insoluble. In the case of Atlas and *DM5*, soluble protein expression was increased or enabled with the addition of the GroEL chaperone system while DnaK and base expression resulted in no or very little soluble protein. While we cannot confirm that co-expression with DnaK in fact decreases the amount of soluble protein (*DM1*, *DM2*, and *DM4*), we do not see increased soluble

expression for any of the candidate proteins in the presence of DnaK as we do for GroEL (*DM5* and Atlas).

### 2.2.3 | Candidate proteins of *Homo sapiens*

The 10 putative human *de novo* proteins were expressed following the same protocol as the *D. melanogaster* proteins by combining the different *E. coli* expression cells, tags, and chaperone systems (Figure S4). We detected a similar trend here as for the *D. melanogaster* proteins (N-terminal GST-tag in *E. coli* T7 express cells; see Table 2). One protein (*HS8*), however, was only weakly expressed with an N-terminal 6xHis-tag but using also *E. coli* T7 express cells. Without the addition of chaperones only *HS7*, *HS8*, and *HS10* were successfully expressed and soluble. After co-expression with chaperones, as described for *D. melanogaster* proteins, two more *H. sapiens*



**FIGURE 5** Western blots with anti-His antibody. Boxes indicate the height of the target protein band: **HS2** (44 kDa): upper bands (lower are degraded protein or double bands) most in DnaK, then GroEL, then base; very insoluble. **HS7** (50 kDa): GroEL best, then base, nothing in DnaK. Possible protein degradation; very soluble. **HS8** (16 kDa): upper bands most in DnaK, then GroEL, then base; very insoluble. **HS9** (42 kDa): upper bands (lower are degraded protein or double bands) most in DnaK, then GroEL, then base; very insoluble. **HS10** (43 kDa): upper bands (lower are degraded protein or double bands) most in GroEL, then DnaK, then base; very insoluble

proteins could be expressed. Unfortunately, *H. sapiens* protein candidates *HS1*, *HS3*, *HS4*, *HS5*, and *HS6* showed no expression at all, even with chaperones. In total, we were able to express 5 out of 10 putative *de novo* proteins following our protocol (see Table 2).

### 2.2.4 | Comparison of different chaperone conditions for *H. sapiens* proteins

Western blots were used for comparison of the three different chaperone expressions (i) base, (ii) GroEL, and (iii) DnaK, as described above. Two out of the five successful candidates (*HS2* and *HS9*) showed very weak or no soluble expression without chaperones, but solubility could be increased with both chaperone systems. *HS8* and *HS10* showed low soluble expression overall, but no change in solubility was visible when co-expressing with either chaperone system. The candidate *de novo* protein *HS7* already showed strong soluble expression at base (Figure 5). However, the addition of GroEL seemed to increase soluble expression further, while DnaK co-expression led to low or no protein being detected. Overall, the trend observed for the *D. melanogaster* proteins was consistent with the trend observed for the *H. sapiens* proteins. GroEL increased soluble expression for most putative *de novo* proteins while DnaK lacked substantial influence on protein solubility.

## 3 | DISCUSSION

*De novo* proteins have first been detected more than a decade ago and the mechanism of their emergence has been studied intensely ever since.<sup>4,12</sup> Still, there are concerns (i) regarding the reliability of their computational identification<sup>24,51,52</sup> and (ii) if and how they code for functional proteins. To shed light on these concerns, *de novo* proteins need to be studied experimentally and theoretically. The handling of *de novo* proteins by heterologous expression and purification is often difficult because solubility is low and purification yields little amounts and potentially unstable proteins. Moreover, identifying the function of these young genes, is another challenging task. In this study, we present a guideline for expressing *de novo* proteins in *E. coli*.

### 3.1 | Expression cells

*E. coli* is the most widely used model organism for recombinant expression. However, foreign proteins can be toxic to *E. coli* by interfering with the physiology or leading to

protein aggregation. This may result in low expression yields, growth defects, or even cell death.<sup>53–55</sup> Optimized expression hosts and plasmids<sup>53–55</sup> or chaperones can be used to overcome the expression issues caused by proteins which are a metabolic burden for the host. Here, we used two different types of the *E. coli* strains (DE3): BL21 Star™ and T7 Express. Both strains resulted in effective protein expression and a relatively high yield of the *de novo* proteins, with T7 Express being the best option. The *de novo* proteins studied here are possibly a toxic, metabolic burden to the *E. coli* cells, suggesting T7 cells are the better choice of expression cell. BL21 Star™ (DE3) contains a T7-RNA-polymerase under control of lacUV5 promoter together with higher mRNA stability. This leads to stable mRNA transcripts and higher amount of target protein. However, BL21 Star™ (DE3) cells have increased basal expression of heterologous genes and cannot express toxic genes. In contrast, the T7 Express cells have a reduced basal expression of target proteins than BL21 Star™ (DE3) cells. Therefore, toxic proteins can be expressed better in T7 cells compared with BL21 Star™ (New England Biolabs).<sup>56</sup>

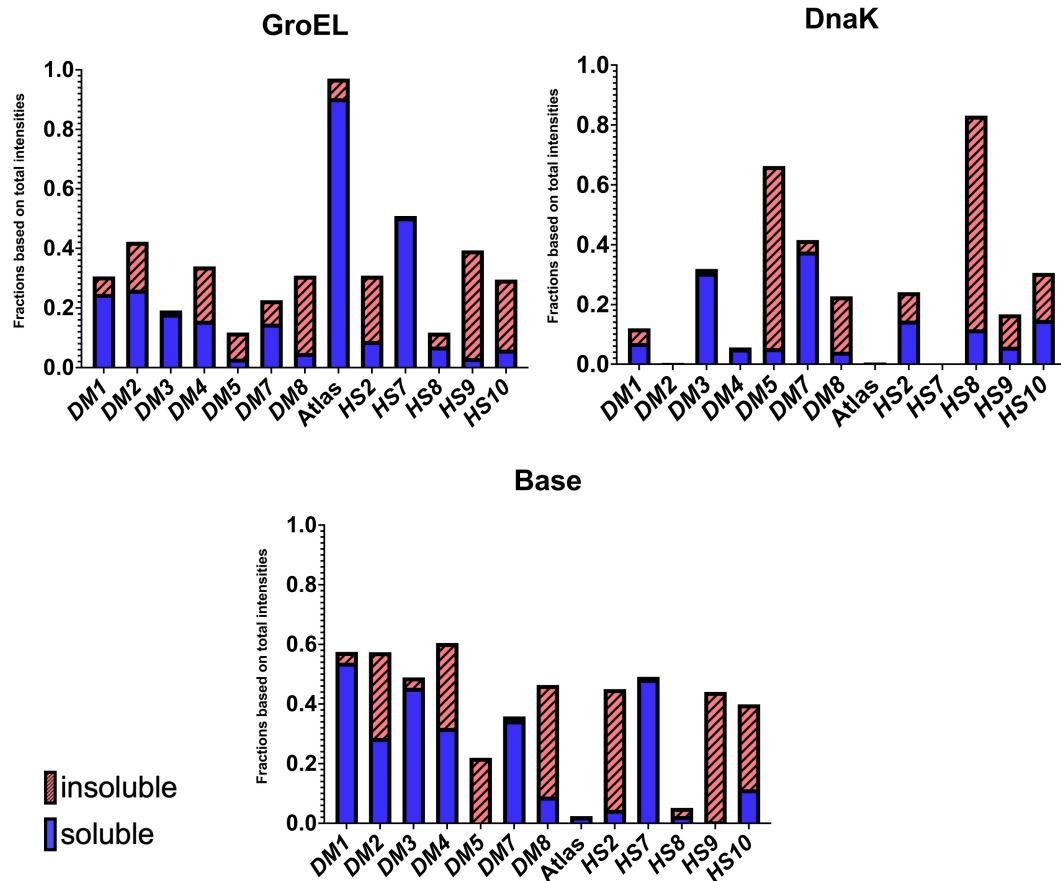
### 3.2 | Comparing different protein tags

Based on our study, an N-terminal GST-tag was the more appropriate choice than a 6xHis-tag. Some *de novo* protein candidates are quite small (8–12 kDa), so a larger tag like GST might already stabilize in a chaperone-like manner.<sup>55,57</sup> However, Atlas and *HS8*, that is, 2/21, were only expressed with an N-terminal 6xHis-tag. With a mass of only 1 kDa, 6xHis-tag is the better choice for further structural characterization using circular dichroism (CD), multi-angle light scattering (MALS) or nuclear magnetic resonance (NMR), since a small tag has less influence on protein folding. In contrast, the larger GST-tag needs to be cleaved for most follow-up experiments. When removing the tag, the *de novo* protein might behave differently and could degrade or aggregate.

### 3.3 | Influence of chaperones on protein expression and solubility

Our Western blot results indicate that GroEL slightly outperforms DnaK in terms of increased protein solubility. In some cases, both chaperone systems increase or enable soluble expression (*HS2* and *HS9*, 2/21) but for most proteins GroEL leads to more soluble protein than DnaK (*DM1*, *DM2*, *DM5*, Atlas, and *HS7*, 5/21) (Figure 6). DnaK requires easily accessible hydrophobic fragments that can be predicted from the protein sequence, while





**FIGURE 6** Fractions of soluble and insoluble expression for candidate *de novo* proteins with expression of GroEL or DnaK and without (base). Values are intensities of Western blot bands

GroEL demands no defined binding motifs. However, in the case of our proteins, we found no connection between predicted DnaK binding sites and influence of DnaK on protein expression level (Figure S2). While GroEL could increase solubility for single *de novo* proteins in 5/21 cases, this system was not successful in Heames et al.,<sup>7</sup> in which we used a library of 1800 putative *de novo* proteins (4–8 kDa) in a cell-free expression system.

In some cases (*DM1*, *DM2*, *HS7*, and *Atlas*), DnaK decreased the solubility of proteins that expressed soluble without chaperones. One could assume that DnaK and target protein expression compete for cellular resources and that for already soluble proteins this has no positive trade-off. Another reason could be DnaK's role as central organizer in the chaperone network of *E. coli*. Binding to DnaK could trigger degradation of the “toxic” target protein but also overexpression of DnaK additionally to the endogenous version could bring the cellular metabolism out of balance.<sup>58,59</sup>

We cannot verify that changes with co-expression of chaperones is solely due to effects of chaperones on putative *de novo* proteins or on overall amount of protein expression. Our main interest here is to optimize

expression for follow-up experiments and not to draw general conclusions on chaperone interaction with *de novo* proteins.

Drawing conclusions from heterologous expression experiments toward *in vivo* interactions of proteins and chaperone systems are fragmentary and can only serve as hypotheses in need of further verification using *in vivo* experiments.<sup>60</sup>

### 3.4 | Comparing putative *de novo* proteins from *D. melanogaster* to *H. sapiens*

In total, we were able to successfully express 13 out of 21 putative *de novo* proteins in *E. coli* cells (eight in *D. melanogaster* and five in *H. sapiens*). For both, *D. melanogaster* and *H. sapiens* candidate putative *de novo* proteins, the combination of GST-tag and *E. coli* T7 Express cells were the best performing (10 out of 13). We performed test expressions and compared the levels of soluble expression for different chaperone combinations shown in Figures 4 and 5. Expression results from putative *de novo* protein candidates *DM5*, *Atlas*, *HS2*, and

*HS9* were in line with our original hypothesis that chaperones enhance solubility of *de novo* proteins in heterologous expression systems. However, the choice of appropriate tag and expression cells in the first step was equally, if not more, important. When using the N-terminal His-tag that proved successful for putative *de novo* protein Gdrd, only two (Atlas and *HS7*) of our candidate proteins were expressed. When switching to the N-terminal GST-tag another seven *D. melanogaster* and four more *H. sapiens* protein candidates were expressed. Unfortunately, we were not able to express 8/21 of the candidate proteins in *E. coli* at all (*HS1*, *HS3* – *HS6*, *DM6*, *DM9*, and *DM10*), despite trying different expression strains, tags and chaperone systems.

### 3.5 | Disorder and secondary structure predictions

When examining the predicted structural properties of the human *de novo* protein candidates, we observe a slight trend toward better expression of the more disordered proteins. This trend can be observed for the IUPred2a disorder predictions (Figure 2) but becomes more apparent for the overall secondary structure predictions (Figure 3). The unsuccessful expression candidates *HS1*, *HS3*, and *HS4* showed a higher predicted  $\alpha$ -helical content of approximately 40% while *HS5* and *HS6* had a higher predicted  $\beta$ -sheet content of around 30%–40% compared with the other human candidate proteins *HS2*, *HS7*, *HS8*, *HS9*, and *HS10* which are predicted to contain over 70% random coils (or 60% disorder). The described differences in predicted secondary structure content and disorder level might be the reason why these putative *de novo* candidates could not be expressed in *E. coli* cells even with the help of chaperones.

For *D. melanogaster* protein candidates, this trend was not observed. Here, several of the proteins with lower disorder predicted (*DM1*, *DM4*, and *DM7*) were expressed solubly without addition of chaperones. Yet, *DM6* (~90% disorder predicted) was not expressed successfully. However, the two proteins with 100% random coils predicted by Porter 5.0 and highest disorder predictions by IUPred2a (*DM8* and *HS10*) did not show any change in solubility when chaperones were co-expressed. Considering that such highly disordered proteins do not need chaperones, this observation was expected.

Deviations of the level of predicted disorder and predicted secondary structures, especially random coils, for each protein can be explained by the differences in IUPred2a and Porter 5.0. IUPred2a provides energy estimations for each amino acid residue resulting in quasi-probabilities of disorder.<sup>46</sup> On the other hand, Porter 5.0

is based on a neural network relying on sequence alignments and co-evolutionary information.<sup>47</sup> These fundamentally different approaches can lead to inconsistent results in some cases (e.g., *HS9* and *DM3*) while not invalidating one another.

## 4 | CONCLUSION AND OUTLOOK

Exemplifying the general trend for soluble *de novo* protein expression is only the first step toward enabling further *in vitro* experiments for functional and structural characterization. Further advancement will lead to efficient and stable purification, followed by functional assays such as peptide phage display to identify binding partners.<sup>61,62</sup> This technique has proven to be useful for high-throughput screening of intrinsically disordered regions for short linear motifs,<sup>63</sup> especially for human proteins. Soluble expression and purification will be crucial for structural characterization via CD, NMR, and Cryo-EM. Due to their small size and high disorder content, only NMR<sup>25</sup> and potentially Cryo-EM<sup>64</sup> will be capable of solving the structure of *de novo* proteins experimentally. Even in light of the recent dawn of computational structure prediction,<sup>65,66</sup> experimental structural and functional determination remains necessary, especially for *de novo* proteins. While contemporary prediction methods can certainly provide a first estimate on structure, the intrinsic nature of *de novo* proteins, with their short length, high disorder content and lack of homology, will demand some scepticism while analyzing such predictions.<sup>67,68</sup> Surprisingly, the AlphaFold prediction of Goddard (GEO12017p1, AlphaFold Protein Structure Database)<sup>69</sup> is in line with its partial experimental characterization<sup>25</sup> and its central helix is predicted with high confidence. Despite the lack of homology, which is a core demand for the MSA generation of AlphaFold2 and a hallmark of *de novo* proteins, one could assume that *de novo* proteins are of such small size that AlphaFold can solve their local folding. This will have to be validated in future studies. This study of 21 putative *de novo* proteins from *H. sapiens* and *D. melanogaster*, including previously *in vivo* characterized putative *de novo* protein Atlas, showed that chaperones may help expressing *de novo* proteins in *E. coli* cells. However, not all putative *de novo* proteins needed chaperones for soluble expression and sometimes even expressed better without. Fusion of the target *de novo* proteins to a GST-tag and using T7 Express cells as hosts proved to be the most successful combination. Our work may serve as a guide to facilitating future analyses of putative *de novo* proteins or other difficult (short and/or disordered) target proteins in *E. coli*.

## 5 | MATERIALS AND METHODS

### 5.1 | Online data availability

All SDS-PAGEs, MS results, Western blots, and scripts are deposited in Zenodo database (<https://doi.org/10.5281/zenodo.6512224>).

### 5.2 | Computational methods

#### 5.2.1 | Candidate selection

We selected a total of 21 putative *de novo* protein candidates. Ten are uncharacterized putative *de novo* proteins from *Homo sapiens*<sup>22</sup> and are referred to here as HS1-10. Ten proteins originate from *Drosophila melanogaster*<sup>21</sup> and are referred to as DM1-10. One is the functionally characterized putative *de novo* protein Atlas from *D. melanogaster*.<sup>44</sup> The 21 candidates contain different levels of disorder and secondary structure elements ( $\alpha$ -helix,  $\beta$ -sheet, and mixture of both) and different sequence lengths (see Figure 2). We selected only candidate sequences without exon/intron structure and without long single amino acid repeats. All putative *de novo* proteins have confirmed expression in their native organism.

#### 5.2.2 | Predictions

We performed disorder predictions with IUPred2a<sup>45,46</sup> using default options *long disorder* for entire proteins. We calculated the average disorder score of the whole sequence and percentage of residues predicted to be disordered. The percentage of disorder was calculated by taking the amount of disordered residues (disorder score > 0.5) and dividing it by the sequence length of the protein. We also predicted average disorder and percentage of disordered residues with a disorder threshold of 0.8 (Figure S1). A python script was used to automate predictions and disorder proportion for all candidates. We performed  $\alpha$ -helix and  $\beta$ -sheet predictions to verify the amount of disordered residues predicted by IUPred2a. Secondary structure predictions were performed with Porter 5.0 (SS3).<sup>47,48</sup> The predicted secondary structure elements for each residue were counted with a Javascript and divided by the total number of residues to obtain a percentage score for each structural element. DnaK binding sites were predicted using the ChaperISM suite (v1) in quantitative mode with default settings.<sup>70</sup>

### 5.3 | Experimental methods

#### 5.3.1 | Cloning of putative *de novo* candidates

Putative *de novo* candidates were synthesized as strings DNA from Twist Bioscience, San Francisco, codon optimized for *E. coli* and without restriction sites used for cloning (BamHI, HindIII, NcoI, XhoI) inside the sequence. The wild-type DNA for Atlas was provided by Geoff Findlay. To introduce restriction sites at the ends, we used different primers (a fasta file containing the DNA sequences and primer used can be found online on Zenodo (<https://doi.org/10.5281/zenodo.6512224>)). For cloning into pHAT2 vector (N-terminal 6xHis) we used restriction enzymes combination of BamHI/XhoI + HindIII, for pETM-30 (N-terminal 6xHis-GST-TEV), we used NcoI+HindIII. Both vectors were from the EMBL vector database, Heidelberg, introduced stop-codon was TAA for all constructs. We digested the PCR product with both restriction enzymes respectively (FastDigest, Thermo Scientific) for 3 h at 37°C. Digest of the vector (1 h, 37°C) was purified from agarose gel (Zymo Research). We ligated both with an insert:vector ratio of 1:4 using Ligase (Thermo Scientific; 1 h, 22°C). The ligation mix was purified (Zymo Research) and 2  $\mu$ l of the purified reaction mix was used to transform into 50  $\mu$ l of chemically competent *E. coli* TOP10 cells. Cells were incubated for 30 min on ice, followed by a 90 sec heat-shock at 42°C. 500  $\mu$ l of LB-Media (5 g yeast extract, 6 g tryptone, 5 g NaCl) was added for recovery and incubated for 1 h at 37°C. After incubation, the resuspended cell pellets were plated on LB-agar containing 50  $\mu$ g/ml ampicillin (AMP, Carl Roth, pHAT2, and EMBL vector database) or Kanamycin (KAN, Carl Roth, pETM-30, and EMBL vector database) and incubated at 37°C over night.

Successful transformation was verified by colony PCR and sequencing at Microsynth, SeqLab, Germany. The plasmid DNA bearing the chaperone combinations GroEL/ES (pGro7) or DnaK/J/GrpE (pKJE) from Takara Biotech chaperone kit<sup>71,72</sup> were first transformed into *E. coli* Top10 cells and then into expression strains (BL21 Star™ (DE3) and T7 Express). Chaperone plasmid bearing cells were made chemically competent (Inoue method)<sup>73,74</sup> and used for transformation with the plasmid containing the target protein sequence. Final expression cells contained two plasmids: chaperone plasmid and target protein plasmid. The chaperone plasmids are chloramphenicol (CAM) resistant, so the double plasmid cells are either AMP + CAM (pHAT2, N-terminal 6xHis-tag) or KAN + CAM (pETM-30, N-terminal GST-tag) resistant.

### 5.3.2 | Test-Expression of candidate *de novo* proteins

To identify in which strain and plasmid proteins were expressed we performed test expressions. 10 ml of LB + AMP + CAM or LB + KAN + CAM were inoculated from a glycerol stock of all three expression cells bearing both plasmids (target protein and chaperone) and grown until turbid (6–8 h, 37°C). We split the solutions into 3 × 3 ml and incubated for 30 min at different temperatures (37°C, 28°C, and 20°C) before adding IPTG (Carl Roth) for a final concentration of 0.5 mM and shaking over night. When using the cells with chaperone plasmids we made the following adjustment: L-arabinose (final concentration 3 mM, Carl Roth) was added from the beginning for immediate induction of chaperone expression. Therefore, after inducing the *de novo* protein expression with IPTG the chaperones were already present in order to help folding the *de novo* proteins.

A total of 500 µL of each cell culture were centrifuged (15,000 rpm, 2 min). Pellets were resuspended and lysed in 50 µl of a mix of Bugbuster and Lysonase (both Merck AG) through vortexing for 10 min. After centrifugation the supernatant was mixed with the same volume of SDS-loading buffer (standard). The pellet was resuspended in 5x diluted Bugbuster, centrifuged, and resuspended in 50 µl SDS-loading buffer. 15 µl of each fraction was loaded on an SDS-PAGE, either 10% Bis-Tris or 12.5% TGS, run on 200 V for 50 min and dyed using ReadyBlue™ staining.

For the final Western blots, the determined optimal combination of strain, expression vector, and chaperone plasmid were used. 20 ml cultures of 2YT + AMP + CAM or 2YT + KAN + CAM were inoculated with 1 ml of the overnight culture. L-arabinose (final concentration 3 mM) was added to the samples, but not to the control without chaperones and grown at 37°C, 180 rpm for 4–6 hr until turbid. The cultures were incubated at 28°C, 180 rpm for 30 min before induction with IPTG (final concentration 0.5 mM) and incubated overnight under these conditions. Final samples were harvested and handled as prior performed test expressions.

### 5.3.3 | Western blot

The SDS-PAGEs were run as described above but without ReadyBlue™ staining. The gel was equilibrated in transfer buffer (20% Methanol) for a few seconds. A polyvinylidene fluoride (PVDF) membrane with a pore size of 0.22 µm was activated by methanol (2 min) and equilibrated in transfer buffer. The semi-dry transfer was performed at 25 V for 30 min using the BioRad standard protocol. The membrane was blocked at room

temperature for 1 hr using 5% bovine serum albumin BSA in phosphate-buffered saline with tween (PBS-T) then washed in PBS-T and incubated for 1 h with anti-His antibody (MA1-21315-HRP) diluted 1:500. For chemiluminescence, 0.5 ml luminol was mixed with 0.5 ml peroxide and distributed evenly on the membrane. Intensities of the different bands were measured in ImageJ after default background subtraction. The different fractions (soluble/insoluble of base, GroEL and DnaK) are calculated as fractions of the intensities of all relevant protein bands to also compare the amount of expression between the different conditions.

### 5.3.4 | Mass spectrometry

Tryptic digest followed by mass spectrometry for peptide detection of the candidate proteins was performed by the Core Unit Proteomics group of Prof. Dr. Simone König, UKM Muenster.

### AUTHOR CONTRIBUTIONS

**Lars A. Eicholt:** Data curation (lead); investigation (lead); validation (lead); visualization (lead); writing – original draft (lead). **Margaux Aubel:** Data curation (equal); investigation (equal); validation (equal); visualization (equal); writing – original draft (equal). **Katrin Berk:** Data curation (supporting); validation (supporting); writing – original draft (supporting). **Erich Bornberg-Bauer:** Conceptualization (supporting); funding acquisition (lead); project administration (supporting); resources (lead); supervision (supporting); writing – original draft (supporting). **Andreas Lange:** Conceptualization (lead); investigation (supporting); project administration (lead); supervision (lead); validation (supporting); writing – original draft (supporting).

### ACKNOWLEDGMENTS

A.L., E.BB, and M.A. were funded by the Volkswagen Stiftung (VWF), grant code 98183. K.B. was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—281125614/GRK2220. Thanks to Colin Jackson who invited us to be part of this special issue. We thank Prof. Geoff Findlay (College of the Holy Cross, Boston, Massachusetts) for the Atlas WT-sequence, Anne Diehl (FMP Berlin) for the BL21 Star™ (DE3) and T7 Express cells, our master students Kai Köstler and Roman Schauerte for their help in the laboratory, and Prof. Simone König from the Core Unit Proteomic facility for performing the tryptic digest and mass spectrometry. We thank Mark Harrison for comments on the manuscript. Open Access funding enabled and organized by Projekt DEAL.



## CONFLICT OF INTERESTS

The authors declare no competing interests.

## ORCID

Lars A. Eicholt  <https://orcid.org/0000-0002-3985-3698>

Margaux Aubel  <https://orcid.org/0000-0003-1653-9441>

Katrin Berk  <https://orcid.org/0000-0002-8734-8047>

Erich Bornberg-Bauer  <https://orcid.org/0000-0002-1826-3576>

Andreas Lange  <https://orcid.org/0000-0003-3871-1986>

## REFERENCES

- Tautz D, Domazet-Lošo T. The evolutionary origin of orphan genes. *Nat Rev Genet.* 2011. ISSN 1471-0056;12(10):692–702. <https://doi.org/10.1038/nrg3053>.
- McLysaght A, Hurst LD. Open questions in the study of de novo genes: What, how and why. *Nat Rev Genet.* 2016. ISSN 1471-0064;17(9):567–578. <https://doi.org/10.1038/nrg.2016.78>.
- Schmitz JF, Bornberg-Bauer E. Fact or fiction: Updates on how protein-coding genes might emerge de novo from previously non-coding DNA. *F1000Research.* 2017. ISSN 2046-1402;6:57. <https://doi.org/10.12688/f1000research.10079.1>.
- Van Van Oss SB, Carvunis A-R. De novo gene birth. *PLOS Genetics.* 2019. ISSN 1553-7404;15(5):e1008160. <https://doi.org/10.1371/journal.pgen.1008160>.
- Rödelsperger C, Prabh N, Sommer RJ. New gene origin and deep taxon phylogenomics: Opportunities and CHALLENGES. *Trends in Genetics.* 2019. ISSN 0168-9525;35(12):914–922. <https://doi.org/10.1016/j.tig.2019.08.007>.
- Bornberg-Bauer E, Hlouchova K, Lange A. Structure and function of naturally evolved de novo proteins. *Curr Opin Struct Biol.* 2021. ISSN 1879-033X;68:175–183. <https://doi.org/10.1016/j.sbi.2020.11.010>.
- Heames B, Buchel F, Aubel M, et al. Experimental characterisation of de novo proteins and their unevolved random-sequence counterparts. *bioRxiv.* 2022. <https://doi.org/10.1101/2022.01.14.476368>.
- Liberles DA, Kolesov G, Dittmar K. Understanding gene duplication through biochemistry and population genetics. Evolution after gene duplication. John Wiley & Sons, Ltd, 2011. ISBN 978-0-470-61990-2; p. 1–21. <https://doi.org/10.1002/9780470619902.ch1>.
- Ohno S. Evolution by gene duplication. Springer-Verlag, New York, 1970. s. <https://doi.org/10.1002/tera.1420090224>.
- Bornberg-Bauer E, Albà MM. Dynamics and adaptive benefits of modular protein evolution. *Current Opinion in Structural Biology.* 2013. ISSN 1879-033X;23(3):459–466. <https://doi.org/10.1016/j.sbi.2013.02.012>.
- Vakirlis N, Carvunis A-R, McLysaght A. Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *eLife.* 2020. ISSN 2050-084X;9:e53500. <https://doi.org/10.7554/eLife.53500>.
- Begun DJ, Lindfors HA, Thompson ME, Holloway AK. Recently evolved genes identified from drosophila Yakuba and D. Erecta accessory gland expressed sequence tags. *Genetics.* 2006. ISSN 0016-6731, 1943-2631;172(3):1675–1681. <https://doi.org/10.1534/genetics.105.050336>.
- Cai J, Zhao R, Jiang H, Wang W. De novo origination of a new protein-coding gene in *saccharomyces cerevisiae*. *Genetics.* 2008. ISSN 0016-6731, 1943-2631;179(1):487–496. <https://doi.org/10.1534/genetics.107.084491>.
- Neme R, Tautz D. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genomics.* 2013. ISSN 1471-2164;14(1):117. <https://doi.org/10.1186/1471-2164-14-117>.
- McLysaght A, Guerzoni D. New genes from non-coding sequence: The role of de novo protein-coding genes in eukaryotic evolutionary innovation. *Philosophical Transactions of the Royal Society B: Biological Sciences.* 2015. ISSN 0962-8436, 1471-2970;370(1678):20140332. <https://doi.org/10.1098/rstb.2014.0332>.
- Schlötterer C. Genes from scratch – The evolutionary fate of de novo genes. *Trends Genet.* 2015. ISSN 0168-9525;31(4):215–219. <https://doi.org/10.1016/j.tig.2015.02.007>.
- Schmitz JF, Ullrich KK, Bornberg-Bauer E. Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. *Nature Ecology & Evolution.* 2018. ISSN 2397-334X;2(10):1626–1632. <https://doi.org/10.1038/s41559-018-0639-7>.
- Vakirlis N, Hebert AS, Oplente DA, et al. A molecular portrait of De novo genes in yeasts. *Mol Biol Evol.* 2018. ISSN 0737-4038;35(3):631–645. <https://doi.org/10.1093/molbev/msx315>.
- Prabh N, Rödelsperger C. De novo, divergence, and mixed origin contribute to the emergence of orphan genes in *Pristionchus Nematodes*. *G3: Genes, Genomes, Genetics.* 2019. ISSN 2160-1836;9(7):g3.400326.2019. <https://doi.org/10.1534/g3.119.400326>.
- Zhang L, Ren Y, Yang T, et al. Rapid evolution of protein diversity by de novo origination in *Oryza*. *Nature Ecology & Evolution.* 2019. ISSN 2397-334X;3(4):679. <https://doi.org/10.1038/s41559-019-0822-5>.
- Heames B, Schmitz J, Bornberg-Bauer E. A continuum of evolving de novo genes drives protein-coding novelty in *Drosophila*. *J Mol Evol.* 2020;88:382–398. <https://doi.org/10.1007/s00239-020-09939-z>.
- Dowling D, Schmitz JF, Bornberg-Bauer E. Stochastic gain and loss of novel transcribed open reading frames in the human lineage. *Genome Biol Evol.* 2020;12:2183–2195. <https://doi.org/10.1093/gbe/evaa194>.
- Grandchamp A, Berk K, Dohmen E, Bornberg-Bauer E. New genomic signals underlying the emergence of human proto-genes. *Genes.* 2022;13(2):284. <https://doi.org/10.3390/genes13020284>.
- Domazet-Lošo T, Carvunis A-R, Albà MM, et al. No evidence for Phylostratigraphic bias impacting inferences on patterns of gene emergence and evolution. *Mol Biol Evol.* 2017. ISSN 0737-4038;34(4):843–856. <https://doi.org/10.1093/molbev/msw284>.
- Lange A, Patel PH, Heames B, et al. Structural and functional characterization of a putative de novo gene in *Drosophila*. *Nat Commun.* 2021. ISSN 2041-1723;12(1):1667. <https://doi.org/10.1038/s41467-021-21667-6>.
- Bungard D, Copple JS, Yan J, et al. Foldability of a natural De novo evolved protein. *Structure.* 2017. ISSN 0969-2126;25(11):1687–1696.e4. <https://doi.org/10.1016/j.str.2017.09.006>.

27. Tokuriki N, Tawfik DS. Chaperonin overexpression promotes genetic variation and enzyme evolution. *Nature*. 2009a;459:668–673. <https://doi.org/10.1038/nature08009>.
28. Tokuriki N, Tawfik DS. Protein dynamism and evolvability. *Science*. 2009b;324:203–207. <https://doi.org/10.1126/science.1169375>.
29. Tokuriki N, Tawfik DS. Stability effects of mutations and protein evolvability. *Curr Opin Struct Biol*. 2009c;19(5):596–604. <https://doi.org/10.1016/j.sbi.2009.08.003>.
30. Jackson C, Toth-Petroczy A, Kolodny R, et al. Adventures on the routes of protein evolution – In memoriam Dan Salah Tawfik (1955–2021). *Journal of Molecular Biology*. 2022. ISSN 0022-2836; 434(7):167462. <https://doi.org/10.1016/j.jmb.2022.167462>.
31. Soskine M, Tawfik DS. Mutational effects and the evolution of new protein functions. *Nat Rev Genet*. 2010;11:572–582. <https://doi.org/10.1038/nrg2808>.
32. Tretyachenko V, Vymětal J, Bednářová L, et al. Random protein sequences can form defined secondary structures and are well-tolerated in vivo. *Scientific Reports*. 2017. ISSN 2045-2322; 7(1):15449. <https://doi.org/10.1038/s41598-017-15635-8>.
33. Gasser B, Saloheimo M, Rinas U, et al. Protein folding and conformational stress in microbial cells producing recombinant proteins: A host comparative overview. *Microb Cell Fact*. 2008; 7:11. <https://doi.org/10.1186/1475-2859-7-11>.
34. Bornberg-Bauer E, Schmitz JF, Heberlein M. Emergence of de novo proteins from ‘dark genomic matter’ by ‘grow slow and moul’t. *Biochem Soc Trans*. 2015;43(5):867–873. <https://doi.org/10.1042/BST20150089>.
35. Finka A, Mattoo RUH, Goloubinoff P. Experimental milestones in the discovery of molecular chaperones as polypeptide unfolding enzymes. *Annu Rev Biochem*. 2016;85:715–742. <https://doi.org/10.1146/annurev-biochem-060815-014124>.
36. Libich DS, Tugarinov V, Clore GM. Intrinsic unfoldase/foldase activity of the chaperonin GroEL directly demonstrated using multinuclear relaxation-based NMR. *Proceedings of the National Academy of Sciences*. 2015;112:8817–8823. <https://doi.org/10.1073/pnas.1510083112>.
37. Lin Z, Madan D, Rye HS. Gro EL stimulates protein folding through forced unfolding. *Nature Structural & Molecular Biology*. 2008;15:303–311. <https://doi.org/10.1038/nsmb.1394>.
38. Thomas JG, Ayling A, Baneyx F. Molecular chaperones, folding catalysts, and the recovery of active recombinant proteins from *E. coli*. To fold or to refold. *Appl Biochem Biotechnol*. 1997; 66(3):197–238. <https://doi.org/10.1007/BF02785589>.
39. Schröder H, Langer T, Hartl FU, Bukau B. Dnak, dnaj and grpe form a cellular chaperone machinery capable of repairing heat-induced protein damage. *The EMBO Journal*. 1993;12: 4137–4144. <https://doi.org/10.1002/j.1460-2075.1993.tb06097.x>.
40. Sharma SS, Rios PDL, Christen P, Lustig A, Goloubinoff P. The kinetic parameters and energy cost of the hsp70 chaperone as a polypeptide unfoldase. *Nature chemical biology*. 2010;6(12): 914–920. <https://doi.org/10.1038/nchembio.455>.
41. Kim YE, Hipp MS, Bracher A, Hayer-Hartl M, Ulrich Hartl F. Molecular chaperone functions in protein folding and proteostasis. *Annu Rev Biochem*. 2013;82:323–355. <https://doi.org/10.1146/annurev-biochem-060208-092442>.
42. Mashaghi A, Bezrukavnikov S, Minde DP, et al. Alternative modes of client binding enable functional plasticity of hsp70. *Nature*. 2016;539:448–451. <https://doi.org/10.1038/nature20137>.
43. Goloubinoff P, Gatenby AA, Lorimer GH. Groe heat-shock proteins promote assembly of foreign prokaryotic ribulose biphosphate carboxylase oligomers in escherichia coli. *Nature*. 1989; 337:44–47. <https://doi.org/10.1038/337044a0>.
44. Rivard EL, Ludwig AG, Patel PH, et al. A putative de novo evolved gene required for spermatid chromatin condensation in *Drosophila melanogaster*. *PLOS Genetics*. 2021. ISSN 1553-7404; 17(9):e1009787. <https://doi.org/10.1371/journal.pgen.1009787>.
45. Erdős G, Dosztányi Z. Analyzing protein disorder with IUPred2A. *Current Protocols in Bioinformatics*. 2020. ISSN 1934-340X;70(1):e99. <https://doi.org/10.1002/cpbi.99>.
46. Mészáros B, Erdős G, Dosztányi Z. IUPred2A: Context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res*. 2018. ISSN 0305-1048;46(W1): W329–W337. <https://doi.org/10.1093/nar/gky384>.
47. Torrisi M, Kaleel M, Pollastri G. Porter 5: Fast, state-of-the-art ab initio prediction of protein secondary structure in 3 and 8 classes. *bioRxiv*. 2018. <https://doi.org/10.1101/289033>.
48. Torrisi M, Kaleel M, Pollastri G. Deeper profiles and cascaded recurrent and convolutional neural networks for state-of-the-art protein secondary structure prediction. *Sci Rep*. 2019;9, Article number: 12374. <https://doi.org/10.1038/s41598-019-48786-x>.
49. Linding, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J., & Russell, R. B. (2003). Protein Disorder Prediction. *Structure*, 11(11), 1453–1459. <https://doi.org/10.1016/j.str.2003.10.002>
50. Hervás, R., & Oroz, J. (2020). Mechanistic Insights into the Role of Molecular Chaperones in Protein Misfolding Diseases: From Molecular Recognition to Amyloid Disassembly. *International Journal of Molecular Sciences*, 21(23), 9186. <https://doi.org/10.3390/ijms21239186>
51. Moyers BA, Zhang J. Phylostratigraphic bias creates spurious patterns of genome evolution. *Molecular Biology and Evolution*. 2015. ISSN 0737-4038, 1537-1719;32(1):258–267. <https://doi.org/10.1093/molbev/msu286>.
52. Weisman CM, Murray AW, Eddy SR. Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes. *Curr. Biol*. 2022;S0960-9822 (22)00721-7. <https://doi.org/10.1016/j.cub101/2022.01.13.476251>.
53. Saïda F, Uzan M, Odaert B, Bontems F. Expression of highly toxic genes in *E. coli*: Special strategies and genetic tools. *Current Protein & Peptide Science*. 2006. ISSN 1389-2037;7(1):47–56. <https://doi.org/10.2174/138920306775474095>.
54. Saïda F. Overview on the expression of toxic gene products in *Escherichia coli*. *Current Protocols in Protein Science*, 2007. ISSN 1934-3663;50:5.19.1-5.19.13. <https://doi.org/10.1002/0471140864.ps0519s50>.
55. Rosano GL, Ceccarelli EA. Recombinant protein expression in *Escherichia coli*: Advances and challenges. *Front Microbiol*. 2014. ISSN 1664-302X;5:172. <https://doi.org/10.3389/fmicb.2014.00172>.
56. New England Biolabs. Datasheet for t7 express competent *e. coli* (high efficiency) (c2566; lot 18). <https://www.nebiolabs.com.au/-/media/catalog/datacards-or-manuals/c2566datasheet-lot18.pdf?rev=234841213ece47a48f9da8de895ca3db&hash=CB482DAE0DA6659F3B5B7618615B4902> (Accessed on February 24, 2022).
57. Harper S, Speicher DW. Purification of proteins fused to glutathione S-transferase. *Methods in Molecular Biology* (Clifton, N.J.). 2011. ISSN 1940-6029;681:259–280. [https://doi.org/10.1007/978-1-60761-913-0\\_14](https://doi.org/10.1007/978-1-60761-913-0_14).

58. Calloni G, Chen T, Schermann SM, et al. Dnak functions as a central hub in the e. coli chaperone network. *Cell Reports*. 2012. ISSN 2211-1247;1(3):251–264. <https://doi.org/10.1016/j.celrep.2011.12.007>.
59. Christensen S, Rämisch S, André I. Dnak response to expression of protein mutants is dependent on translation rate and stability. *bioRxiv*. 2021. <https://doi.org/10.1101/2021.09.29.462496>.
60. Niwa T, Kanamori T, Ueda T, Taguchi H. Global analysis of chaperone effects using a reconstituted cell-free translation system. *Proc Natl Acad Sci U S A*. 2012;109(23):8937–8942. <https://doi.org/10.1073/pnas.1201380109>.
61. Sundell GN, Ivarsson Y. Interaction analysis through proteomic phage display. *BioMed Research International*. 2014. ISSN 2314-6141;2014:176172. <https://doi.org/10.1155/2014/176172>.
62. Ivarsson Y, Arnold R, McLaughlin M, et al. Large-scale interaction profiling of PDZ domains through proteomic peptide-phage display using human and viral phage peptidomes. *Proceedings of the National Academy of Sciences*. 2014. ISSN 0027-8424, 1091-6490; 111(7):2542–2547. <https://doi.org/10.1073/pnas.1312296111>.
63. Ali M, Simonetti L, Ivarsson Y. Screening intrinsically disordered regions for short linear binding motifs. In: Kragelund BB, Skriver K, editors. *Intrinsically disordered proteins: Methods and protocols, methods in molecular biology*. US, New York, NY: Springer, 2020. ISBN 978-1-07-160524-0; p. 529–552. [https://doi.org/10.1007/978-1-0716-0524-0\\_27](https://doi.org/10.1007/978-1-0716-0524-0_27).
64. Chiu Y-H, Ko KT, Yang T-J, et al. Direct visualization of a 26 kda protein by cryo-electron microscopy aided by a small scaffold protein. *Biochemistry*. 2021;60(14):1075–1079. <https://doi.org/10.1021/acs.biochem.0c00961>.
65. Jumper JM, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with alphafold. *Nature*. 2021;596:583–589. <https://doi.org/10.1038/s41586-021-03819-2>.
66. Baek M, DiMaio F, Anishchenko I, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*. 2021;373(6557):871–876. <https://doi.org/10.1126/science.abj8754>.
67. Ruff KM, Pappu RV. Alphafold and implications for intrinsically disordered proteins. *J Mol Biol*. 2021;433(20):167208. <https://doi.org/10.1016/j.jmb.2021.167208>.
68. Monzon V, Haft DH, Bateman A. Folding the unfoldable: Using alphafold to explore spurious proteins. *Bioinformatics Advances*. 2022;2(1):vbab043. <https://doi.org/10.1093/bioadv/vbab043>.
69. Varadi M, Anyango S, Deshpande M, et al. AlphaFold protein structure database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*. 2021. ISSN 0305-1048;50(D1):D439–D444. <https://doi.org/10.1093/nar/gkab1061>.
70. Gutierrez MBB, Bonorino C, Rigo MM. Chaperism: Improved chaperone binding prediction using position-independent scoring matrices. *Bioinformatics*. 2020;36(3):735–741. <https://doi.org/10.1093/bioinformatics/btz670>.
71. Nishihara K, Kanemori M, Kitagawa M, Yanagi H, Yura T. Chaperone coexpression plasmids: Differential and synergistic roles of dnak-dnaj-grpe and groel-groes in assisting folding of an allergen of japanese cedar pollen, cryj2, in escherichia coli. *Appl Environ Microbiol*. 1998;64:1694–1699. <https://doi.org/10.1128/AEM.64.5.1694-1699.1998>.
72. Nishihara K, Kanemori M, Yanagi H, Yura T. Overexpression of trigger factor prevents aggregation of recombinant proteins in escherichia coli. *Appl Environ Microbiol*. 2000;66:884–889. <https://doi.org/10.1128/AEM.66.3.884-889.2000>.
73. Inoue H, Nojima H, Okayama H. High efficiency transformation of Escherichia coli with plasmids. *Gene*. 1990. ISSN 0378-1119; 96(1):23–28. [https://doi.org/10.1016/0378-1119\(90\)90336-p](https://doi.org/10.1016/0378-1119(90)90336-p).
74. Sambrook J, Russell DW. The Inoue method for preparation and transformation of competent E. Coli: “Ultra-competent” cells. *Cold Spring Harbor Protocols*. 2006. ISSN 1940-3402, 1559-6095;2006(1):pdb.prot3944. <https://doi.org/10.1101/pdb.prot3944>.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Eicholt LA, Aubel M, Berk K, Bornberg-Bauer E, Lange A. Heterologous expression of naturally evolved putative *de novo* proteins with chaperones. *Protein Science*. 2022; 31(8):e4371. <https://doi.org/10.1002/pro.4371>