Charles University

Faculty of Science

Study programme: Molecular and Cellular Biology, Genetics and Virology

Mgr. Štěpán Herynek

OPTIMIZATION OF RECOMBINANT PROTEIN PRODUCTION IN *E. COLI* EXPRESSION SYSTEM BY GENETIC ENGINEERING METHODS

OPTIMALIZACE PRODUKCE REKOMBINANTNÍCH PROTEINŮ V EXPRESNÍM SYSTÉMU *E. COLI* METODAMI GENOVÉHO INŽENÝRSTVÍ

Doctoral thesis

Supervisor: prof. Ing. Bohdan Schneider, DSc.

Laboratory of Biomolecular Recognition

Institute of Biotechnology of the Czech Academy of Sciences, v. v. i.

Prague, 2024

Declaration:

I hereby declare that I have completed this PhD thesis independently and thesis is based on my own research. All publications and other used sources have been properly cited. This thesis has not been submitted previously in order to obtain the same or any other academic degree.

In Vestec 25. 3. 2025

Acknowledgements

I would like to thank my supervisor prof. Bohdan Schneider, and my unofficial consultant RNDr. Pavel Mikulecký, PhD for their leadership, unrelenting optimism and patience with my project.

I would like to thank all my colleagues from the Laboratory of Biomolecular Recognition at the Institute of Biotechnology, Academy of Sciences of the Czech Republic, for their help and for creating such a pleasant workplace. I am especially grateful to Jakub Svoboda for the illustrations he created to help me present my project, for his assistance with data analysis, and, along with my colleague RNDr. Maroš Huličiak, PhD, for making our PhD studies not only challenging but also an unforgettable experience. I am also grateful to our lab manager, RNDr. Lubica Škultétyová, PhD, for keeping up the good mood in the lab while ensuring everything ran smoothly. Finally, yet importantly, I would like to thank my family for their support during my studies.

This work was supported by project nr. LX22NPO5102 (MEYS): Financed by European Union – Next Generation EU.

Abstract

In this study, I introduce an innovative approach to optimize sequences coding for the N-termini of recombinant proteins to enhance their production yield in *Escherichia coli*. Previous findings by others have indicated that specific sequences immediately following the initiation codon, collectively referred to here as the N-terminal sequences, have a profound influence on protein expression. While these earlier efforts have mainly relied on a limited selection of rationally designed sequences, I employed a directed evolution-based approach that screens large libraries of diversified N-terminal sequences. To streamline the identification of constructs with optimized expression levels, I fused a GFP marker to the C-terminus of the target proteins and used fluorescence-activated cell sorting to isolate cells showing higher expression levels based on their fluorescence intensity. This systematic workflow enabled me to achieve up to a 30-fold increase in the production yield of soluble recombinant proteins across various constructs. In this research, I built upon methodologies and experience from prior studies on engineering small binding proteins ("scaffolds"), which highlighted the importance of the N-terminal region for recombinant protein production and modified the purpose and construction of DNA libraries for the yield optimization. This work thus bridges methodologies in protein scaffold engineering with N-terminal sequence optimization and underscores the critical role of N-terminal composition of constructs used to produce recombinant proteins.

Abstrakt v češtině

V této práci představuji inovativní přístup k optimalizaci sekvencí kódujících N-konec rekombinantních proteinů, díky čemuž je možné zvýšit jejich výtěžnost v expresním systému *Escherichia coli*. Předchozí studie ukazují, že specifické sekvence bezprostředně následující iniciační kodon, souhrnně označované v této práci jako N-terminální sekvence, mají zásadní vliv na expresi rekombinantních proteinů. Zatímco předchozí výzkumy se zaměřily především na omezený výběr uměle navržených sekvencí, já jsem použil přístup založený na řízené evoluci, který umožňuje prověření rozsáhlých knihoven různorodých N-terminálních sekvencí. Pro zjednodušení identifikace konstruktů s optimalizovanou expresí jsem připojil na C-konce studovaných proteinů GFP, což umožnilo použití buněčné separace založená na principu průtokové cytometrie k izolaci buněk s vyšší úrovní produkce na základě pozorované intenzity fluorescence. Tento systematický postup umožnil dosáhnout až 30násobného zvýšení výtěžků různých rozpustných rekombinantních proteinů. Během tohoto výzkumu jsem navázal na metodiky z předchozího výzkumu umělých malých vazebných proteinů, během kterého se mimo jiné ukázala důležitost N-terminální oblasti pro produkci rekombinantních proteinů. Pro vývoj vazebných proteinů byl aplikován přístup řízené evoluce s využitím DNA knihoven. Tato práce tak propojuje metodologie inženýrství vazebných proteinů s optimalizací N-koncových sekvencí a poukazuje na klíčovou roli složení N-koncové části produkovaných konstruktů.

List of abbreviations

AI	Artificial intelligence
c3GFP	Cycle 3 GFP (we refer to this specific version as a "GFP" for simplicity)
E. coli	*Escherichia coli*
EF-G	Elongation factor G
endA	Endonuclease I
FACS	Fluorescence activated cell sorting
FSC	Forward scatter
GFP	Green fluorescent protein
GST	Glutathione S-transferase
His-tag	Histidine tag
Ig	Immunoglobulin
IL	Interleukin
IFN	Interferon
JAK	Janus kinase
ML	Machine learning
mRNA	messenger RNA
PDB	Protein Data Bank
PTMs	Post translational modifications
PYP	Photoactive yellow protein
RBS	Ribosome binding site
SSC	Side scatter
STAT	Signal transducer and activator of transcription
S. cerevisiae	*Saccharomyces cerevisiae*
SECIS	Selenocysteine insertion sequence
Th	T helper
tRNA	transfer RNA
TrxA	Thioredoxin A
UTR	Untranslated region
WT	Wild type

Contents

1 Introduction 1

1.1 Translation in bacteria 2

1.2 *Escherichia coli* as a host for protein expression 3

1.3 Optimization of protein production by manipulating the N-terminal sequence 4

1.3.1 Importance of the N-terminus composition in protein scaffolds 5

1.3.1.1 Antibodies 5

1.3.1.2 Protein scaffolds 7

1.3.1.2.1 Protein scaffolds vs antibodies 8

1.3.1.3 IL10 as a target 8

1.3.2 mRNA and its impact on the yield of the protein product 11

1.3.3 Nascent protein elements and their impact on the recombinant protein yield 14

1.3.4 Codon usage and its impact on the yield of the protein product 17

1.3.5 Optimizing the N-terminal sequence in order to increase the yield of the protein product 19

1.3.6 Consequences of altered translation speed on protein folding 20

1.3.7 Reporter molecule for directed evolution based selection of N-terminal sequences 20

1.3.8 Introduction to FACS 21

2 Aims and objectives 25

3 Results 26

3.1 Additional unpublished results 27

4 Discussion 32

4.1 Protein scaffolds and evolution of IL10 binder 32

4.1.1 Protein scaffolds and their N-terminal composition importance 33

4.2 Optimizing the sequence coding the N-terminus 34

4.2.1 Successful production optimization of IFNb 35

4.2.2 GFP cleavage during optimization of TrxA production 35

4.2.3 Impact of randomized DNA library size on protein yield 37

4.2.4 Challenges in optimizing well expressed and hard to express proteins 38

4.2.5 N-terminal library decreases the yield before sorting 40

4.3 From unexplained mechanisms to practical use of N-terminal libraries 40

4.4 My research approach in the context of modern biology: Prioritizing practical outcomes over mechanistic descriptions 42

5 Conclusions 44

6 References 45

7 Attachments 55

1 Introduction

Recombinant protein production is the first step in many scientific projects. If this step fails, the entire project may yield no usable results, regardless of the strength or originality of the initial hypothesis. Consequently, careful attention to the initial protocol of protein expression is essential. Over the years, this process has been optimized numerous times in many laboratories, targeting various pathways and testing multiple conditions. Among available host organisms, *Escherichia coli (E. coli)* remains the most widely used for recombinant protein expression, primarily due to its well-established workflows and cost effectiveness. Accordingly, this study focuses on optimizing recombinant protein expression in *E. coli* as a host organism.

To increase the yield of a desired product, I manipulated the N-terminal sequences of expressed genes. The specific composition of sequences encoding the initial amino acids significantly affects the initiation and progression of the elongation phase in translation. Although several studies have investigated this, their findings are inconsistent. I choose not to repeat the effort of many who attempted to rationalize composition of the N-terminal residues by a rational-design approach. Instead, I decided to make use of techniques of directed evolution previously used in our laboratory to produce high-affinity binders [1, 2] and create *de novo* protein mimicking function of interferon lambda [3].

All directed evolution methods require designing a sequentially highly variable DNA library. In my project, the DNA library needed to be introduced into the region coding for the N-terminal part of a protein, whose expression I attempted to increase. To streamline monitoring how different DNA library sequences influence the final protein yield, I fused the protein gene with a fluorescent marker. Cells exhibiting high fluorescence were isolated from the population using fluorescence activated cell sorting (FACS). I then compared the expression levels between cells carrying the wild type (WT) protein and those with N-terminally enhanced variants. This approach successfully identified protein constructs with higher yield. Alongside the positive results, I also discuss here limitations encountered with the method.

1.1 Translation in bacteria

Translation represents the final stage of gene expression, where the coding sequence of mRNA is converted into the amino acid sequence of a protein [4]. It is a highly conserved and energy-demanding process. In bacteria, the production and upkeep of ribosomes and associated translational machinery consume over 40 % of the cellular energy [5].

The main components of translation are codons on messenger RNA (mRNA), transfer RNA (tRNA), and ribosomes. Codons consist of three nucleotides, allowing for 64 possible combinations. Since there are only 20 “canonical” or “natural” amino acids, several are encoded by multiple codons, a phenomenon referred to as genetic code degeneracy. Of the 64 codons, 61 encode specific amino acids, while the remaining three act as termination signals for translation. tRNA serves as a transfer molecule, linking nucleic acid sequence information to amino acids. Each tRNA molecule is chemically bound to a specific amino acid and contains a trinucleotide anticodon complementary to the corresponding codon on mRNA. Ribosomes, along with translational factors, provide the environment necessary for protein synthesis by facilitating interactions between mRNA codons and tRNA anticodons. Composed of two subunits made of proteins and ribosomal RNA (rRNA), ribosomes possess peptidyl transferase activity, enabling the formation of peptide bonds between adjacent amino acids during translation.

Translation process involves four key phases: initiation, elongation, termination, and ribosome recycling [4]. Initiation begins translation and includes events prior to the first peptide bond formation. It starts with the recognition of the start codon, AUG, which encodes methionine, the first amino acid in all proteins. The initiation complex forms when a ribosome binds to mRNA and pairs with a methionine-carrying tRNA containing the complementary anticodon. Elongation is the stepwise addition of amino acids to the growing polypeptide chain, guided by codon-anticodon interactions as the ribosome moves along the mRNA. Elongation factors assist in tRNA entry, and efficiency is enhanced by polysomes, where multiple ribosomes translate the same mRNA simultaneously. Termination occurs when a stop codon (UAG, UAA, or UGA) signals the release of the completed polypeptide chain. The ribosome then dissociates from the mRNA. [6]. Ribosome recycling is the process of disassembling post-termination ribosomes to prepare them for reuse in translation. After termination, the ribosome releases the mRNA and tRNA, followed by the

splitting of the ribosomal subunits [4]. Overview of the translation process along with the main components of translation is shown in Figure 1.

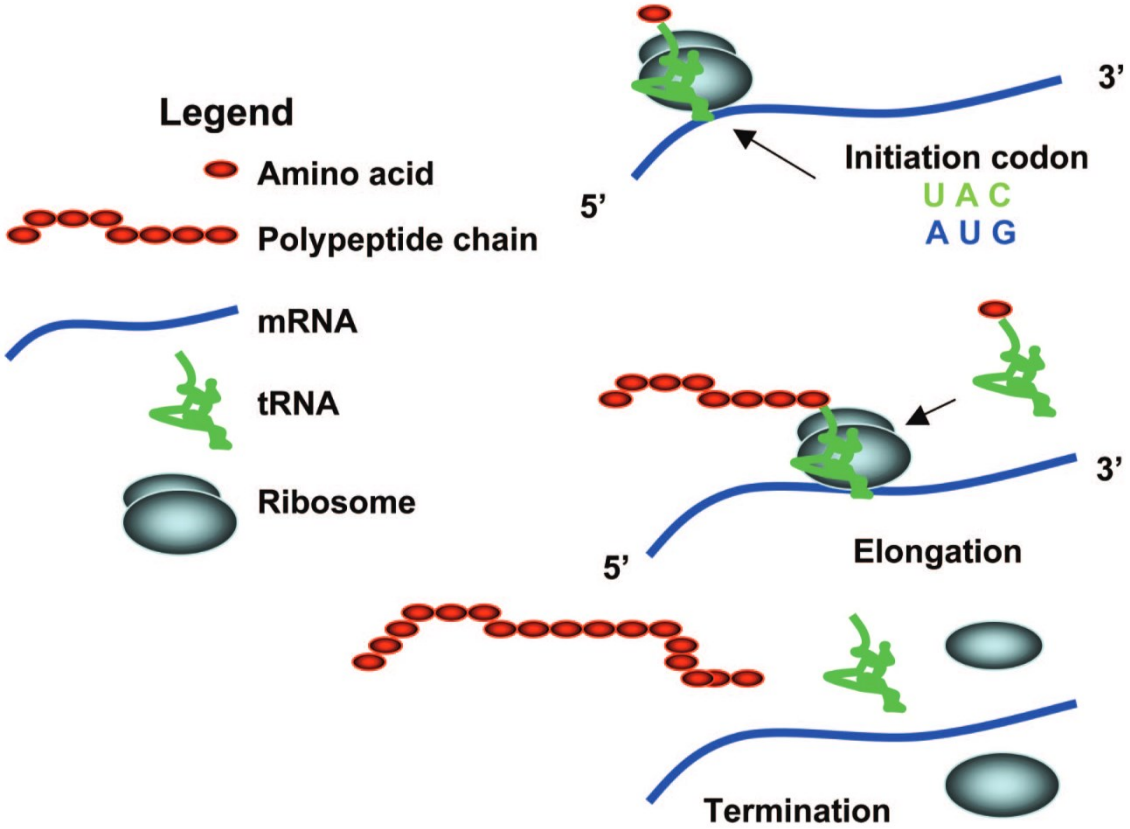


Figure 1. Schematic overview of the translation process. The process begins with initiation (top), where the ribosome complex binds to the mRNA, allowing the entry of a tRNA carrying the anticodon (UAC) that pairs with the start codon (AUG). Next, during elongation (middle), the ribosome moves along the mRNA in the 5' to 3' direction, facilitating the sequential addition of amino acids to the growing polypeptide chain through codon-anticodon interactions. Finally, in termination (bottom), translation halts upon reaching a stop codon (UAG, UAA, or UGA), leading to ribosome disassembly and release of the completed polypeptide. The ribosome complex can then be recycled for subsequent rounds of translation. This figure was adapted from [6].

In biotechnology, translation is a critical biological process that is harnessed in recombinant protein production. This involves utilizing the host organism's translational machinery to synthesize desired proteins from foreign genes.

1.2 *Escherichia coli* as a host for protein expression

Recombinant protein production in *E. coli* revolutionized biotechnological research and made its way into industrial applications. Discoveries alongside recombinant protein production in *E. coli*

set a number of milestones in the field of molecular biology. The inception of recombinant protein production in *E. coli* traces back to the 1970s, marked by Paul Berg's creation of the first recombinant DNA molecule in 1972 [7], earning him a share of the 1980 Nobel Prize in Chemistry for his fundamental studies of the biochemistry of nucleic acids, with particular regard to recombinant DNA [8]. One of the early landmarks is the pioneering synthesis of somatostatin fused to β-galactosidase, the first ever recombinant protein prepared in *E. coli* [9], followed by the production of the human growth hormone, the first product in its non-precursor form produced in *E. coli* host in 1979 [10]. Importance of recombinant protein preparation and usage of *E. coli* as a host has skyrocketed afterwards. This bacterium suddenly became an important powerhouse for synthetizing important molecules, for example biopharmaceuticals such as insulin [11, 12].

E. coli production platform facilitates the efficient expression of diverse proteins, driving advancements in medicine, biotechnology, and beyond. Heterologous gene expression in *E. coli* is a key step for lot of protein-based research. Some of the main advantages of this expression system are rapid cell culture growth, easy gene expression, cost-effectivity, commercial availability, and low equipment requirement. Advantages and disadvantages of *E. coli* as a host for protein expression were discussed in numerous reviews [13-25].

1.3 Optimization of protein production by manipulating the N-terminal sequence

Since this process of translation is so massively used in biotechnologies and research, it has been the object of many optimization experiments targeting several sub-processes. Translation is regulated by several different mechanisms, mostly taking place during the initiation phase. Additionally, several mechanisms that regulate the pace of translation focus on the elongation as well as the termination stage [26].

One of the primary focuses of this study is the sequence composition of the N-terminal region. Numerous studies have demonstrated the importance of the codons immediately following the start codon as a regulatory mechanism in translation [27-36]. The mechanisms driving differences in protein yield following N-terminal sequence manipulation remain an active area of investigation. Sequence composition can influence processes as early as DNA-to-mRNA transcription, although this effect is not consistently observed [30]. The expression regulation could also appear during

later stages, for example by interactions between the nascent peptide and the ribosome. Although several studies suggest potential mechanisms, the overall process remains incompletely understood [37].

1.3.1 Importance of the N-terminus composition in protein scaffolds

Prior research in our lab on small non-antibody protein scaffolds designed to bind interleukin 10 (IL10) revealed the crucial role of N-terminal sequence composition in protein production and function. These scaffolds represent a promising class of artificial high affinity binders, offering potential advantages over traditional antibodies in terms of stability, specificity, and manufacturability [38, 39].

1.3.1.1 Antibodies

Antibodies are highly selective and adaptable molecules that can bind to a wide range of targets. They serve as natural biomolecular scaffolds, with highly variable antigen binding sites evolved to recognize diverse foreign substances and pathogens. Beyond their *in vivo* role, both natural and engineered antibodies are essential tools in research, with over 10 000 different molecules commercially available. Additionally, several antibody-based drugs are approved, primarily for cancer and autoimmune treatments, with many more undergoing clinical trials [40]. Antibody based binders are essential in research and clinical trials, with many of them being Federal Drug Administration (FDA) approved biologicals. Several characteristics of antibodies, such as their large size, cross reactivity, and the need for animal immunization during preparation, have driven the development of binders with alternative structures [41, 42].

In the adaptive immune response, three types of molecules can bind antigens: immunoglobulins (antibodies), T cell receptors (TCR), and major histocompatibility complex (MHC) proteins. Among these, antibodies exhibit the broadest antigen recognition capacity, superior discrimination between related antigens, and generally the highest binding strength. Unlike TCRs and MHC molecules, which recognize only linear epitopes, antibodies can bind both linear and conformational epitopes across a diverse range of molecules, including proteins, lipids, polysaccharides, and small chemicals [43].

Produced by B-cells, antibodies play a crucial role in immune defense. Also known as immunoglobulins, they are Y-shaped glycoprotein molecules composed of two functional regions—the light and heavy chains. Immunoglobulins are not restricted to plasma but are also present in various body fluids and tissues, including urine, cerebrospinal fluid, milk, saliva, tears, lymph nodes, and the spleen. They are classified into five types—IgM, IgD, IgG, IgA, and IgE—based on their heavy chain composition in constant regions, while their variable regions determine antigen specificity [42] [44]. Structure of an antibody molecule is shown in Figure 2.

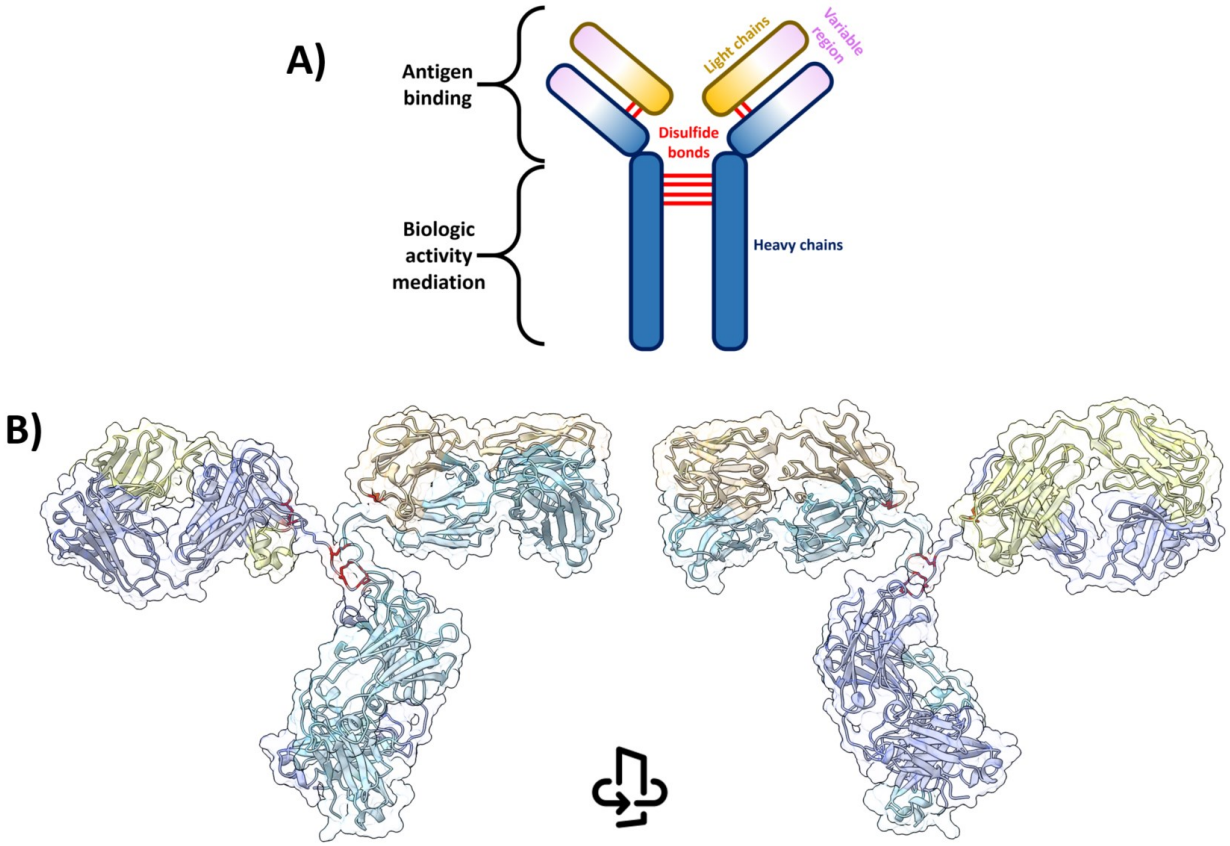


Figure 2 Antibody structure. A) Simplified schematic of an antibody molecule. The Y-shaped antibody consists of four polypeptide chains — two heavy (blue) and two light (orange) — linked by disulfide bonds (red). The upper part, including the variable region (pink), binds antigens, while the lower part mediates biological activity. B) Immunoglobulin structure (IgG) visualized using ChimeraX software [45, 46] and protein data bank (PDB) structure 1IGT [47]. The two heavy chains are depicted in different shades of blue, while the light chains are shown in varying shades of yellow. Disulfide bonds, which connect the heavy chains to each other and link the light chains to the heavy chains, are highlighted in red.

In biotechnology, the complicated composition of antibodies leads to unstable domain association and complicates cloning procedure for recombinant expression. The constant region of antibodies

drives immune functions but can cause unwanted interactions in some biopharmaceutical applications. Producing full size, post translation modifications (PTMs) containing antibodies is costly and slow, requiring complex eukaryotic cell cultures with limited production capacity. These disadvantages create need for an antibody alternative, which could be protein scaffolds [40].

1.3.1.2 Protein scaffolds

The term "protein scaffold" refers to the ability of certain molecules to form stable, natural protein structures that can be modified to alter an existing binding site or create a new one for a specific target. Originally, the term "scaffold" described a mechanical structure that supports workers and equipment during construction, repair, or maintenance. Similarly, in molecular biology, protein scaffolds provide a structural framework that can be engineered for various functional applications. Since *de novo* protein design has struggled to produce practically useful biomolecules due to the unresolved protein folding problem, researchers have instead focused on modifying structurally stable polypeptide frameworks. By reshaping less critical regions of these proteins, they aim to introduce novel functions while preserving structural integrity [48].

An ideal protein scaffold should have a rigid folding unit that brings together multiple exposed loops to form an extended interface for strong target binding. Ideally, small globular proteins are preferred due to their stability and well-defined structure. However, since protein folding is highly cooperative, a suitable scaffold must tolerate a certain number of amino acid substitutions in its loop regions without losing its stable tertiary structure. While secondary structure packing primarily determines overall protein topology, surface loops and turns also play a crucial role in maintaining the stability of the folded state. An optimal protein scaffold should have structurally separated stability and overall conformation of the polypeptide from the local shape and molecular recognition function of its active site, allowing modifications without disrupting the integrity of the protein structure. Additionally, two key criteria define an applicable scaffold. First, it should possess a well-defined hydrophobic core that is structurally superimposable among members of its protein family and significantly contributes to the free energy of folding, and second, it should feature a solvent-accessible active site or binding pocket that is spatially distinct from the core and ideally involved in recognizing different target molecules [48].

Numerous small protein scaffolds have already shown utility in diverse applications, such as academic research, clinical imaging, diagnostics, and therapeutics, with some currently advancing through clinical trials [49-51]. Here, scaffold's applicability was evaluated by generating high-affinity variants for a target protein using *in vitro* directed evolution. Directed evolution is a protein engineering technique that emulates natural selection to optimize proteins or nucleic acids for a specific, user-defined purpose [52].

1.3.1.2.1 Protein scaffolds vs antibodies

In contrast to antibodies, small non-antibody protein scaffolds offer several advantages. Often derived from well characterized proteins as separated protein domains or from a robust and small soluble monomeric protein, these scaffolds are generally smaller than antibodies. This enhances tissue penetration, which is indispensable attribute for clinical applications. Additionally, they are typically more stable and less costly to produce, as they do not require PTMs and have simpler structures, making them suitable for efficient production in *E. coli*. This streamlines production, lowers costs, and eliminates the need for animal immunization, which supports the ethical considerations in scientific research [53].

1.3.1.3 IL10 as a target

As a target for the protein scaffold research, IL10 was chosen. IL10 is a medically significant cytokine, which was first identified over 30 years ago as a cytokine synthesis inhibitory factor, secreted by T helper (Th) 2 cell clones, which suppress cytokine production by Th1 cells [54]. Later the production of this cytokine by other CD4 and also CD8 T cells was soon demonstrated. Shortly after IL10 was discovered, its diverse functions were identified in mice. In addition to inhibiting cytokine production, IL10 was found to promote mast cell and thymocyte growth and activate B cells [55].

IL10 belongs along with IL19, IL20, IL22, IL24, IL26, IL28A, IL28B, and IL29, into IL10 cytokine family. The main features that group these cytokines into this family are shared common receptor usage, downstream signaling, and high similarity in secondary structure, represented by six bundled alpha helixes [56]. Structure of IL10 is shown in Figure 3.

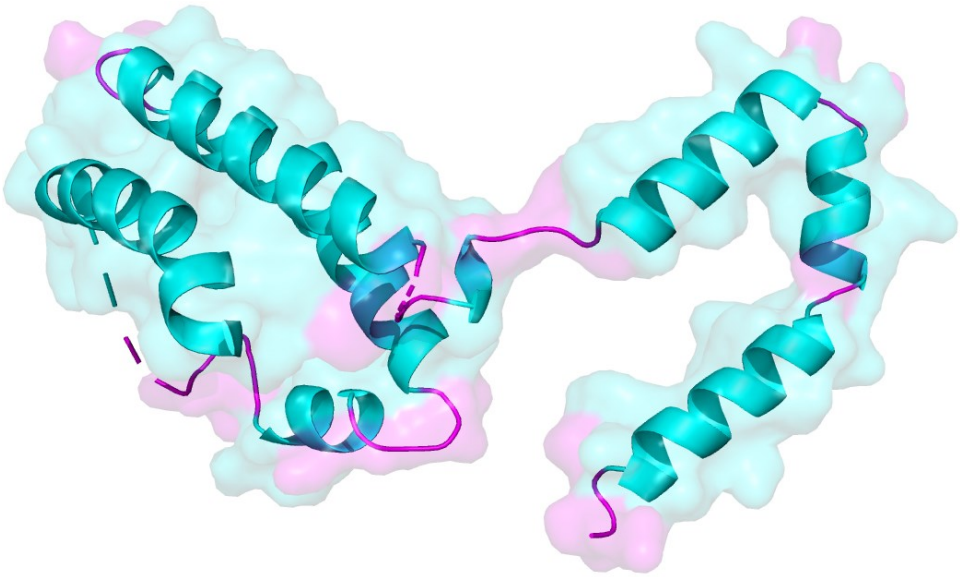


Figure 3. IL10 (single monomeric subunit) structure visualized using PyMOL software [57] and PDB structure 1INR [58]. Alpha helixes are shown in cyan and loops in purple.

IL-10 forms a homodimer and binds to two high-affinity IL10R1 receptors. This complex then recruits two IL10R2 receptors, forming a tetrameric receptor complex. Signal transduction occurs through the Janus kinase (JAK)-Signal transducer and activator of transcription (STAT) pathway, commonly known as the JAK-STAT pathway. IL10 homodimer binding to the receptor complex activates the receptor associated kinases JAK1 and Tyk2, leading to STAT3 phosphorylation and dimerization [59].

IL10 plays a key role in regulating immune responses by inhibiting cytokine production and reducing the antigen presenting capacity of macrophages and dendritic cells. It downregulates costimulatory molecules such as major histocompatibility complex Class II (commonly referred to as MHC Class II), modulates chemokine secretion, and helps control T cell activation. IL10 also affects the immune environment by limiting the effects of cytokines IFNγ, IL2, tumor necrosis factor α, and IL4, while enhancing the survival of B and T cells and activating natural killer cells. These properties establish IL10 as a key mediator of the anti-inflammatory response. Indeed, IL10 deficiency in mice leads to spontaneous inflammation, exaggerated immune responses, and greater resistance to certain infections, underscoring the role of IL10 in immune balance and prevention of autoimmune like conditions [59].

The identification of IL10 homologs in viral genomes provided evidence that viruses may exploit IL10's immunosuppressive properties to evade the host immune response. Several studies have shown that these viral IL10 homologs can mimic the activity of host IL10 by inhibiting cytokine production and modulating immune cell function. Since this discovery, multiple viruses have been found to encode IL10-like proteins, highlighting a common strategy used by pathogens to suppress immune activation and promote persistence within the host [55]. Overexpression of IL10 is connected with different types of cancer [60, 61]. Therefore, IL10 expression is tightly regulated to maintain immune balance and prevent diseases associated with its excess or deficiency. This regulation involves both general and cell specific mechanisms, including the activation of specific signaling pathways, the expression and activation of key transcription factors, and post transcriptional and epigenetic modifications, all of which ensure the controlled production of IL10 [55].

Genetic studies have further linked impaired IL10 responses to Inflammatory Bowel Disease (IBD) susceptibility in humans. While systemic administration of recombinant IL10 in clinical trials has not significantly improved patient outcomes, alternative strategies have been explored. One promising approach involves engineering bacteria, such as *Lactococcus lactis* and *Bifidobacterium bifidum*, to produce IL10 for targeted intestinal delivery. Preclinical studies in mouse models have demonstrated the efficacy of this method in reducing colitis symptoms, and early trials in Crohn's disease patients have shown potential benefits. Further refinements in bacterial based IL10 therapy could offer a viable treatment option for IBD [55].

IL10 has been seen as immunosuppressive, promoting tumor immune evasion. Studies have shown that IL10 overexpression impairs tumor control, while blocking its receptor enhances chemotherapy efficacy. High IL10 levels in tumors are often linked to poor prognosis. However, recent results suggest that IL10 can have antitumor effects by increasing CD8 T cell infiltration, boosting IFNγ production, and enhancing T cell memory. Pegylated IL10 has shown promise as an immunotherapy, stimulating CD8 T cell immunity in patients with solid tumors. Despite these benefits, IL10 may also contribute to immune exhaustion, particularly through regulatory T cells. The varying effects of IL10 depend on factors such as the immune cell targets and T cell response stages, highlighting the need for careful consideration of its role in cancer treatment [55].

IL10 plays a crucial role in immune regulation and is closely linked to rheumatoid arthritis (RA). Patients with RA have fewer IL10 producing regulatory B cells, which also exhibit a reduced ability to suppress inflammatory IFNγ production in CD4+ T cells. This deficiency contributes to persistent immune activation, making IL10 and regulatory B cells potential therapeutic targets in RA [62]. Another cytokine with anti-inflammatory properties in RA is interferon beta (IFNβ), which was one of the target proteins for my study presented in this thesis. IFNβ helps regulate inflammation by reducing key pro-inflammatory cytokines such as TNFα, IL1β, and IL6 while promoting anti-inflammatory mediators like IL1Ra and IL10. In animal models, IFNβ treatment has shown promising results, decreasing synovial inflammation and protecting against joint damage. However, clinical trials in RA patients have been unsuccessful, likely due to pharmacokinetic challenges. Targeted delivery strategies may be necessary to harness therapeutic potential of IFNβ in RA [63].

IL10 was chosen as a target protein for the protein scaffold research for its medical relevance. Differences in protein production were observed when making sequence changes at the N- and C-terminus during this process. This unexpected finding led me to further investigate how sequence modifications influence protein yield.

1.3.2 mRNA and its impact on the yield of the protein product

Beyond encoding proteins, mRNA contains regulatory elements that influence protein synthesis at multiple stages [64]. It is also one of the molecules considered to have control over the speed of translation. The first regulatory step is the DNA transcription into mRNA. This process has a direct impact on the level of mRNA. Increasing the mRNA level could lead to a higher yield of the protein after translation. This raises the question of whether modifying the N-terminal sequence could enhance mRNA levels and subsequently improve recombinant protein expression. It has been reported that the mRNA levels of the enhanced protein were not higher compared to the wild type (WT) variant [36], moreover it has been shown, that they were notably lower [30], suggesting that mRNA level is not responsible for the product enrichment.

Another attribute of mRNA that can perhaps alter the protein yield is its secondary structure. The formation of mRNA secondary structures such as hairpins and pseudoknots could affect the

progress of the ribosome [65, 66]. The occurrence of mRNA structure regulation is well documented in multiple cases.

For example, the incorporation of selenocysteine relies on the mRNA stem loop structure formed at the selenocysteine insertion sequence (SECIS). Then a complex of selenocystenyl-tRNA, GTP and a specialized elongation factor (SELB in bacteria) binds to the mRNA stem loop. This interaction enables delivery of selenocystenyl-tRNA to the ribosomal A site in time, resulting in recoding of the UGA stop codon to UGA-selenocysteine codon. This mechanism of selenocysteine insertion is an excellent example of how an mRNA structure and its interaction with an RNA-binding protein can alter decoding [67]. This process is illustrated in Figure 4.

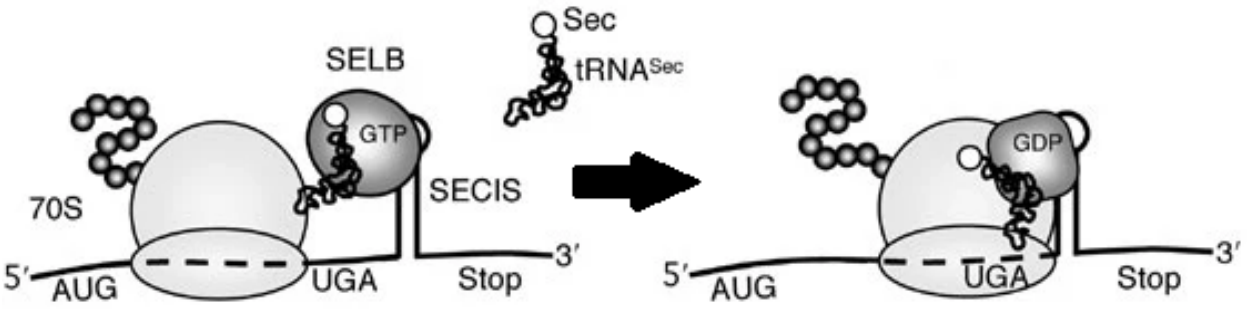


Figure 4. Selenocysteine incorporation. SELB binds to the SECIS element, Sec-tRNASec, and GTP, forming a complex downstream of the UGA-selenocysteine codon. The binding of SELB to the SECIS element triggers a conformational change in SELB, depicted as a shape alteration, facilitating the precise delivery of Sec-tRNASec to the ribosomal A site at the position of the UGA-selenocysteine codon. This figure was edited and taken from [68].

During translation, the mRNA structure needs to be unfolded before mRNA is decoded by ribosomes. Single-stranded mRNA passes through a five to six nucleotide long entrance channel formed by three ribosomal proteins (called S3, S4, and S5 in bacteria) [69, 70]. This narrow mRNA entrance channel prevents RNA duplexes from entering. Consequently, structured mRNA must unfold around two codons before reaching the ribosome A site, where the decoding takes place [71]. Unfolding requires energy, likely provided by elongation factor G (EF-G), which hydrolyzes GTP for translocation. Ribosome translocation on mRNA may involve a conformational change in EF-G that primes the mRNA tRNA complex for translocation [72].

This mechanism of translation speed alteration plays its role at the ribosome binding site (RBS) and its close regions as well as at the beginning of the open reading frame [35, 73-76]. Indeed, it has been shown by Katz *et al.*, that there is a trend across various eubacteria wherein the content of mRNA secondary structures decreased towards the 5' end of the analyzed genes [77]. This is in concordance with findings not only in *E. coli* host, but in other organisms, even in dsDNA viruses [74, 78, 79].

The stability of mRNA structures is usually estimated using nearest-neighbor minimum-free energy model. The nearest-neighbor model predicts nucleic acid stability by considering how each base pair is influenced by its neighboring base pairs. Stability depends on the sequence and orientation of these pairs. The model includes parameters for duplex initiation, accounting for effects like strand ends and ion interactions. It distinguishes between duplexes with terminal G/C and A/T pairs and adds an adjustment for symmetrical sequences. Overall, the nearest-neighbor model combines sequence-specific and general factors to estimate the stability of DNA or RNA duplexes [80]. Applying this calculation on a sequence near Shine-Dalgarno sequence and start codon shows an inverse relationship with protein expression [81, 82]. The specific amino acid sequence can be maintained while changing the nucleotide sequence and codon usage, which is often the target of optimization experiments [83-85].

By targeting the N-terminal part of the sequence specifically, the sequence composition can be optimized using the TISIGNER tool. This bioinformatics tool is designed to minimize the probability of mRNA secondary structure formation with a goal to design optimal initial codons of the gene, while keeping the same primary protein structure [86]. In contrast, Looman *et al*. systematically analyzed the effect of all possible codons following the start codon on recombinant protein yield, reporting up to a 15-fold disparity in expression levels [87].

Charneski *et al.* also investigated structure of mRNA as a potential key regulatory element of the translation. In their study, mRNA structure was examined in comparison with other potential regulators, mainly properties of amino acids of the nascent peptide, which are discussed later in this thesis. To elucidate the contribution of the mRNA structure to the translation speed regulation, authors used empirically determined (rather than computationally predicted) mRNA structure data of protein coding sequences from a *Sacharomyces cerevisiae* (*S. cerevisiae)*. Sections of the gene

showing single stranded conformation followed immediately by a double stranded stretch were chosen for this experiment to elucidate how the translation speeds changes in proximity of these regions. To estimate the translation rate, relative ribosomal density was calculated. Authors also claim that they investigated not only mRNA stretches showing higher probability of secondary structure formation, but also sections of the transcript before the structured parts, since the ribosomes can theoretically stall there as well. Ribosome pausing was confirmed to occur in both cases. However, the authors concluded that the impact of mRNA secondary structure on the translation rate is relatively marginal. They observed that presence of the mRNA structure alone could not consistently explain the occurrence of the slowest translated regions, even in the slowest and fastest translated regions [88].

1.3.3 Nascent protein elements and their impact on the recombinant protein yield

The amino acid composition of the nascent protein is another factor that can influence translation speed. However, the role of specific amino acids within the growing polypeptide as regulators of translation speed has been relatively underexplored [88].

Bivona *et al.* have identified that the crucial component is specifically the second amino acid of the nascent protein. The outcome of this investigation remains consistent regardless of the specific codon used to represent the amino acid. Employing various *E. coli* strains with a pET vector, Bivona *et al*. altered the second amino acid of the extracellular domain of the immunoglobulin α (Igα) using all 20 natural amino acids for screening. The findings indicated that the amino acids at the second position exerted a 10-fold influence on recombinant protein expression [27]. The importance of the second amino acid position has been found crucial before, but the authors determined the second nucleotide triplet as the crucial factor instead of the incorporated amino acid [89]. Voges *et al.* conducted a study focusing primarily on the mRNA and its influence on translation efficiency. Their findings also revealed that presence of a specific amino acid following the start methionine altered expressed protein yield. Specifically, they observed that constructs with high levels of expression tended to have serine, threonine, asparagine, or histidine as the second amino acid. In contrast, constructs with lower expression levels more commonly featured alanine, valine, glycine, cysteine, proline, or arginine in the second position [90].

Some findings indicate that the key property of the amino acid in the nascent protein affecting the speed of translation could be the charge of amino acids passing through an exit tunnel of a ribosome [91, 92]. The ribosomal exit tunnel, approximately 10 nm long and 1.5 nm wide, connects the peptidyl transferase center to the solvent environment and can hold 30–40 residues of an extended α-helical polypeptide. Its electrostatic potential is not uniform, varying in magnitude along the tunnel's length, but it is generally negative. The exit tunnel is closely associated with the nascent protein and plays an active role in the translation process. A schematic view of the ribosome and its exit tunnel is shown in Figure 5.

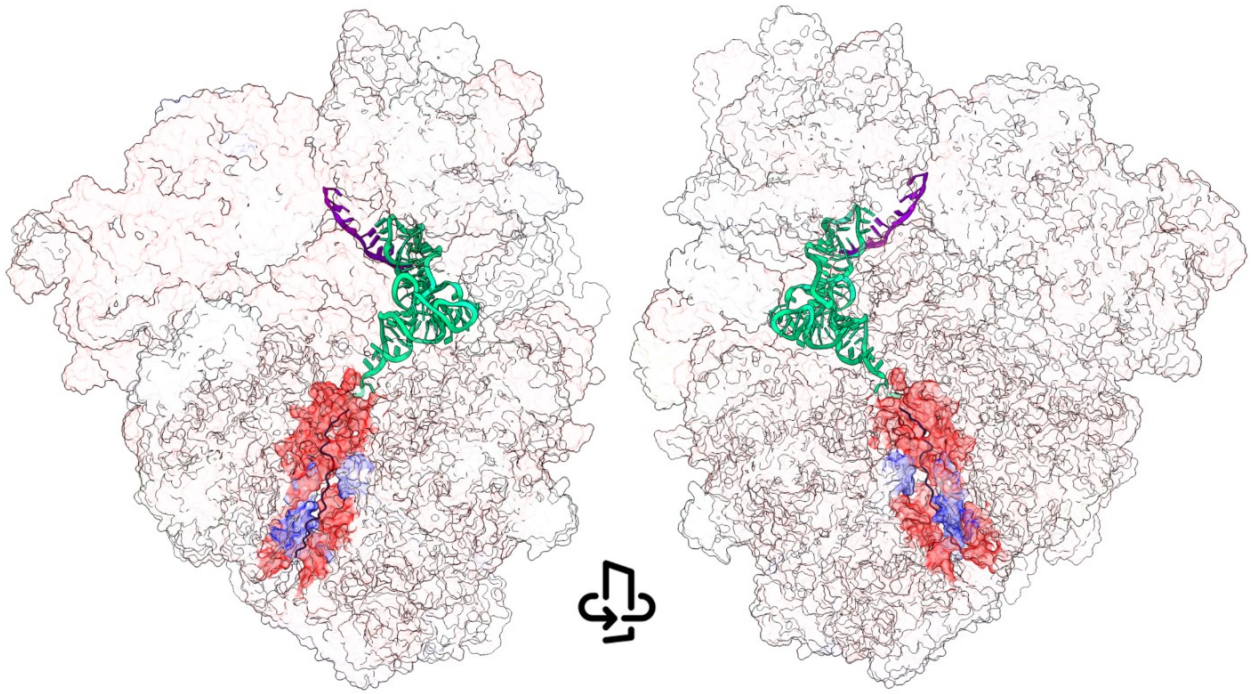


Figure 5. Schematic representation of the ribosome and its exit tunnel, depicted from two opposite perspectives created using ChimeraX software [45, 46] and a PDB ID structure 3JBU [93]. The Gly-tRNA molecule is shown in Medium Spring Green, while the mRNA is depicted in Dark Violet. The coloration of the ribosome's exit tunnel is based on the charge of interacting molecules: red indicates negative charge and blue indicates positive charge. The overall charge is predominantly negative due to the abundance of negatively charged rRNA within the ribosome structure. Positive charge regions are attributed to ribosomal proteins that interact closely with the exit tunnel, such as uL4, uL22, uL23, and uL24. The nascent protein traversing the ribosome exit tunnel is shown as a Midnight Blue line.

An electrostatic potential in the tunnel can have a range of effects. It can affect the concentration of ions, promote and stabilize formation of helical structures, alter the orientation and pKa of charged side chains of both nascent peptides and ribosomal proteins that are facing the tunnel, and the organization of polar molecules such as water. These consequences might influence the rates of elongation phase of translation. Lu *et al.* showed, that positively charged residues such as arginine or lysine cause transient arrest (pausing) of the ribosome, before the nascent protein is fully synthetized [91, 92, 94, 95]. If the charge of the newly formed protein plays a key role in ribosomal translation efficiency, the distribution of charges throughout a protein sequence could be influenced not just by structural and functional needs, but also by their impact on the translation process. It has already been shown that there is a greater presence of positively charged residues at both N- and C-terminus of proteins [96, 97]. Tuller *et al.* suggested that the so-called ribosomal ramp is an inherent component of the translation initiation and is important for avoiding ribosomal jams [98]. This translation control mechanism is similar to the regulation achieved using rare codons, as discussed later.

The cyclic amino acid proline stands out for its unique structure. The incorporation of proline into a nascent peptide chain presents unique challenges due to its distinct structure. While this structure plays a crucial role in facilitating protein folding, it significantly impedes the rate of peptide bond formation during translation. Compared to other amino acids, proline is incorporated into nascent peptides at much slower rates, which can result in ribosome stalling. The cyclic structure of proline hinders its proper positioning in the ribosome's amino acid binding pocket and disrupts the alignment of the nascent peptide chain in the ribosomal exit tunnel. As a result, proline binds in a different way compared to the other amino acids, further contributing to the difficulty of its incorporation into growing peptide chains [99].

Charnetski *et al.*, who previously ruled out mRNA structure as the key factor in translation control, also presumed that the properties of specific amino acids could play a crucial role in the regulation of the translation rate. They examined the impact of positively charged residues (lysine, arginine, and histidine) in the nascent protein. These residues were expected to cause ribosome stalling after their incorporation into the nascent peptide, likely due to interactions between the charged amino acids and the ribosome's charged exit tunnel. Authors claim, that a single positive charge could

slow the ribosome regardless of specific codon composition. They also show that histidine produces weaker slowing of the ribosomes compared to arginine or lysine, perhaps because the pKa of histidine and hence overall net charge is in physiological conditions lower compared to other positively charged amino acids. Authors conclude, that this increase in ribosomal occupancy cannot be explained by physiochemical properties of other amino acids, like hydrophobicity, or polarity. They also imply that a positive charge is more often associated with a higher occupancy window within each transcript than the other considered factors [88]. Requiao *et al.* analyzed 551 705 protein sequences from the UniProt Swiss-Prot database and found a higher prevalence of positively charged amino acids within the first 30 residues of a protein. The 30 amino acid span was chosen, because it corresponds to the amount of amino acids that could fit within the ribosome exit tunnel. Their analysis also shows and enrichment of positively charged residues at the C-terminus, even though the difference in charge is not as significant as on the N-terminus [100].

Charneski *et al.* postulated that the enrichment of positively charged residues at both ends of proteins could be due to the presence of positive regions near the transmembrane protein segments. These basic areas, located on the cytoplasmic side of the membrane, control protein orientation through electrostatic interactions with the inner phospholipid layer. By examining the first thirty amino acids of *E. coli* and *S. cerevisiae* proteins, they discovered that membrane proteins, but not cytoplasmic proteins, contain basic N-terminus. They also showed that the specific arrangement of arginine or lysine at the N- and C-terminus may result from a natural inclination for positively charged regions to cluster near the ends of proteins. Overall, these findings suggest there is no clear consensus on whether translation efficiency affects the distribution of charges in the primary protein structure [101].

1.3.4 Codon usage and its impact on the yield of the protein product

The process by which codons are redefined as a result of dynamic interactions with adjacent mRNA elements is known as recoding. This phenomenon is widespread across all domains of life and is notably prevalent in viruses [64]. Positive correlation between the amount of tRNA and the occurrence of respective codons was previously found for most *E. coli* genes, showing the importance of specific codon usage [102, 103]. It is hypothesized that codons corresponding to less frequent tRNAs are translated at reduced rates due to the ribosome pausing until the required

tRNA becomes available. This phenomenon is thought to account for the preferential usage of codons recognized by the most abundant tRNAs in highly expressed genes [104].

The situation and usage of codons at the N-terminus of the nascent protein is quite different. Several studies suggest that the recombinant protein yield can be influenced by codon usage at the N-terminus of the produced protein [105, 106]. This is consistent with the previous results showing that rare codons such as AGG, the least used codon in *E. coli*, appear much more frequently in the first 25 codons [32, 107], even though overall codon usage usually correlates with the abundance of tRNAs [102, 108]. The bias in codon usage was experimentally shown on the example of lacZ protein and streptokinase, which has a relatively high number of rare codons in its sequence (30 %). The relatively rare AGG triplet coding arginine, was implemented in different positions, resulting in overexpression, when placed into the second position (again emphasizing the importance of this position as discussed previously [27]). Whenever such rare codons were placed further in the gene, it limited or prevented the expression completely [32, 109]. It has also been shown that placing more common arginine codons in the second and third position was less effective at supporting gene expression compared to the rarer codons [109]. This aligns with prior research indicating that the strategic positioning of the rare codons like AGG or AGA, which are linked to low gene expression within a gene, significantly influences protein production. [107, 110].

It is also worth noting that implementing a rare codon into the second position is not always beneficial for the protein production. Pai *et al.* further determined that the presence of repetitive CAT codons, which encode HisTag, following an AGG triplet facilitated the production of human erythropoietin production in *E. coli* [111].

A plausible explanation for this phenomenon could be that the rare codon serves as a retarder for ribosomes translating the mRNA. The mRNA is simultaneously translated by multiple ribosomes and the ribosomes could interfere with each other. This may even lead to their dissociation from the template. Since different codons are translated at different rates, those which are uncommon could provide a spacing between the ribosomes when placed at the beginning of the translated gene by delaying them and thus prevent the ribosome collisions [112].

1.3.5 Optimizing the N-terminal sequence in order to increase the yield of the protein product

Despite the uncertainty surrounding the precise mechanism and molecules involved in regulating protein synthesis based on the N-terminus sequence of nascent protein, existing research is at least uniform in referring to the crucial role played by the composition of this region in protein production. The regulation is most likely affected by multiple factors, which are not easily distinguishable, and probably differs for individual proteins. The uncertainty about which specific mechanism influences recombinant protein production in relation to its N-terminal sequence has been expressed in the literature by Verma *et al.* [33]. Similarly, Cambray *et al.* suggest that several mechanisms remain undiscovered, further complicating the understanding of this phenomenon [37].

Even proving specific mechanisms in specific cases can be a challenging task, since for example rare codons are generally rich in A+T content, leading to reduced secondary structures in mRNAs and likely enhancing initiation rates. This makes it hard to decide which of the mentioned mechanisms takes the main place during the regulation [28, 37]. As with any other research, it is important to avoid potential misinterpretation of the results. For example, the preference for N-terminal positive charge, observed almost exclusively in weakly expressed membrane proteins. While this sequence trait promotes association with the cytosolic membrane leaflet rather than ribosomal slowing, it is linked to both effects [97, 101].

Even though the prospects for elucidating the molecular mechanism are not optimistic, the recombinant protein yield can still be manipulated by optimization of the N-terminal sequence. One of the approaches, taken by Ojima-Kato *et al.* was to implement a short N-terminal protein sequence coding a small number of amino acids common to the N-termini of highly expressed proteins in *E. coli*. The N-terminal sequence MSKIK (representing amino acid sequence methionine, serine, lysine, isoleucine and lysine) was selected by this approach and proven to enhance the protein yield. Authors claim in their publication, that the specific mechanism underlying the enhancement is not clear. It is worth to highlighting that this sequence included the uncommon ATA codon. Additionally, the authors noted that mRNA levels of proteins tagged with the MSKIK sequence were 4-5 times lower compared to a construct with a WT N-terminal sequence [30]. Six years later the same authors proposed that their MSKIK sequence prevents or

releases ribosomal stalling resulting in higher protein yield, while they also ruled out the influence of codon usage [113].

1.3.6 Consequences of altered translation speed on protein folding

Both slowed and accelerated translation rates affect protein folding and accelerated translation may lead to misfolded product. Komar *et al.* demonstrated this phenomenon on a chloramphenicol acetyltransferase by introducing silent mutations using frequent codons to replace those of low frequency to alter the speed of translation at specific regions of the gene. Results of this study have shown that codon replacement for more common variants abolished particular translational pauses, resulting in a different speed of translation across the gene. Outcomes of this study show a reduced specific activity in comparison with that of the WT protein. The overall difference in specific activity (ratio of activity over protein amount) was found to be 20 %, which suggests that a replacement of certain synonymous codons could lead to the alteration of ribosome traffic on mRNA at a selected place [114], which is in concordance with the previous findings by Young *et al*. [115].

1.3.7 Reporter molecule for directed evolution based selection of N-terminal sequences

In this study, I applied an optimization strategy to refine the N-terminal sequence composition. To achieve this, I cloned a short DNA library immediately after the start codon of each construct.

To evaluate translation efficiency, I introduced a fluorescent reporter molecule, cycle 3 GFP (c3GFP), at the C-terminus of each gene of interest. Reporter molecules serve as detectable biomarkers that help track the presence, quantity, or activity of a target process - in this case, translation efficiency. The structure of this GFP mutant is shown in Figure 6.

c3GFP was chosen as a reporter molecule for all constructs discussed in this thesis. c3GFP is a 26.9 kDa β-barrel protein derived from GFP. Excitation maximum of the mutant is reached at 520 nm and emission maximum is at 650 nm. Compared to a WT GFP molecule the c3GFP mutant has three point mutations that are responsible for increased fluorescence (F99S, M153T, and V163A). These mutations also facilitate proper maturation of the protein at 37° C and the molecule shows lower tendency to aggregate [116, 117]. The c3GFP mutant originated from a DNA

shuffling experiment done by Crameri *et al.* [116]. I commonly refer to this molecule as "GFP" in this study.

Figure 6. Structure of the cycle 3 GFP made with Swiss-model using WT GFP as a template (sequence identity 97.90 %) [118] and visualized using PyMOL software [57]. Three mutations (F99S, M153T, and V163A), that make the mutant different from the WT are shown in red.

With this setup, increased fluorescence corresponds to increased production of the protein of interest in each cell caused by the N-terminal library.

1.3.8 Introduction to FACS

FACS is an advanced technique built upon the principles of flow cytometry, a powerful tool for the multiparametric analysis of particles, most commonly live cells, suspended in a solution. Flow cytometry was initially limited to detecting just one parameter—the size of the particles being analyzed. However, modern flow cytometers are much more sophisticated and can measure multiple characteristics of a single cell as it passes through the instrument, such as size, granularity, and fluorescence. These measurements are based on light scattering and the fluorescence properties of the cells, providing a detailed profile of each individual cell.

The key components of a flow cytometer include the fluidic system, optics, detectors, and computer interface. The fluidic system transports cells one at a time through the instrument, while the optics and detectors handle the collection and measurement of data. FACS builds on flow cytometry by not only analyzing but also sorting cells based on these measured characteristics.

In the fluidic system, cells move through the cytometer in a highly controlled manner. To achieve the precision required for accurate measurements, two fluids are employed: sheath fluid (usually a phosphate-buffered saline) and pressurized sample fluid. The sheath fluid flows through the core of the instrument, creating a steady, pressurized stream. The cell suspension is then introduced into the center of this flow, forming what is known as coaxial flow. Due to the pressure difference between the sheath fluid and the sample, the cells align in a single file, ensuring they pass individually through the laser beam, which is crucial for precise measurement. Schematic overview of the fluidic system is shown in Figure 7.

The optics system is responsible for generating the light necessary for measurement and collecting the light that interacts with the cells. A laser serves as the excitation source, and lenses focus and direct the light towards the cells. Light scattered from the cells is measured in two directions: Forward Scatter (FSC), which gives information about cell size, and Side Scatter (SSC), which reveals the internal complexity or granularity of the cell. These scatter signals are independent of any fluorescence the cell may emit. A schematic of the interaction between laser beams and passing cells is shown in Figure 8.

Fluorescent signals are collected from the fluorophores present within the sample, often introduced through techniques like transfection (e.g., the GFP in this study), fluorescent dyes, or fluorescently labeled antibodies. The collection optics use mirrors, lenses, and filters to direct specific wavelengths of the collected light to their respective detectors, enabling the detailed analysis of the fluorescent properties of each cell. Light scattered by measured particles or emitted by fluorophores is filtered by set of band pass filters and quantified by conversion to volts using photodetectors. Two main types of photodetectors with different sensitivity are used – photodiodes and photomultipliers. This analog signal is then converted to a digital signal that can be displayed as a plot or histogram during the analysis.

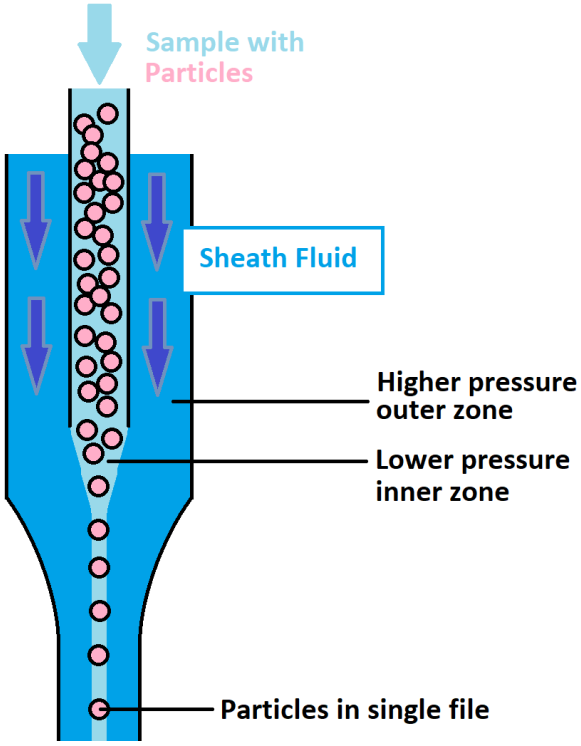


Figure 7. Schematic of the flow cytometry fluidic system. The sample containing particles (shown in pink) is injected into the instrument and surrounded by sheath fluid (shown in dark blue), creating a higher-pressure outer zone. This pressure difference focuses the particles into a single file for precise analysis as they pass through the flow chamber.

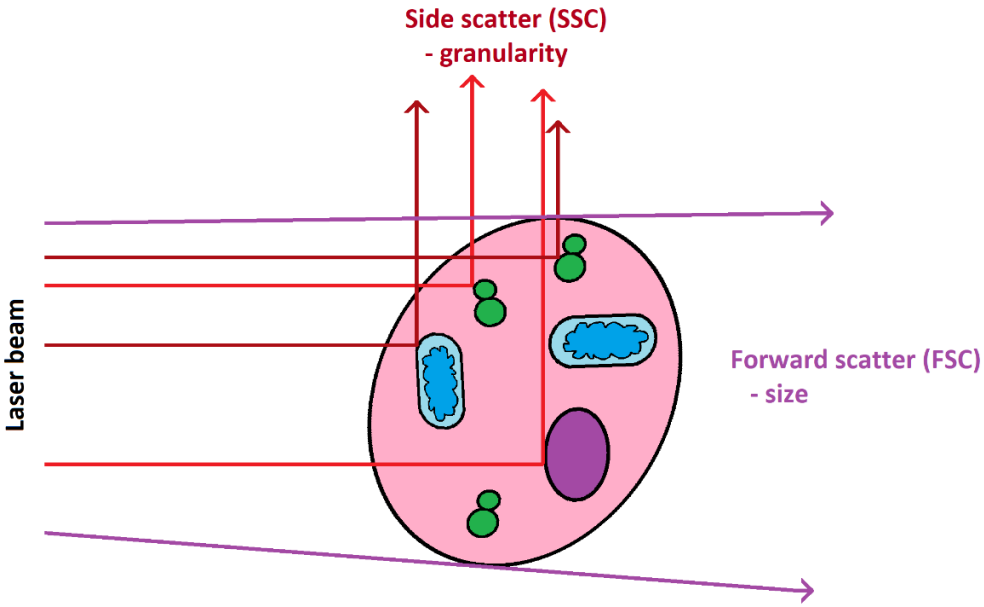


Figure 8. Schematic view of the interaction of laser beams with a living cell passing through. SSC is collected usually at 90° angle relative to the laser beam providing information about the granularity and complexity of the cell. FSC is collected along the same axis as the laser beam and it provides information about the size of the measured particles.

Some flow cytometers are equipped with an electrostatic sorting device, which is a crucial equipment for fulfilling the goals of this project. This device is responsible for capturing and sorting particles passing through the cytometer based on their properties defined by the user (in case of this project the correct size and fluorescence of *E. coli* cells). The cell suspension is injected through a vibrating nozzle, which is responsible for creating small droplets from the continuous stream. The droplets pass through more laser beams and are charged by a charging electrode. Electrostatic deflection of charged droplets that contain analyzed cells is used to guide the droplets into prepared collection tubes through electrostatic field [119, 120].

2 Aims and objectives

Objective of this thesis is to investigate how the N-terminal region regulates recombinant protein production in *Escherichia coli* and to optimize protein yield using a directed evolution-based approach. This research was inspired by previous work in our lab on developing small, non-antibody protein scaffolds, during which the importance of the N-terminal sequence composition as a regulatory element for translation became apparent. To explore this observation further, I designed a systematic study utilizing DNA libraries to manipulate protein expression levels and employed FACS to identify optimal variants.

To achieve these goals, this study focuses on:

- Developing a systematic approach for designing and optimizing small, non-antibody protein scaffolds, including engineering a novel scaffold with tunable binding surfaces to optimize interaction with human IL10.
- Investigating the regulatory role of N-terminal sequence composition by fusing various short sequences to different proteins.
- Constructing a DNA library to assess its impact on recombinant protein yield and selecting optimal variants using FACS.
- Exploring the limits and applicability of this approach across different target proteins, including development and optimization of the protocol to determine yield increase by N-terminal manipulations.

3 Results

Results presented in this dissertation thesis were published in two articles in peer reviewed journals. The first article, *Protein Binder (ProBi) as a New Class of Structurally Robust Non-Antibody Protein Scaffold for Directed Evolution*, describes development of IL10 binding small non-antibody protein scaffold using directed evolution. During this study, the importance of the N-terminal sequence composition as a regulatory element for translation became apparent and therefore I decided to implement similar approach using directed evolution, random DNA libraries and cell sorting to optimize the sequence coding for the N-terminus composition. Results of this project are presented in the second article, *Increasing Recombinant Protein Production in E. coli via FACS-Based Selection of N-terminal Coding DNA Libraries*. My contribution to these studies are listed below.

Protein Binder (ProBi) as a New Class of Structurally Robust Non-Antibody Protein Scaffold for Directed Evolution

> My contribution: Cloning, expression and purification, inhibitory assay

Increasing Recombinant Protein Production in E. coli via FACS-Based Selection of N-terminal Coding DNA Libraries

> My contribution: experiment planning, cloning and expression of all used constructs, flow cytometry experiments, data analysis, preparation of the original draft of the article

The results of this thesis are presented in two scientific articles, which are included as attachments in the appendix. These publications form the foundation of this study, complemented by several unpublished results. These findings are further discussed in the discussion section, where their implications and connections are explored in detail.

3.1 Additional unpublished results

The following figures (Figures 9–15) illustrate experimental results that complement the findings presented in both scientific articles. These figures provide further data and insights, which I discuss later in the thesis, offering a broader perspective on the study.

Figure 9: Illustrates the impact of His-tag position (N-terminal vs C-terminal) on protein expression, showing successful production with the N-terminal tag and stalled expression with the C-terminal tag for the 4PSF and 1N3Y scaffold candidates.

Figure 10: Shows the western blot analysis of GFP cleavage using different linkers and production conditions, highlighting the effectiveness of the new constructs in preventing unwanted cleavage and demonstrating the impact of shorter production times and higher temperatures on reducing GFP cleavage.

Figure 11: Displays SDS-PAGE analysis showing the expression levels of GFP constructs with different-sized DNA sequences coding for the N-terminus, revealing that the smallest library (encoding three amino acids) was the least effective, even reducing GFP expression.

Figures 12 & 13: Show the results of optimizing the production of photoactive yellow protein (PYP) and GFP, highlighting the limited success in enhancing expression for proteins that are already well-expressed. Figure 12 illustrates the gating strategy used during sorting, while Figure 13 presents SDS-PAGE analysis of the protein yields.

Figure 14: Shows SDS-PAGE analysis of GFP production after sorting with and without the addition of the N-terminal library. The sorted cells demonstrated relatively high GFP production, while the unsorted cells produced very low levels, highlighting the effectiveness of FACS selection and the impact of the DNA library on the final yield.

Figure 15: Displays a pie chart illustrating the analysis of 105 plasmid sequences collected from non-fluorescent cells. The chart shows the distribution of sequence types, including mixed signals, empty cloning vectors, frameshift mutations, and stop codons, confirming the validity of the experimental setup.

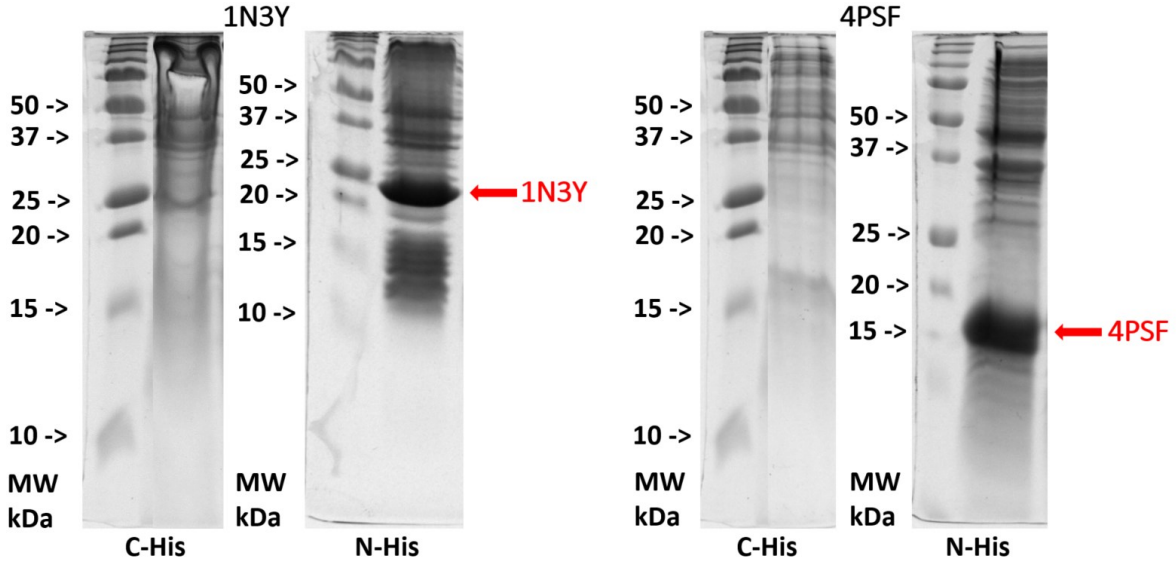


Figure 9. SDS-PAGE analysis of 1N3Y and 4PSF scaffold candidates with both C-terminal (left) and N-terminal (right) His-tag. Expected size of the 1N3Y construct was 22 kDa and 4PSF 15 kDa.

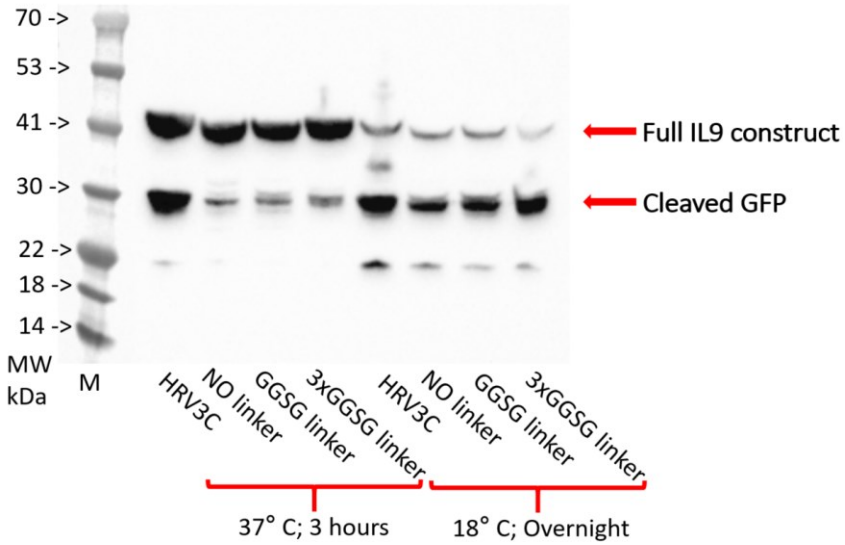
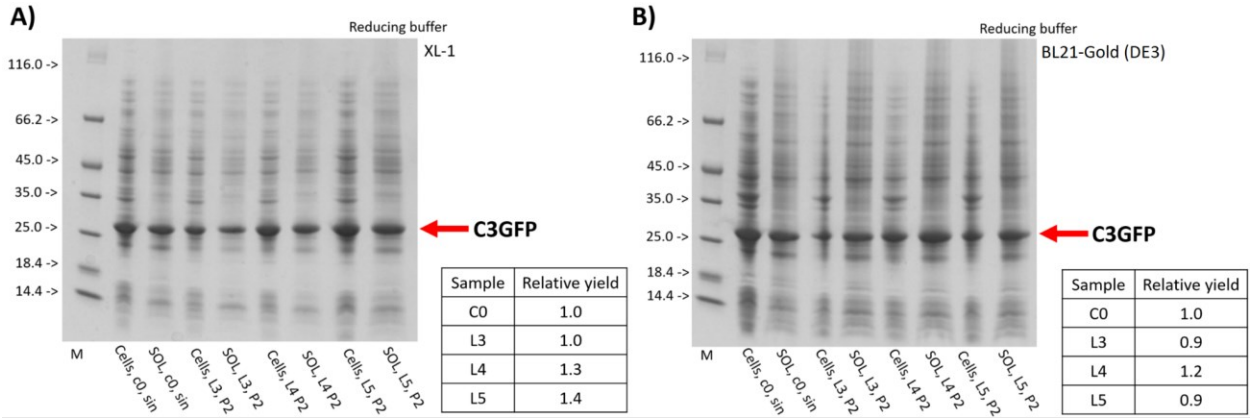


Figure 10. Western blot analysis of the GFP molecule cleavage using different linkers and temperatures for the recombinant protein production. Expected size of the full construct is 42.3 kDa and expected size of the cleaved GFP is 27.5 kDa. Anti-His-tag antibody conjugated with peroxidase was used for the detection targeting the His-tag at the C-terminus of the whole construct. First lane labeled M contains marker, the next four lanes contain samples from the 3 hour production at 37° C and the next four lanes contain samples from the overnight production at 18° C. HRV3C lanes contain the original pKIK sample with the HRV3C cleavage site between the IL9 and GFP gene. NO linker lanes contain the same sample with the cleavage site removed, GGSG linker lanes contain the sample with GGSG linker added and 3xGGSG lanes contain the sample with three times GGSG linker added.



Sample	Relative yield
C0	1.0
L3	1.0
L4	1.3
L5	1.4

Sample	Relative yield
C0	1.0
L3	0.9
L4	1.2
L5	0.9

Figure 11. SDS-PAGE analysis of the produced constructs. Digested cells (labeled "Cells") and soluble fractions (labeled "SOL") of desired constructs. The expected size is 27.5 kDa. Tables show the relative yield calculated from the soluble fractions using ImageJ. A) The GFP construct produced in XL-1 cells. B) GFP construct produced in BL21-Gold (DE3) cells. Expected size of the whole construct is 27.7 kDa.

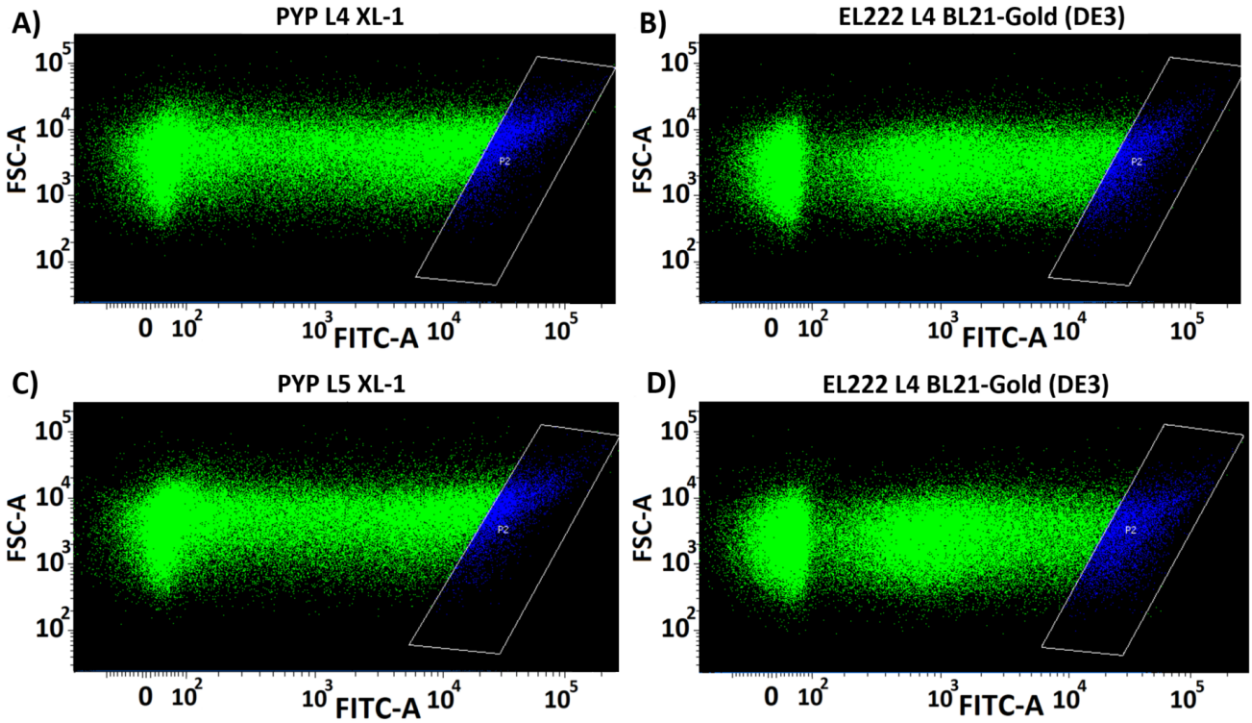


Figure 12. Gate design during PYP-GFP sorting using both *E. coli* cell strains. A) PYP construct with library coding 4 amino acids in *E. coli* XL-1 cells. B) PYP construct with library coding 4 amino acids in *E. coli* BL21-Gold (DE3) cells. C) PYP construct with library coding 5 amino acids in *E. coli* XL-1 cells. D) PYP construct with library coding 5 amino acids in *E. coli* BL21-Gold (DE3) cells.

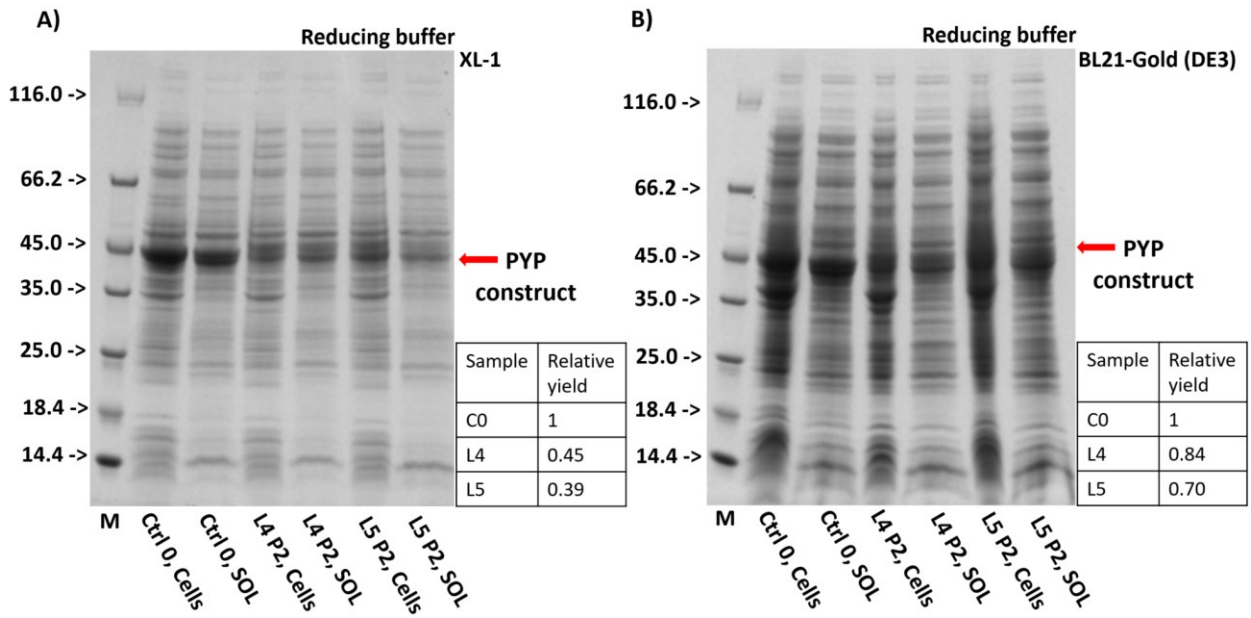


Figure 13. SDS-PAGE analysis of PYP yield enrichment in both *E. coli* strains. The expected size of the PYP construct is 41.4 kDa (+ 4 or 5 amino acids delivered from the library). A) PYP construct produced in *E. coli* XL-1 strain. B) PYP construct produced in *E. coli* BL21-Gold (DE3) strain.

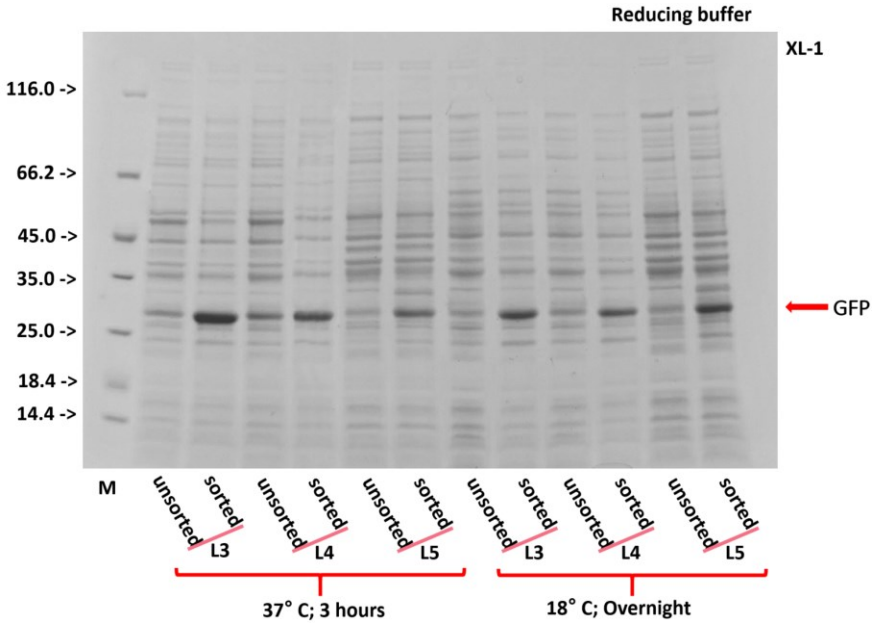


Figure 14. SDS-PAGE analysis of the sorted GFP produced by *E. coli* XL-1. Expected size of the construct is 27.5 kDa (+ 4 or 5 amino acids derived from the N-terminal library). Libraries containing samples are labeled L3, L4 and L5 for three, four and five amino acid libraries, respectively.

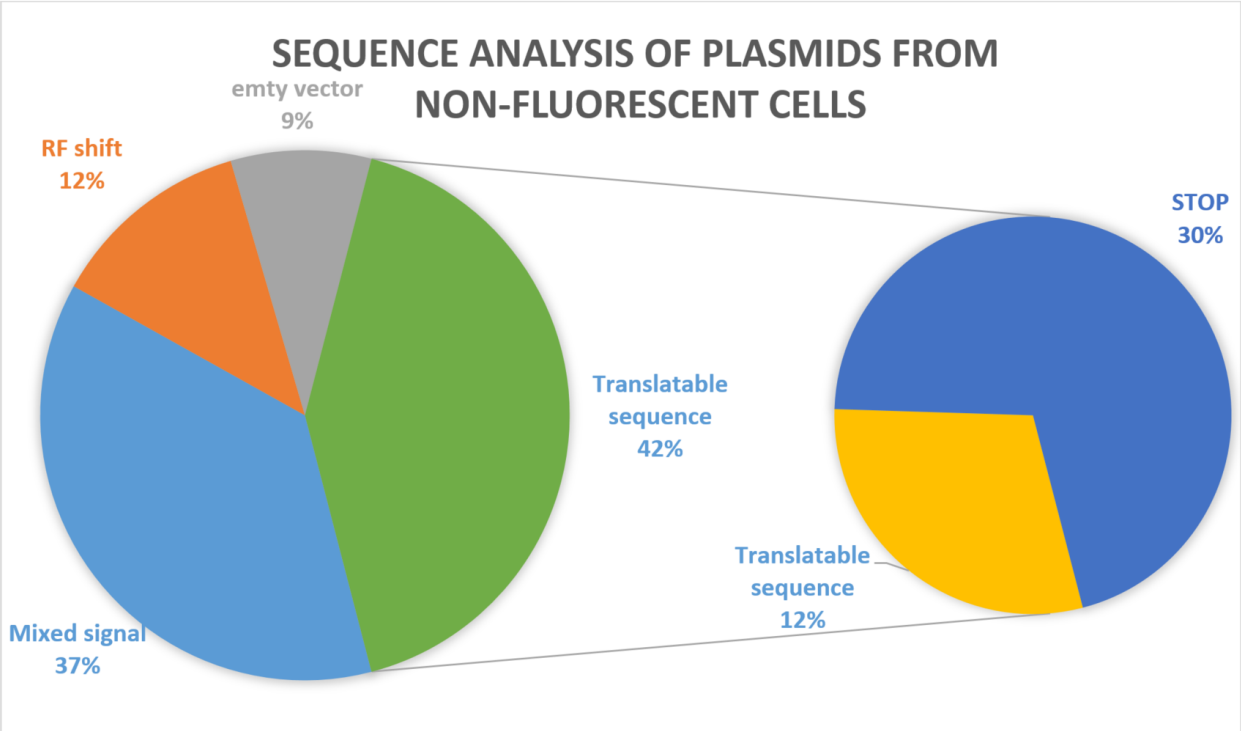


Figure 15. Pie chart displaying the analysis of 105 sequences from the DNA library that inhibit GFP production. Of these sequences, 9% were identified as empty cloning vectors lacking GFP (shown in grey on the left), 12% contained a reading frameshift mutation (shown in orange on the left), and 37% were unreadable due to mixed signals (shown in light blue on the left). The remaining 42% of sequences were readable (shown in green on the left). Among these readable sequences, 70% contained a stop codon, accounting for 30% of the total sequencing results (shown in dark blue on the right).

4 Discussion

4.1 Protein scaffolds and evolution of IL10 binder

Small scaffold derived binders offer several advantages over traditional antibodies. Their significantly smaller size, with a molecular weight of 15 - 20 kDa compared to the 150 kDa of a typical monoclonal antibody, contributes to improved tissue penetration and these proteins can have other beneficial properties such as optimized stability. Unlike antibodies, scaffold based proteins do not need to undergo posttranslational modifications, allowing them to be efficiently expressed in bacterial systems such as *E. coli*. This significantly reduces production costs. Additionally, their development eliminates the ethical concerns associated with animal immunization, further driving interest in novel scaffold molecules [40].

To create a small non-antibody protein binder, ribosome display was utilized to evolve a scaffold trained to bind human IL10 with nanomolar affinity. This study presents a proof-of-concept methodology for identifying protein structures that can be evolved into new protein scaffold constructs. Experimental evidence demonstrated that at least one of these structures could be adapted into a binder for a medically important target—IL10—with nanomolar affinity. No interference with the IL10 signaling pathway was observed, suggesting that the engineered molecule binds IL10 outside the receptor-binding site.

The resulting scaffold was called ProBi, but due to intellectual property reasons, it had to be later renamed to 57aBi. This molecule is based on the best protein binder candidate, which was the human PIH1 domain containing protein 1 (UniProt ID Q9NWS0, structure of PDB ID 4PSF [121]). Several engineered non-antibody scaffolds have been developed as alternatives to antibodies, each utilizing distinct structural elements for molecular recognition. Affibodies, derived from the Z domain of *Staphylococcus aureus* protein A, employ a three-helix bundle, while designed ankyrin repeat proteins use modular α-helical repeats. Affilins, based on human γ-crystallin or ubiquitin, rely on β-sheet structures to present binding surfaces. These scaffolds have been successfully applied in research, diagnostics, and therapeutics [40].

The 57aBi molecule presents a novel monomeric scaffold that expands the landscape of synthetic binders. Unlike many other existing scaffolds, 57aBi features two independently mutable surface

patches, allowing for flexible engineering of binding properties. Its selection followed a systematic computational and experimental pipeline to ensure favorable solubility, thermostability, and efficient recombinant expression in *E. coli*.

As an initial step, a comprehensive search of the Protein Data Bank was conducted using specific selection criteria [122]. The ideal scaffold needed a well-characterized, high-resolution structure, the ability to be expressed in a soluble form in bacterial strains, and low immunogenicity and toxicity. To prevent folding complications during bacterial expression, proteins lacking cysteines were prioritized. The shortlisted candidates were experimentally evaluated for their biophysical properties, ensuring efficient production, mutability, and high solubility. Thermal stability was also assessed, with the top candidate, 4PSF, demonstrating exceptional stability above 60 °C, providing an advantage over traditional antibodies.

By successfully developing high-affinity IL10 binders, this work demonstrated the potential of 57aBi as a versatile and efficient alternative to existing scaffolds. With its robust structural framework and optimized expression properties, 57aBi complements current synthetic binders and opens new opportunities for protein engineering.

4.1.1 Protein scaffolds and their N-terminal composition importance

During the research on IL10 binders and scaffolds in general, the position of the His-tag (either at the N-terminus or C-terminus) significantly impacted the expression of the constructs in the *E. coli* BL21 (DE3) strain. The best candidates, protein constructs inspired by the PDB entries 4PSF, 1N3Y, 4I3B, and 1W2I, were cloned with both N-terminal and C-terminal His-tag variants. When the His-tag was placed at the N-terminus, the proteins were successfully produced. However, when the tag was moved to the C-terminus, protein expression stalled completely, as shown for 4PSF and 1N3Y candidates in Figure 9. This behavior was consistent across several constructs intended for IL10 binder development, specifically 4PSF, 1N3Y, and 4I3B. In contrast, the 1W2I construct did not express at all in the *E. coli* BL21 (DE3) strain but did express in the *E. coli* C41 (DE3) and C43 (DE3) strains, strains commonly used to produce toxic proteins.

This finding shows the importance of the N-terminal composition of produced proteins. Indeed, it has been shown before that repetitive CAT triplets coding for His-tag downstream from an AGG codon placed in the +2 position helped to restart the production of human erythropoietin, which was originally stalled by the placement of the AGG codon in this position [111]. NGG codons, including AGG, are known to be associated with reduced gene expression in *E. coli* [110].

The observed change in gene expression of target proteins in the IL10 binder study could also point to the importance of the C-terminus rather than the N-terminus. The composition of the C-terminus in relation to gene expression has been discussed in the literature as well [123, 124], but evidence supporting the importance of the N-terminus is predominant. This is why I decided to focus on the N-terminus recognizing the significance of the N-terminal sequence composition in protein production, as has also been demonstrated in numerous studies [27-36]. A similar approach was applied using directed evolution with random DNA libraries and cell sorting. Instead of focusing on improving binding affinity through *in vitro* evolution, the goal was to enhance recombinant protein yield by selecting the optimal N-terminal sequence for various proteins.

4.2 Optimizing the sequence coding the N-terminus

Similarly to the development of protein binders, directed evolution was used for optimizing the N-terminus. The experimental design utilized a randomized DNA library coding for the N-terminal region and FACS selection for *E. coli* cells showing increased production of target protein fused to a GFP reporter. Completely random DNA libraries (instead of the NNK libraries as used for the IL10 binder development) were utilized to include possibility of observing the direct impact of nucleic acid usage and interaction on the recombinant protein yield.

I chose the c3GFP as a reporter molecule for all my constructs. Compared to a WT GFP molecule the c3GFP includes several mutations resulting in increased fluorescence. These mutations also facilitate proper maturation of the protein at 37° C and the molecule shows lower tendency to aggregate [116, 117]. The c3GFP mutant was derived from a DNA shuffling experiment, similar to the approach used in this study nearly 30 years later. However, instead of employing FACS, the authors of the previous study conducted a visual screen under UV excitation [116].

4.2.1 Successful production optimization of IFNb

The most significant result of this study regarding increased recombinant protein production was obtained during the optimization of soluble IFNb production. IFNb is a critical target for production optimization. Even though the recombinant bacterial product lacks glycosylation, it is still biologically active [125]. Multiple clinical trials have been conducted with recombinant human IFNb, and it has been approved for the treatment of multiple sclerosis [126, 127] and it also showed a potential as adjuvant in prophylactic vaccines against infectious diseases [128, 129]. Recombinant human IFNb is produced in *E. coli* and Chinese Hamster Ovary cell lines to meet the current demand [125]. Given the demand for recombinant IFNb, the outlined methodology for optimizing recombinant protein production could be of significant interest to various academic institutions and industry partners. While a potential limitation is the addition of a few amino acids at the N-terminus, which may affect its suitability as a therapeutic, this modification is unlikely to hinder its biological activity. As such, the product could still be valuable for further research, pending appropriate biological testing.

Besides IFNb I was able to enhance yield of other constructs as well, including thioredoxin A (TrxA), glutathione S-transferase (GST) and EL222 light-regulated bacterial DNA-binding transcription factor. During these experiments, several key observations were made, which are discussed below.

4.2.2 GFP cleavage during optimization of TrxA production

In certain cases, a spontaneous cleavage of the GFP reporter from the rest of the construct was observed, which typically resulted in increased fluorescence of the cells during cell sorting. The cleavage can occur even during recombinant protein production post-sorting. To avoid this false-positive selection of clones with cleaved GFP, a different gating strategy was developed as described below.

This issue was encountered while attempting to enhance the production of several proteins and further investigation was carried out using human IL9. An HRV3C protease cleavage site was introduced in the original plasmid, which was initially intended to cleave the GFP after selection,

if needed. To avoid unwanted cleavage, the HRV3C cleavage site was replaced with a tetrapeptide GGSG linker and a dodecamer (3xGGSG) linker. A version without any linker between the IL9 gene and the GFP was also created. These three constructs were expressed and their expression compared to the performance of the original vector. The ratio of cleaved to not cleaved versions was compared using a western blot with an anti-His-tag antibody conjugated with peroxidase, targeting the His-tag at the C-terminus of the whole construct. The western blot analysis revealed improved coherence of all three new constructs, effectively preventing cleavage, while also showing that using shorter production time (3 hours at 37° C instead of overnight production at 18° C) lowers the chance of the GFP cleavage as well. The western blot analysis is shown in Figure 10. Based on the results of this experiment, the HRV3C cleavage site was removed. In addition, higher production temperatures were adopted with shorter production times, coupled with an improved gating strategy. These combined enhancements significantly reduced the occurrence of GFP cleavage in subsequent experiments, almost eliminating the problem. This comprehensive approach ensured results that are more reliable and streamlined the experimental processes.

To the best of my knowledge, the current literature does not address this issue. While Cambray *et al.* conducted an extensive study on N-terminal sequence optimization using a methodology similar to the one presented in this study, including synthetic library sequences fused to the N-terminus of the target protein, they did not address the cleavage problem [37]. This is because their study focused solely on super folder GFP without any additional fused proteins. The same applies for a study by Verma *et al.*, who conducted a comprehensive research on efficiency of protein synthesis, but focused mainly on one GFP derivative [33]. Other studies on N-terminal sequence optimization have also been limited to fluorescent proteins, and thus did not need to consider the cleavage of fluorescent partners [28, 81]. Furthermore, studies by Looman *et al.* and Bivona *et al.* focused on optimizing the production of non-fluorescent recombinant proteins using a rational design approach, without employing high-throughput directed evolution methods. As a result, they did not require fluorescent partners and avoided the cleavage issue [27, 87]. Thangadurai *et al.* conducted experiments on both a fluorescent protein (GFP) and a non-fluorescent protein (streptokinase), but they also avoided the cleavage issue by examining the proteins separately while investigating the effects of rare codons in the N-terminal region [32].

Although the current literature does not address this issue, the observation of GFP cleavage is crucial to the optimization method proposed in this thesis. Notably, GFP cleavage reoccurred when the cells from the gate representing cleaved GFP containing cells were grown again after cell sorting, suggesting that the short sequence encoding the N-terminus directly affects the integrity of the entire construct.

4.2.3 Impact of randomized DNA library size on protein yield

Based on the current state of knowledge, three distinct library sizes were designed for this study: nine, twelve, and fifteen nucleotides encoding three, four, and five N-terminal amino acids, respectively. These sizes were chosen with the expectation that shorter libraries might produce better protein variants for future experiments.

Generally, the position of the second amino acid (coded by a codon following the initiation ATG codon) received a lot of attention in the literature [27, 87, 111]. Bivona *et al.* conducted a systematic study highlighting the importance of this position by replacing the second amino acid of recombinant Igα with each of the twenty canonical amino acids. Their findings showed a significant impact on protein yield, with up to a ten-fold difference in production [27]. Other studies also examined longer region, such as Goodman *et al.*, whose research covered the first eleven codons of 137 endogenous essential *E. coli* genes [28] or 756 randomly generated initial thirteen codons analyzed by Voges *et al.* [90]. Based on these findings, Verma *et al.* focused on the first three to five amino acids, which is what I present in my study as well [33]. Importance of codons in the second, third and fifth position were also studied by Gonzalez *et al.* [110]. Ohno *et al.* investigated the role of minor codons in the region of a gene encoding the N-terminus and reached a notable conclusion: "*When the distance between the initiation codon and the minor codon is greater than 50–60 codons, a queue of ribosomes at the minor codon does not block the entry of a ribosome in the translation process*" [103]. This result suggests that a much longer segment of the sequence influences the regulation of translation. However, extending the variable region by up to sixty codons would significantly alter the target protein, making it less suitable for further research applications. Additionally, such optimization would require modifying codons within the sequence of the actual protein product, which would result in a similar research approach utilized by Bhandari *et al.*, who emphasized mRNA folding over codon rarity [86].

To evaluate the impact of these libraries, cells expressing GFP with the different sized DNA sequences coding for the N-terminus were sorted. The SDS-PAGE analysis and yield differences, as shown in Figure 11, reveal that the smallest library, encoding just three amino acids, was the least effective. In fact, for GFP the expression was even reduced.

Based on these results, I decided not to proceed with the twelve and fifteen nucleotide libraries, which code for four and five amino acids, respectively, as the smallest library (nine nucleotides, encoding three amino acids) yielded the least favorable outcomes. I believe that the addition of one or two extra amino acids from the larger libraries will still produce usable proteins for subsequent experiments, and that such small changes in the N-terminus are unlikely to significantly affect the results.

4.2.4 Challenges in optimizing well expressed and hard to express proteins

I observed that this methodology significantly increased yields for some target proteins, while it was less effective or ineffective for others. This limitation was particularly evident in proteins that are already well expressed in *E. coli*, such as GFP and PYP. I hypothesize that the N-terminal sequences of these proteins is already near its optimal composition, leaving only a little room for further improvement. This is in agreement with other results of this study, where decreased protein expression was observed after the cell sorting procedure for some of the GFP and all PYP constructs. Unsuccessful optimization of the PYP production ended up in obtaining 39 - 84 % of the WT variant of the protein. Results of this experiment are shown in Figure 12 (FACS) and 13 (SDS-PAGE analysis). Results of the optimization of GST production (data not shown) are in concordance with this hypothesis, since the GST construct should be reasonably well expressed, as GST fusion is often used for expression and purification of proteins produced in *E. coli* [130, 131]. The soluble GST construct production was increased by only 3.3-fold, which is much lower compared to IFNb discussed above. Voges *et al.* also came to a similar conclusion while optimizing production of proteins that already showed high expression [90]. This is also why I did not choose the protein scaffold candidates presented in this thesis for optimization using this process, as one of the key selection criteria was ease of production in *E. coli.*

Based on a series of experiments in which I was unable to increase the yield of already well-expressed recombinant proteins, I decided to evaluate the impact of adding a DNA library and the effectiveness of cell sorting. To investigate this, GFP fused to the N-terminal library was produced and the cells were sorted using standard protocols. GFP was chosen for its simplicity, as fusion proteins often require more process optimization. After sorting, I compared the GFP yield between sorted and unsorted cells, with the same number of singlet cells counted using FACS. The results for *E. coli* XL-1, which provided the most prominent data, are shown in Figure 14 (SDS-PAGE). The sorted cells clearly exhibited relatively high GFP production, as expected. Surprisingly, the unsorted cells produced very low levels of the GFP construct. These results suggest that the addition of the DNA library significantly reduces the yield of the final product and highlights the efficiency of FACS selection.

To investigate what occurred in the cells harvested from the non-fluorescent gate, I collected and sequenced 105 plasmids from these cells. Among the 105 sequences, 39 showed mixed signals, 9 were empty cloning vectors without GFP (which can happen during cloning when attempting to preserve high library complexity by harvesting as many clones as possible), 13 contained a library version with a frameshift mutation, and 44 had sequences derived from the library (containing 17 unique sequences, with some of them repeating not more than 5 times). Of the 44 correctly sized sequences, 31 contained a stop codon, confirming the validity of the experimental setup. These findings are presented as a pie chart in the Figure 15.

Conversely, proteins that are difficult to express in *E. coli* also presented challenges. For instance, I was unable to produce erythroferrone and IL9. IL9 proved difficult to optimize due to its complex structure, which includes disulfide bonds and other features not easily accommodated by a prokaryotic expression system. N-terminal sequence optimization, while effective in many cases, could not overcome these structural and system-related barriers, limiting the ability to enhance the production of such proteins. The overall expression of IL9 was too low, even after cell sorting, likely due to the bacterial expression system's inability to accommodate the IL9 structure. I was able to detect the IL9 product using western blot technique and experiments with this construct delivered valuable results while solving the GFP cleavage problem as already discussed. While the initial experiment offered valuable insights, I ultimately determined that the construct was not

suitable for further investigation. The low quantity and questionable quality of the desired product were the reasons not to continue with this construct further.

4.2.5 N-terminal library decreases the yield before sorting

Counterintuitively, the introduction of the N-terminal library initially led to a significant decrease in protein production. GFP (without any fusion) production was suppressed, necessitating a sorting process to restore it. As already discussed, GFP is an example of a protein, which is easily produced in *E. coli*. Therefore, it makes sense that changing randomly the N-terminus composition decreases its production level. The experiment also demonstrated effectiveness of the sorting procedure in selecting cells with higher protein yields as well as potential to increase the recombinant protein yield. Results supporting this hypothesis are shown in Figure 14, where I compare production levels of sorted and not sorted cells containing the DNA library vector, instead of using WT constructs without randomized DNA as a control.

4.3 From unexplained mechanisms to practical use of N-terminal libraries

Since N-terminal sequence effects are protein-specific and cannot be applied universally, I concluded that establishing a universal explanation was not only beyond the scope of this study, but generally an unrealistic goal. Instead, I shifted my focus toward a more practical application - utilizing N-terminal libraries as a tool to enhance recombinant protein production. This shift in focus allowed me to prioritize the practical outcomes of my research, aiming to optimize yield rather than solely understanding the underlying mechanisms driving the increase. This is in conclusion with extensive study by Cambray *et al.*, where the authors claim that they analyzed 244 000 N-terminal sequences but were not able to describe the mechanism behind the yield increase. They investigated eight intrinsic sequence properties reported to be the main factors affecting translation efficiency. These predictors describe nucleotide content of coding sequences, patterns of the codon usage, hydrophobicity of the coded polypeptide and stability of mRNA secondary structures tiled along the transcript. The authors conclude: "*Several of our main findings are negative. Over our entire dataset, we were able to explain just over half (53%) of the total variance in protein production (...) the lack of explanation points to fundamental mechanisms yet to be discovered*" [37]. This corresponds to a conclusion with research by Verma *et al.*, who

conclude one year later: "*The expression levels of 213,708 eGFP variants with randomized nucleotide and amino acid sequences in those positions resulted in substantial differences in fluorescence and protein levels. The effect of the assayed sequences was dependent on both nucleotide and amino acid sequence, which suggests that a combination of tRNA, mRNA, ribosome, and nascent polypeptide chain interactions define the efficiency of protein synthesis at the very N-terminus.*" [33].

These two comprehensive studies provide a large amount of data and explore various potential mechanisms that could explain the yield changes. However, both studies primarily focused on a single target, a variant of GFP, and their conclusions remain unclear [33, 37]. Introducing additional proteins into the analysis adds another layer of complexity, which is why I believe that determining the exact mechanism behind this phenomenon is beyond the scope of my current work. On the other hand, altering the N-terminal sequence could still be pragmatically used as a biotechnological tool for optimizing recombinant protein production.

I initially considered publishing some of my optimized specific sequences obtained through the optimization of IFNb but ultimately decided against it due to their limited generalizability. Factors such as bacterial strain, vector, target gene, and codon usage significantly influence sequence usability, meaning the presented sequences may not be applicable beyond my specific experimental conditions. I analyzed 20 obtained sequences and realized that they exhibited unique and unexpected features. Notably, I observed the presence of three stop codons within these sequences, which at first seems counterintuitive, as stop codons typically terminate translation. It is also worth noting, that there were different versions of the stop codons present in these sequences at different positions (TAA and TGA), while specifically TAA stop codon was found to be by far the most preferred stop codon for highly expressed genes in bacteria, highlighting its efficiency [132]. The presence of these stop codons suggests that factors other than the nascent protein itself, such as mRNA structure, may play a role in translation regulation, since the translation may start from a different start codon present at the beginning of the IFNb gene but after the randomized library. From calculations I did using the TISIGNER tool, I found no strong correlation between the sequences containing stop codons and a low probability of mRNA secondary structure formation [86]. Bacteria is also known to have polycistronic mRNA, so this finding is not overly

surprising. Perhaps a surprising part of the observation is that adding sequences containing stop codons can be on some occasions related to a higher expression of the target gene.

I hypothesize that these N-terminal stop codons may slow down the initial steps of translation, serving as a regulatory element. This effect could mimic the impact of rare codons, which similarly influence translation kinetics in unexpected ways [32, 107] as I discuss in the introduction section. The native IFNb gene, which follows the optimized N-terminal sequence, still contains its start codon (ATG) immediately after the inserted region. This second start codon could initiate actual translation of the target gene, suggesting that the optimized N-terminal sequence may not be translated into the protein itself, but it forms a part of the untranslated region (UTR). The pattern start codon – stop codon – start codon in a short DNA N-terminal segment introduces an additional layer of complexity in understanding the molecular mechanisms underlying yield optimization.

Considering the published research in context of my own experimental results, I shifted my focus towards utilizing N-terminal libraries to enhance protein production, rather than continuing to explore the underlying mechanism driving the yield increase.

4.4 My research approach in the context of modern biology: Prioritizing practical outcomes over mechanistic descriptions

My research on optimizing protein production, despite not fully understanding the underlying mechanisms, reflects the shift of approach to research in biology discussed by Victor de Lorenzo [133]. In his work, de Lorenzo discusses machine learning (ML) and artificial intelligence (AI) in relation to molecular biology research. De Lorenzo points out that traditional mechanistic approaches, which seek to understand biological phenomena from first principles, are often not sufficient when faced with the complexity and dynamic nature of complex biological systems. This mirrors the challenges I encounter in protein production, where the exact molecular mechanisms driving optimization remain elusive. Yet, as de Lorenzo emphasizes, AI driven methods like AlphaFold have shown that practical advancements can be achieved by relying on patterns and empirical data rather than fully grasping the mechanistic details [133]. In my case, similar principles apply: I optimize protein production by leveraging observed outcomes, without needing to first decipher the intricate processes involved.

This shift away from purely mechanistic understanding represents a broader trend in biology, as exemplified by AI driven models, which increasingly guide complex decision making and prediction in fields like protein folding, gene regulation, and metabolic engineering. My research fits neatly into this framework. By capitalizing on empirical effects, I am able to enhance protein yields, mirroring the way that ML models make reliable predictions even without complete understanding of the systems involved. As de Lorenzo notes, "*Molecular biology reduces the complexity of a given phenomenon to a point where rigorous logic can be applied, experimental results unambiguously interpreted, and conclusions fixed as permanent pieces of knowledge*" [133]. However, in many modern biological contexts, including this work on N-terminal sequence optimization, achieving practical outcomes may not require this degree of mechanistic detail understanding. Instead, the probabilistic and data driven approaches that de Lorenzo describes allow researchers to focus on outcomes, advancing scientific knowledge even when the finer mechanistic details remain unexplored.

Ultimately, my study reinforces the idea that in many cases, achieving practical, beneficial outcomes in biology is possible without fully understanding the intricate details of the underlying processes. This approach, in line with de Lorenzo's argument, highlights the power of experience-based optimization, particularly when addressing complex challenges, such as large-scale protein production. As the biological sciences continue to embrace AI and ML as essential tools, the role of empirical, data driven methodologies will likely grow, enabling scientists to address pressing biological problems in real time, without being hindered by the need for complete mechanistic clarity.

In light of these insights, ML and AI could also be employed to optimize N-terminal sequences for protein production. By using high throughput techniques like next generation sequencing, large datasets of optimized sequences could be generated, providing the input for AI driven analysis. These models could identify patterns between sequence variations and protein expression efficiency, even without fully understanding the underlying mechanisms. This data-driven approach could significantly accelerate the optimization of N-terminal sequences, aligning with the probabilistic methods highlighted by de Lorenzo [133].

5 Conclusions

In this study, I present the results of two related projects. The first project focused on developing an IL10 binder based on a newly developed 57aBi protein scaffold using a directed evolution approach that leveraged high-throughput screening facilitated by usage of DNA libraries. During this process, a key factor influencing recombinant protein production emerged—the composition of the gene sequence, particularly the coding region for the N-terminus, proved to be crucial. This discovery inspired the second project, which aimed to optimize N-terminal sequences to enhance recombinant protein yield, again utilizing directed evolution.

Building on this insight, I systematically optimized gene sequences using a directed evolution based methodology, allowing me to screen large pools of sequence variants and identify those that enhance soluble protein expression. This approach ultimately led to protein yield increases of up to 30-fold. These findings complement the systematic scaffold selection process employed in the first study, in which a new non-antibody scaffold was developed through structural analysis, computational screening, and experimental validation. Using the 57aBi scaffold, high-affinity binders against IL10 were successfully generated, demonstrating both its adaptability and the effectiveness of selection strategies based on DNA libraries.

Both studies relied on large-scale sequence diversification and high-throughput screening to address key challenges in protein engineering—whether in designing a novel binding scaffold or optimizing expression efficiency. Directed evolution principles were applied at different stages of protein development, leading to the identification of functional IL10 binders in the first project and enhanced production of multiple target proteins in the second.

6 References

1. Pham, P.N., et al., *Protein Binder (ProBi) as a New Class of Structurally Robust Non-Antibody Protein Scaffold for Directed Evolution.* Viruses, 2021. **13**(2).

2. Huliciak, M., et al., *Combined in vitro and cell-based selection display method producing specific binders against IL-9 receptor in high yields.* FEBS J, 2023. **290**(11): p. 2993-3005.

3. Kolarova, L., et al., *De novo developed protein binders mimicking Interferon lambda signaling.* FEBS J, 2022. **289**(9): p. 2672-2684.

4. Rodnina, M.V., *Translation in Prokaryotes.* Cold Spring Harbor Perspectives in Biology, 2018. **10**(9).

5. Li, G.W., et al., *Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources.* Cell, 2014. **157**(3): p. 624-635.

6. Wong, H.R., *Translation.* Critical Care Medicine, 2005. **33**(12): p. S404-S406.

7. Jackson, D.A., P. Berg, and R.H. Symons, *Biochemical Method for Inserting New Genetic Information into DNA of Simian Virus 40 - Circular Sv40 DNA Molecules Containing Lambda Phage Genes and Galactose Operon of Escherichia-Coli.* Proceedings of the National Academy of Sciences of the United States of America, 1972. **69**(10): p. 2904-&.

8. Berg, P., *Nobel Lecture. NobelPrize.org. Nobel Prize Outreach AB 2024. Wed. 27 Mar 2024.* <https://www.nobelprize.org/prizes/chemistry/1980/berg/lecture/> 1980.

9. Itakura, K., et al., *Expression in Escherichia coli of a chemically synthesized gene for the hormone somatostatin.* Science, 1977. **198**(4321): p. 1056-63.

10. Goeddel, D.V., et al., *Direct expression in Escherichia coli of a DNA sequence coding for human growth hormone.* Nature, 1979. **281**(5732): p. 544-8.

11. Goeddel, D.V., et al., *Expression in Escherichia-Coli of Chemically Synthesized Genes for Human Insulin.* Proceedings of the National Academy of Sciences of the United States of America, 1979. **76**(1): p. 106-110.

12. Crea, R., et al., *Chemical Synthesis of Genes for Human Insulin.* Proceedings of the National Academy of Sciences of the United States of America, 1978. **75**(12): p. 5765-5769.

13. Baneyx, F., *Recombinant protein expression in Escherichia coli.* Current Opinion in Biotechnology, 1999. **10**(5): p. 411-421.

14. Zhang, Y., et al., *Expression of eukaryotic proteins in soluble form in Escherichia coli.* Protein Expression and Purification, 1998. **12**(2): p. 159-165.

15. Schein, C.H., *Production of Soluble Recombinant Proteins in Bacteria.* Bio-Technology, 1989. **7**(11): p. 1141-1147.

16. Lavallie, E.R. and J.M. Mccoy, *Gene Fusion Expression Systems Escherichia-Coli.* Current Opinion in Biotechnology, 1995. **6**(5): p. 501-506.

17. Sorensen, H.P. and K.K. Mortensen, *Advanced genetic strategies for recombinant protein expression in Escherichia coli.* Journal of Biotechnology, 2005. **115**(2): p. 113-128.

18. Rosano, G.L., E.S. Morales, and E.A. Ceccarelli, *New tools for recombinant protein production in Escherichia coli: A 5-year update.* Protein Science, 2019. **28**(8): p. 1412-1422.

19. Baeshen, M.N., et al., *Production of Biopharmaceuticals in E. coli: Current Scenario and Future Perspectives.* Journal of Microbiology and Biotechnology, 2015. **25**(7): p. 953-962.

20. Gopal, G.J. and A. Kumar, *Strategies for the Production of Recombinant Protein in Escherichia coli.* Protein Journal, 2013. **32**(6): p. 419-425.

21. Jia, B. and C.O. Jeon, *High-throughput recombinant protein expression in Escherichia coli: current status and future perspectives.* Open Biology, 2016. **6**(8).

22. Overton, T.W., *Recombinant protein production in bacterial hosts.* Drug Discovery Today, 2014. **19**(5): p. 590-601.

23. Makrides, S.C., *Strategies for achieving high-level expression of genes in Escherichia coli.* Microbiological Reviews, 1996. **60**(3): p. 512-+.

24. Rosano, G.L. and E.A. Ceccarelli, *Recombinant protein expression in Escherichia coli: advances and challenges.* Front Microbiol, 2014. **5**: p. 172.

25. Stevens, R.C., *Design of high-throughput methods of protein production for structural biology.* Structure, 2000. **8**(9): p. R177-R185.

26. Hershey, J.W.B., N. Sonenberg, and M.B. Mathews, *Principles of Translational Control.* Cold Spring Harbor Perspectives in Biology, 2019. **11**(9).

27. Bivona, L., et al., *Influence of the second amino acid on recombinant protein expression.* Protein Expression and Purification, 2010. **74**(2): p. 248-256.

28. Goodman, D.B., G.M. Church, and S. Kosuri, *Causes and Effects of N-Terminal Codon Bias in Bacterial Genes.* Science, 2013. **342**(6157): p. 475-479.

29. Gutierrez, G., L. Marquez, and A. Marin, *Preference for guanosine at first codon position in highly expressed Escherichia coli genes. A relationship with translational efficiency.* Nucleic Acids Research, 1996. **24**(13): p. 2525-2527.

30. Ojima-Kato, T., S. Nagai, and H. Nakano, *N-terminal SKIK peptide tag markedly improves expression of difficult-to-express proteins in Escherichia coli and Saccharomyces cerevisiae.* Journal of Bioscience and Bioengineering, 2017. **123**(5): p. 540-546.

31. Stenstrom, C.M., et al., *Codon bias at the 3 '-side of the initiation codon is correlated with translation initiation efficiency in Escherichia coli.* Gene, 2001. **263**(1-2): p. 273-284.

32. Thangadurai, C., et al., *Rare codon priority and its position specificity at the 5 ' of the gene modulates heterologous protein expression in Escherichia coli.* Biochemical and Biophysical Research Communications, 2008. **376**(4): p. 647-652.

33. Verma, M., et al., *A short translational ramp determines the efficiency of protein synthesis.* Nature Communications, 2019. **10**.

34. Stenstrom, C.M. and L.A. Isaksson, *Influences on translation initiation and early elongation by the messenger RNA region flanking the initiation codon at the 3 ' side.* Gene, 2002. **288**(1-2): p. 1-8.

35. Allert, M., J.C. Cox, and H.W. Hellinga, *Multifactorial Determinants of Protein Expression in Prokaryotic Open Reading Frames.* Journal of Molecular Biology, 2010. **402**(5): p. 905-918.

36. Puri, N., et al., *Effect of the codon following the ATG start site on the expression of ovine growth hormone in Escherichia coli.* Protein Expr Purif, 1999. **17**(2): p. 215-23.

37. Cambray, G., J.C. Guimaraes, and A.P. Arkin, *Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in Escherichia coli.* Nature Biotechnology, 2018. **36**(10): p. 1005-+.

38. Hey, T., et al., *Artificial, non-antibody binding proteins for pharmaceutical and industrial applications.* Trends in Biotechnology, 2005. **23**(10): p. 514-522.

39. Simeon, R. and Z.L. Chen, *In vitro-engineered non-antibody protein therapeutics.* Protein & Cell, 2018. **9**(1): p. 3-14.

40. Skerra, A., *Alternative non-antibody scaffolds for molecular recognition.* Curr Opin Biotechnol, 2007. **18**(4): p. 295-304.

41. Banta, S., K. Dooley, and O. Shur, *Replacing antibodies: engineering new binding proteins.* Annu Rev Biomed Eng, 2013. **15**: p. 93-113.

42. Charles A Janeway, J., et al., *Immunobiology: The Immune System in Health and Disease*. 5th edition ed. 2001, New York: Garland Sciences.

43. Moulds, J.M., *Introduction to antibodies and complement.* Transfus Apher Sci, 2009. **40**(3): p. 185-8.

44. Mestecky, J., *Structure of antibodies.* J Oral Pathol, 1972. **1**(6): p. 288-300.

45. Meng, E.C., et al., *UCSF ChimeraX: Tools for structure building and analysis.* Protein Science, 2023. **32**(11).

46. Pettersen, E.F., et al., *UCSF ChimeraX: Structure visualization for researchers, educators, and developers.* Protein Science, 2021. **30**(1): p. 70-82.

47. Harris, L.J., et al., *Refined structure of an intact IgG2a monoclonal antibody.* Biochemistry, 1997. **36**(7): p. 1581-97.

48. Skerra, A., *Engineered protein scaffolds for molecular recognition.* J Mol Recognit, 2000. **13**(4): p. 167-87.

49. Bedford, R., et al., *Alternative reagents to antibodies in imaging applications.* Biophys Rev, 2017. **9**(4): p. 299-308.

50. Rothe, C. and A. Skerra, *Anticalin((R)) Proteins as Therapeutic Agents in Human Diseases.* BioDrugs, 2018. **32**(3): p. 233-243.

51. Vazquez-Lombardi, R., et al., *Challenges and opportunities for non-antibody scaffold drugs.* Drug Discov Today, 2015. **20**(10): p. 1271-83.

52. Lutz, S., *Beyond directed evolution--semi-rational protein engineering and design.* Curr Opin Biotechnol, 2010. **21**(6): p. 734-43.

53. Gebauer, M. and A. Skerra, *Engineered protein scaffolds as next-generation antibody therapeutics.* Curr Opin Chem Biol, 2009. **13**(3): p. 245-55.

54. Fiorentino, D.F., M.W. Bond, and T.R. Mosmann, *2 Types of Mouse T-Helper Cell .4. Th2 Clones Secrete a Factor That Inhibits Cytokine Production by Th1 Clones.* Journal of Experimental Medicine, 1989. **170**(6): p. 2081-2095.

55. Saraiva, M., P. Vieira, and A. O'Garra, *Biology and therapeutic potential of interleukin-10.* J Exp Med, 2020. **217**(1).

56. Walter, M.R. and T.L. Nagabhushan, *Crystal structure of interleukin 10 reveals an interferon gamma-like fold.* Biochemistry, 1995. **34**(38): p. 12118-25.

57. *The PyMOL Molecular Graphics System, Version 2.3.0, Schrödinger, LLC.*

58. Walter, M.R. and T.L. Nagabhushan, *Crystal-Structure of Interleukin-10 Reveals an Interferon Gamma-Like Fold.* Biochemistry, 1995. **34**(38): p. 12118-12125.

59. Pestka, S., et al., *Interleukin-10 and related cytokines and receptors.* Annu Rev Immunol, 2004. **22**: p. 929-79.

60. Krugerkrasagakes, S., et al., *Expression of Interleukin-10 in Human-Melanoma.* British Journal of Cancer, 1994. **70**(6): p. 1182-1185.

61. Lech-Maranda, E., et al., *Interleukin-10 gene promoter polymorphisms influence the clinical outcome of diffuse large B-cell lymphoma.* Blood, 2004. **103**(9): p. 3529-3534.

62. Banko, Z., et al., *Induction and Differentiation of IL-10-Producing Regulatory B Cells from Healthy Blood Donors and Rheumatoid Arthritis Patients.* J Immunol, 2017. **198**(4): p. 1512-1520.

63. Tak, P.P., *IFN-beta in rheumatoid arthritis.* Front Biosci, 2004. **9**: p. 3242-7.

64. Choi, J., et al., *How Messenger RNA and Nascent Chain Sequences Regulate Translation Elongation.* Annual Review of Biochemistry, Vol 87, 2018. **87**: p. 421-449.

65. Somogyi, P., et al., *Ribosomal Pausing during Translation of an Rna Pseudoknot.* Molecular and Cellular Biology, 1993. **13**(11): p. 6931-6940.

66. Wen, J.D., et al., *Following translation by single ribosomes one codon at a time.* Nature, 2008. **452**(7187): p. 598-603.

67. Fischer, N., et al., *The pathway to GTPase activation of elongation factor SelB on the ribosome.* Nature, 2016. **540**(7631): p. 80-+.

68. Chavatte, L., B.A. Brown, and D.M. Driscoll, *Ribosomal protein L30 is a component of the UGA-selenocysteine recoding machinery in eukaryotes.* Nature Structural & Molecular Biology, 2005. **12**(5): p. 408-416.

69. Yusupova, G.Z., et al., *The path of messenger RNA through the ribosome.* Cell, 2001. **106**(2): p. 233-241.

70. Jenner, L.B., et al., *Structural aspects of messenger RNA reading frame maintenance by the ribosome.* Nature Structural & Molecular Biology, 2010. **17**(5): p. 555-U48.

71. Chen, C.L., et al., *Dynamics of translation by single ribosomes through mRNA secondary structures.* Nature Structural & Molecular Biology, 2013. **20**(5): p. 582-+.

72. Chen, C.L., et al., *Elongation factor G initiates translocation through a power stroke.* Proceedings of the National Academy of Sciences of the United States of America, 2016. **113**(27): p. 7515-7520.

73. Kozak, M., *Regulation of translation via mRNA structure in prokaryotes and eukaryotes.* Gene, 2005. **361**: p. 13-37.

74. Tuller, T., et al., *Translation efficiency is determined by both codon bias and folding energy.* Proceedings of the National Academy of Sciences of the United States of America, 2010. **107**(8): p. 3645-3650.

75. Del Campo, C., et al., *Secondary Structure across the Bacterial Transcriptome Reveals Versatile Roles in mRNA Regulation and Function.* PLoS Genet, 2015. **11**(10): p. e1005613.

76. Keller, T.E., et al., *Reduced mRNA secondary-structure stability near the start codon indicates functional genes in prokaryotes.* Genome Biol Evol, 2012. **4**(2): p. 80-8.

77. Katz, L. and C.B. Burge, *Widespread selection for local RNA secondary structure in coding regions of bacterial genes.* Genome Research, 2003. **13**(9): p. 2042-2051.

78. Zhou, T. and C.O. Wilke, *Reduced stability of mRNA secondary structure near the translation-initiation site in dsDNA viruses.* Bmc Evolutionary Biology, 2011. **11**.

79. Gu, W., T. Zhou, and C.O. Wilke, *A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes.* PLoS Comput Biol, 2010. **6**(2): p. e1000664.

80. SantaLucia, J., *A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics.* Proceedings of the National Academy of Sciences of the United States of America, 1998. **95**(4): p. 1460-1465.

81. Kudla, G., et al., *Coding-sequence determinants of gene expression in Escherichia coli.* Science, 2009. **324**(5924): p. 255-8.

82. de Smit, M.H. and J. van Duin, *Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis.* Proc Natl Acad Sci U S A, 1990. **87**(19): p. 7668-72.

83. López, J.L., et al., *Codon Usage Optimization in the Prokaryotic Tree of Life: How Synonymous Codons Are Differentially Selected in Sequence Domains with Different Expression Levels and Degrees of Conservation.* Mbio, 2020. **11**(4).

84. Plotkin, J.B. and G. Kudla, *Synonymous but not the same: the causes and consequences of codon bias.* Nat Rev Genet, 2011. **12**(1): p. 32-42.

85. Ikemura, T., *Codon usage and tRNA content in unicellular and multicellular organisms.* Mol Biol Evol, 1985. **2**(1): p. 13-34.

86. Bhandari, B.K., C.S. Lim, and P.P. Gardner, *TISIGNER.com: web services for improving recombinant protein production.* Nucleic Acids Res, 2021. **49**(W1): p. W654-W661.

87. Looman, A.C., et al., *Influence of the codon following the AUG initiation codon on the expression of a modified lacZ gene in Escherichia coli.* EMBO J, 1987. **6**(8): p. 2489-92.

88. Charneski, C.A. and L.D. Hurst, *Positively charged residues are the major determinants of ribosomal velocity.* PLoS Biol, 2013. **11**(3): p. e1001508.

89. Looman, A.C., et al., *Influence of the Codon Following the Aug Initiation Codon on the Expression of a Modified Lacz Gene in Escherichia-Coli.* Embo Journal, 1987. **6**(8): p. 2489-2492.

90. Voges, D., et al., *Analyzing and enhancing mRNA translational efficiency in an Escherichia coli in vitro expression system.* Biochemical and Biophysical Research Communications, 2004. **318**(2): p. 601-614.

91. Lu, J.L. and C. Deutsch, *Electrostatics in the Ribosomal Tunnel Modulate Chain Elongation Rates.* Journal of Molecular Biology, 2008. **384**(1): p. 73-86.

92. Tenson, T. and M. Ehrenberg, *Regulatory nascent peptides in the ribosomal tunnel.* Cell, 2002. **108**(5): p. 591-594.

93. Zhang, J., et al., *Mechanisms of ribosome stalling by SecM at multiple elongation steps.* Elife, 2015. **4**.

94. Lu, J.L., W.R. Kobertz, and C. Deutsch, *Mapping the electrostatic potential within the ribosomal exit tunnel.* Journal of Molecular Biology, 2007. **371**(5): p. 1378-1391.

95. Ito, K. and S. Chiba, *Arrest Peptides: Cis-Acting Modulators of Translation.* Annual Review of Biochemistry, Vol 82, 2013. **82**: p. 171-202.

96. Berezovsky, I.N., et al., *Amino acid composition of protein termini are biased in different manners.* Protein Engineering, 1999. **12**(1): p. 23-30.

97. Tuller, T., et al., *Composite effects of gene determinants on the translation speed and density of ribosomes.* Genome Biology, 2011. **12**(11).

98. Tuller, T., et al., *An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation.* Cell, 2010. **141**(2): p. 344-354.

99. Melnikov, S., et al., *Molecular insights into protein synthesis with proline residues.* EMBO Rep, 2016. **17**(12): p. 1776-1784.

100. Requiao, R.D., et al., *Protein charge distribution in proteomes and its impact on translation.* PLoS Comput Biol, 2017. **13**(5): p. e1005549.

101. Charneski, C.A. and L.D. Hurst, *Positive Charge Loading at Protein Termini Is Due to Membrane Protein Topology, Not a Translational Ramp.* Molecular Biology and Evolution, 2014. **31**(1): p. 70-84.

102. Ikemura, T., *Correlation between the Abundance of Escherichia-Coli Transfer-Rnas and the Occurrence of the Respective Codons in Its Protein Genes - a Proposal for a Synonymous Codon Choice That Is Optimal for the Escherichia-Coli Translational System.* Journal of Molecular Biology, 1981. **151**(3): p. 389-409.

103. Ohno, H., et al., *Preferential usage of some minor codons in bacteria.* Gene, 2001. **276**(1-2): p. 107-115.

104. Grantham, R., et al., *Codon Catalog Usage Is a Genome Strategy Modulated for Gene Expressivity.* Nucleic Acids Research, 1981. **9**(1): p. R43-R74.

105. dos Reis, M., R. Savva, and L. Wernisch, *Solving the riddle of codon usage preferences: a test for translational selection.* Nucleic Acids Research, 2004. **32**(17): p. 5036-5044.

106. Pop, C., et al., *Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation.* Molecular Systems Biology, 2014. **10**(12).

107. Chen, G.F.T. and M. Inouye, *Suppression of the Negative Effect of Minor Arginine Codons on Gene-Expression - Preferential Usage of Minor Codons within the 1st 25 Codons of the Escherichia-Coli Genes.* Nucleic Acids Research, 1990. **18**(6): p. 1465-1473.

108. Ikemura, T., *Codon Usage and Transfer-Rna Content in Unicellular and Multicellular Organisms.* Molecular Biology and Evolution, 1985. **2**(1): p. 13-34.

109. Zamora-Romo, E., et al., *Efficient expression of gene variants that harbour AGA codons next to the initiation codon.* Nucleic Acids Research, 2007. **35**(17): p. 5966-5974.

110. de Valdivia, E.I.G. and L.A. Isaksson, *A codon window in mRNA downstream of the initiation codon where NGG codons give strongly reduced gene expression in Escherichia coli.* Nucleic Acids Research, 2004. **32**(17): p. 5198-5205.

111. Pai, V.R. and V. Murugan, *A synergistic effect of suppressive CGG codon in+2 position and downstream CAT repeats for efficient heterologous protein expression in Escherichia coli.* Biochemical and Biophysical Research Communications, 2008. **371**(3): p. 380-384.

112. Mitarai, N., K. Sneppen, and S. Pedersen, *Ribosome collisions and translation efficiency: Optimization by codon usage and mRNA destabilization.* Journal of Molecular Biology, 2008. **382**(1): p. 236-245.

113. Ojima-Kato, T., et al., *Nascent MSKIK peptide cancels ribosomal stalling by arrest peptides in Escherichia coli.* Journal of Biological Chemistry, 2023. **299**(5).

114. Komar, A.A., T. Lesnik, and C. Reiss, *Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation.* FEBS Lett, 1999. **462**(3): p. 387-91.

115. Young, J.C. and D.W. Andrews, *The signal recognition particle receptor alpha subunit assembles co-translationally on the endoplasmic reticulum membrane during an mRNA-encoded translation pause in vitro.* Embo Journal, 1996. **15**(1): p. 172-181.

116. Crameri A., W.E.A., Tate E., Stemmer W. P. C., *Improved Green Fluorescence Protein by Molecular Evolution Using DNA Shuffling.* Nature Biotechnology, 1996. **14**: p. 5.

117. Fukuda, H., M. Arai, and K. Kuwajima, *Folding of green fluorescent protein and the cycle3 mutant.* Biochemistry, 2000. **39**(39): p. 12025-32.

118. Waterhouse, A., et al., *SWISS-MODEL: homology modelling of protein structures and complexes.* Nucleic Acids Res, 2018. **46**(W1): p. W296-W303.

119. Adan, A., et al., *Flow cytometry: basic principles and applications.* Critical Reviews in Biotechnology, 2017. **37**(2): p. 163-176.

120. McKinnon, K.M., *Flow Cytometry: An Overview.* Curr Protoc Immunol, 2018. **120**: p. 5 1 1-5 1 11.

121. Horejsi, Z., et al., *Phosphorylation-dependent PIH1D1 interactions define substrate specificity of the R2TP cochaperone complex.* Cell Rep, 2014. **7**(1): p. 19-26.

122. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.

123. Weber, M., et al., *Impact of C-terminal amino acid composition on protein expression in bacteria.* Mol Syst Biol, 2020. **16**(5): p. e9208.

124. Bjornsson, A., S. Mottagui-Tabar, and L.A. Isaksson, *Structure of the C-terminal end of the nascent peptide influences translation termination.* EMBO J, 1996. **15**(7): p. 1696-704.

125. Morowvat, M.H., et al., *Optimization of Fermentation Conditions for Recombinant Human Interferon Beta Production by Escherichia coli Using the Response Surface Methodology.* Jundishapur J Microbiol, 2015. **8**(4): p. e16236.

126. Sottini, A., et al., *Interferon-beta therapy monitoring in multiple sclerosis patients.* Endocr Metab Immune Disord Drug Targets, 2009. **9**(1): p. 14-28.

127. Lim, S.Y. and C.S. Constantinescu, *Current and future disease-modifying therapies in multiple sclerosis.* Int J Clin Pract, 2010. **64**(5): p. 637-50.

128. Toporovski, R., M.P. Morrow, and D.B. Weiner, *Interferons as potential adjuvants in prophylactic vaccines.* Expert Opin Biol Ther, 2010. **10**(10): p. 1489-500.

129. Bracci, L., et al., *Type I interferons as vaccine adjuvants against infectious diseases and cancer.* Expert Rev Vaccines, 2008. **7**(3): p. 373-81.

130. Smith, D.B., and K S Johnson, *Single-step purification of polypeptides expressed in Escherichia coli as fusions with glutathione S-transferase.* Gene, 1998. **67**: p. 31-40.

131. Zhan, Y., X. Song, and G.W. Zhou, *Structural analysis of regulatory protein domains using GST-fusion proteins.* Gene, 2001. **281**(1-2): p. 1-9.

132. Korkmaz, G., et al., *Comprehensive Analysis of Stop Codon Usage in Bacteria and Its Correlation with Release Factor Abundance.* Journal of Biological Chemistry, 2014. **289**(44): p. 30334-30342.

133. de Lorenzo, V., *The principle of uncertainty in biology: Will machine learning/artificial intelligence lead to the end of mechanistic studies?* PLoS Biol, 2024. **22**(2): p. e3002495.

7 Attachments

This section includes two scientific articles that present the results discussed in this thesis.

1) Pham, P.N., et al., *Protein Binder (ProBi) as a New Class of Structurally Robust Non-Antibody Protein Scaffold for Directed Evolution.* Viruses, 2021. **13**(2).

2) Herynek, S., et al., *Increasing recombinant protein production in E. coli via FACS-based selection of N-terminal coding DNA libraries.* FEBS J, 2025. **292**(5): p. 1070-1085.

viruses

MDPI

Article

Protein Binder (ProBi) as a New Class of Structurally Robust Non-Antibody Protein Scaffold for Directed Evolution

Phuong Ngoc Pham, Maroš Huličiak, Lada Biedermannová, Jiří Černý, Tatsiana Charnavets, Gustavo Fuertes, Štěpán Herynek, Lucie Kolářová, Petr Kolenko, Jiří Pavlíček, Jiří Zahradník, Pavel Mikulecky * and Bohdan Schneider *

Institute of Biotechnology of the Czech Academy of Sciences, BIOCEV, CZ-25250 Vestec, Czech Republic; phuong@ibt.cas.cz (P.N.P.); maros.huliciak@ibt.cas.cz (M.H.); Lada.Biedermannova@ibt.cas.cz (L.B.); jiri.cerny@ibt.cas.cz (J.Č.); tatsiana.charnavets@ibt.cas.cz (T.C.); gustavo.fuertes@ibt.cas.cz (G.F.); Stepan.Herynek@ibt.cas.cz (Š.H.); kolarovalucie007@gmail.com (L.K.); kolenko@ibt.cas.cz (P.K.); jiri.pavlicek@ibt.cas.cz (J.P.); jiri.zahradnik@ibt.cas.cz (J.Z.)

* Correspondence: mikulecky@outlook.com (P.M.); bohdan.schneider@gmail.com (B.S.)

Abstract: Engineered small non-antibody protein scaffolds are a promising alternative to antibodies and are especially attractive for use in protein therapeutics and diagnostics. The advantages include smaller size and a more robust, single-domain structural framework with a defined binding surface amenable to mutation. This calls for a more systematic approach in designing new scaffolds suitable for use in one or more methods of directed evolution. We hereby describe a process based on an analysis of protein structures from the Protein Data Bank and their experimental examination. The candidate protein scaffolds were subjected to a thorough screening including computational evaluation of the mutability, and experimental determination of their expression yield in *E. coli*, solubility, and thermostability. In the next step, we examined several variants of the candidate scaffolds including their wild types and alanine mutants. We proved the applicability of this systematic procedure by selecting a monomeric single-domain human protein with a fold different from previously known scaffolds. The newly developed scaffold, called ProBi (Protein Binder), contains two independently mutable surface patches. We demonstrated its functionality by training it as a binder against human interleukin-10, a medically important cytokine. The procedure yielded scaffold-related variants with nanomolar affinity.

Keywords: directed evolution; protein scaffold; protein engineering; computational saturation; ribosome display; interleukin-10

check for updates

Citation: Pham, P.N.; Huličiak, M.; Biedermannová, L.; Černý, J.; Charnavets, T.; Fuertes, G.; Herynek, Š; Kolářová, L.; Kolenko, P.; Pavlíček, J.; et al. Protein Binder (ProBi) as a New Class of Structurally Robust Non-Antibody Protein Scaffold for Directed Evolution. *Viruses* **2021**, *13*, 190. https://doi.org/10.3390/v13020190

Academic Editor: Petri Susi
Received: 19 December 2020
Accepted: 23 January 2021
Published: 27 January 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Many biological functions depend on specific protein-protein interactions. Protein engineering offers the possibility to tune these interactions by developing de novo binding partners [1–3] or by mutating the interaction partners using computational design [4,5]. A powerful tool of protein engineering is to generate protein binders by the in vitro directed selection techniques [6–8] or to use evolution-based approaches to increase the stability of recombinant proteins [9–11].

A prototype of a binding protein is an antibody, a highly selective and adaptive molecule capable of binding to a huge spectrum of partners. The antibody-based binders have been indispensable not only in research experiments but also in clinical trials and tens of them are biologicals approved by Federal Drug Administration. However, several suboptimal properties of these molecules such as their large size (~150 kDa), cross-reactivity, and necessity of animal immunization during the preparation, motivated the development of binders with alternative structures [12,13]. These novel artificial high-affinity binders are called "small non-antibody protein scaffolds" [14–16]. The advantage of scaffolds over antibodies is their typically higher stability and lower cost of production. The most widely

used scaffolds are designed Ankyrin repeats (dARPins), lipocalin domain (anticalins), and Ig-binding domain of Staphylococcal Protein A (affibodies) [17–19]. Some of the small protein scaffolds were already proven to be useful in a wide range of applications, from academic research to clinical imaging, diagnostics, or therapeutics [20–23], and others are even in clinical trials [24,25]. In our previous work, we used the albumin-binding domain (ABD) of Streptococcal Protein G as a scaffold protein. We engineered its binding against human IFNγ [26] and interleukin-23 receptor [27] using computational design and ribosome display.

The exploration of non-antibody protein scaffolds is still a young field compared to the research of antibodies and the full potential of scaffolds has not yet been uncovered. Therefore, we are convinced that there is ample space to expand the current portfolio of protein scaffolds so that they can complement antibodies and interact with new target molecules. Therefore, a systematic method to streamline the development of new scaffolds with desired properties would be beneficial but relatively few attempts can be found in the literature [18,28,29]. In general, a good small protein scaffold should display a high stability, together with flexibility of binding. The desired properties thus include tolerance to mutations during diversification, thermostability, in vivo integrity, and non-immunogenicity. Moreover, the variants of the protein scaffold must be efficiently evolvable to achieve high affinity and specificity towards target proteins. Therefore, the properties of the potential small protein scaffold must be thoroughly evaluated.

In this work, we present a systematic stepwise procedure for the selection of new and structurally robust protein scaffolds. The procedure starts with a wide search of suitable structures in the Protein Data Bank (PDB) [30], followed by a series of computational and experimental screens to select the most robust candidates as novel, small protein scaffolds. We demonstrate the applicability of our method by selecting a protein scaffold called ProBi (Protein Binder) that fulfills all the criteria of stability and mutability, and comprises not one, but two independently mutable surface patches.

We tested the applicability of the ProBi scaffold by generating its variants with high affinity to a target protein using in vitro directed evolution techniques. As the target, we selected a medically important protein, interleukin-10 (IL-10), a member of the cytokine superfamily. IL-10 is an immune repressor [31], with numerous investigations suggesting its major impact in inflammatory, malignant, and autoimmune diseases. IL-10 overexpression was found in certain tumors and is considered to promote tumor development [32]. Also, some viral homologs can bind to the IL-10 receptor 1 [33] and mimic the immunosuppressive effects of IL-10. Thus, proteins binding IL-10 and interfering with its signaling have a great potential for diagnostic and therapeutic purposes.

To generate ProBi-based binders to IL-10, we created a DNA library by randomization of 10 amino acids on one of the ProBi surface patches and used the ribosome display technique to screen for ProBi variants with the highest affinity. We successfully generated binders of the newly developed ProBi protein scaffold with affinities of ~10 and ~200 nM against IL-10.

2. Materials and Methods

2.1. Selection of Suitable Protein Structures from the Protein Data Bank

The Protein Data Bank (PDB) [30] was searched for potential protein scaffolds using the following criteria: (1) small size (molecular weight in range of 10–25 kDa); (2) monomeric protein with a reasonably high-resolution X-ray structure (<3.0 Å); and (3) produced in Expression Organism of *E. coli* for ease of expression. Proteins matching these initial search criteria were then manually curated to meet additional rules: structure (protein fold class) (1) should differ from the previously published protein scaffolds; (2) should differ from the other selected structures in the set; (3) should contain a low number of cysteine residues and disulfide bonds, (4) should not be toxic or immunogenic, (5) should not contain any ligand or cofactor, and (6) should be soluble and easy to purify (according to the source literature).

2.2. In Silico Identification of Mutable Surface Patches on Protein Scaffolds Candidates

(a) Multiple sequence alignment. For each of the scaffold candidate proteins, a similarity search was performed for the related amino acid sequences from different species using the UniProt BLAST service (www.uniprot.org/blast) with default parameters (Target database = UniProtKB, E-Threshold = 10, Matrix = blosum62, Filtering = None, Gapped = yes, Hits = 250). The identical sequences were subsequently removed using a script and the remaining ones were aligned using Clustal W [34] in the UGENE program [35]. The conservation of each sequence position was calculated.

(b) In silico saturation mutagenesis. The mutability of the surface amino acid residues was evaluated by in silico mutation scanning using the FoldX program [36] by applying the "positionscan" FoldX Keywords: In all proteins, each position was substituted by all of the 20 standard amino acid residues, and the corresponding free energy differences ($\Delta\Delta G$) between the mutant and the wild-type (WT) structure were calculated; the calculations included self-mutations leading to $\Delta\Delta G = 0$. Positions in which most of the mutations were stabilizing ($\Delta\Delta G < 0$), or only slightly destabilizing ($\Delta\Delta G < 0.5$ kcal/mol), were identified and the mutability score for each position was calculated as a percentage of mutations fulfilling these criteria.

(c) Sequence and structural evaluation. The mutable residues, identified by multiple sequence alignment and in silico saturation mutagenesis, were located on the 3D structure of each protein scaffold candidate. We then visually searched for continuous surface regions consisting of mutable residues. The regions where the mutable residues were close together in both sequential and structural space we termed the mutable surface patch. Ideally, the patch should contain 10–12 residues to fulfill the theoretical complexity suitable for ribosome display. If the identified patch contained less than 10–12 residues with the highest mutability score, we completed the patch by a small number of neighboring surface residues with slightly lower mutability scores.

2.3. Production of Proteins

(a) Cloning of recombinant proteins. The genes coding the selected five protein scaffold candidates–4PSF, 1N3Y, 4I3B, 1W2I, and 2F3L, were ordered in the form of DNA strings (ThermoFisher Scientific, Waltham, MA, USA) with codons optimized for expression in *Escherichia coli* but without a stop codon. The DNA strings were cloned into the pET-26b(+) vector using NdeI and XhoI restriction enzymes in frame with C-terminal His-tag. Also, the DNA strings were used as templates for polymerase chain reaction (PCR) to add the N-terminal His-tag and stop codon. These N-terminally tagged proteins were cloned into the pET-26b(+) vector using NdeI and XhoI restriction enzymes. Final constructs were labeled as pET26b-4PSF-WT, etc.

The genes coding the variants of three protein scaffold candidates–4PSF, 1N3Y, and 4I3B, with all mutable residues changed to alanine amino acid, referred to as allAla mutants, were ordered in the form of DNA strings with N-terminal His-tag and C-terminal stop codon. The DNA strings were digested by NdeI and XhoI restriction enzymes and cloned into the pET-26b(+) vector. Final constructs were labeled as pET26b-4PSF-allAla, etc.

The ProBi-WT protein and its variants were amplified by the PCR method using the ProBi-cloning-for and ProBi-cloning-rev primers (Table S1). The created fragments were cleaved by NcoI and BamHI restriction enzymes and cloned into the pRDVsm or pETsm vectors (Figure S1).

The genes encoding the interleukin-10 (residues 19–179 of UniProt entry P22301) and interleukin-29 (residues 20–200 of UniProt entry Q8IU54) were ordered in the form of a DNA string with codons optimized for expression in *Drosophila melanogaster*. The DNA was cleaved by BglII and XhoI restriction enzymes and cloned into a pMTH vector (a modified version of pMT-BiP-V5-His_A vector where recognition site for the AgeI restriction enzyme is exchanged for XhoI enzyme). The final construct contained IL-10 or IL-29 protein with N-terminal BiP signal peptide and C-terminal His-tag. Final constructs were labeled as pMTH-IL10cd and pMTH-IL29cd.

(b) Expression and purification of recombinant proteins. The recombinant protein scaffold candidates were produced in *E. coli* BL21(DE3) strain. In addition, the expression of 1W3I scaffold was tested in *E. coli* C41(DE3) and C43(DE3) strains. The bacteria were cultivated in LB medium (Sigma-Aldrich, St. Louis, MO, USA) at 37 °C until OD600 = 0.6, followed by the addition of 1 mM isopropyl-beta-D-thiogalactopyranoside (IPTG) for induction of expression. Then the cultivation continued for 4 h at 37 and 30 °C or overnight at 25 and 16 °C. The cells were harvested by centrifugation (5000× *g*, 15 min, 4 °C) and stored at −20 °C. The cells were resuspended in LH buffer (20 mM Na-Phosphate buffer, pH 7.3, 100 mM NaCl), and disrupted by ultrasound. The soluble fraction was separated by centrifugation (40,000× *g*, 20 min, 4 °C) and passed over the Ni-NTA agarose beads equilibrated in LH buffer. The beads were washed by LH buffer supplemented by 10 mM imidazole and proteins were eluted by LH buffer containing 200 mM imidazole.

The recombinant ProBi variants were produced in *E. coli* BL21-Gold(DE3) in LB medium (Sigma-Aldrich) at 37 °C until OD600 = 0.6, followed by the addition of 1 mM IPTG for induction of expression. Then cultivation continued overnight at 18 °C. The cells were harvested by centrifugation (5000× *g*, 15 min, 4 °C) and stored at −20 °C. The cells were resuspended in LS buffer (50 mM Tris, pH 8.0, 150 mM NaCl, 1 mM EDTA), and disrupted by ultrasound. The soluble fraction was separated by centrifugation (40,000× *g*, 20 min, 4 °C) and passed over the StrepTactin XT resin equilibrated in LS buffer. The resin was washed by LS buffer and proteins were eluted by BXT buffer (100 mM Tris, pH 8.0, 150 mM NaCl, 1 mM EDTA, 50 mM biotin).

The constructs pMTH-IL10cd and pMTH-IL29cd for expression of recombinant IL-10 and IL-29, respectively, in insect S2 cells were produced and purified similarly as IFNgR2 in the previous report [37]. Briefly, S2 cells were cultivated in Insect-XPRESSTM Protein-free Insect Cell Medium (Lonza) and protein expression was induced by the addition of CuSO4. IL-10 and IL-29 proteins were secreted into the medium and stored at −80 °C.

All proteins were further purified to homogeneity by size exclusion chromatography at 16 °C using Superdex 75 10/300 Increase, Superdex 200 10/300 Increase, or Superdex 75 16/600 column (GE Healthcare, Chicago, IL, USA). The column was equilibrated in phosphate-buffered saline (PBS) pH 7.4 (for protein scaffold proteins and all Ala mutants), 20 mM Tris, pH 7.5, 100 NaCl (for preliminary crystallization trials), or in 50 mM Tris, pH 8.0, 300 mM NaCl (for ProBi variants and IL-10 protein). Samples were analyzed by 15% sodium dodecyl sulphate–polyacrylamide gel electrophoresis (SDS-PAGE).

2.4. Biophysical Characterization of Recombinant Proteins

(a) Oligomerization and melting temperature of protein scaffold candidates and allAla mutants. The oligomeric state of proteins and their monodispersity was measured by dynamic light scattering (DLS) using a ZetaSizer Nano ZS90 (Malvern Panalytical, Malvern, UK) instrument and quartz cuvette ZEN2112. The secondary structure and folding of proteins were analyzed by circular dichroism (CD) spectra using a Chirascan Plus spectrometer (Applied Photophysics, Surrey, UK) in steps of 1 mm over the wavelength range 200–260 nm. Samples were diluted with PBS buffer, pH 7.4, to a concentration of 0.2 mg/mL. The individual spectra were measured in a 0.01 cm path-length quartz cell at a temperature of 25 °C. The CD signal was recorded as the ellipticity, and the resulting spectra were buffer-subtracted. The melting temperatures of proteins were evaluated using either the CD spectrometer or nanoDSF technology. The CD melting measurements were performed by using the 0.1 cm path-length quartz cell. The samples were heated from 20 to 85 °C, and the sample absorption was recorded at 280 nm in 1 °C increments at a rate of 0.5 °C/min. The nanoDSF technology was implemented in the Prometheus NT.48 instrument (NanoTemper, München, Germany). The samples were loaded into standard capillaries and heated from 20 to 95 °C at a rate of 1 °C/min. The melting temperatures (Tm) were estimated from the first derivative of the melting curves.

(b) Testing protein suitability for affinity measurements. The applicability of protein scaffold candidates and allAla mutants for affinity measurements using the microscale

thermophoresis (MST) and surface plasmon resonance (SPR) was tested. For MST, proteins were labeled using the Monolith Protein Labeling Kit RED-NHS (Amine Reactive) according to the manufacturer's protocol. The samples were loaded into three types of capillary (standard, hydrophilic, and hydrophobic) and inserted into the Monolith NT.115 instrument (NanoTemper). The capillary scan was performed immediately, and then samples were incubated in the instrument for 1 h and re-measured. An alternative approach was implemented for SPR. The GLC sensor chip was inserted into the ProteOn XPR36 instrument, washed but not activated. The samples in different channels were passed over the un-activated chip to estimate the non-specific binding of proteins to the surface of the chip itself. The running buffer (PBS, pH 7.4, 0.005% Tween-20) was used in the control channel.

2.5. Crystallization and Structure Determination

The crystallization solution included J61 at a concentration of 10 mg/mL in a buffer containing 100 mM NaCl, 20 mM Tris, pH 8.0. Crystals were prepared using the hanging-drop vapor-diffusion method with the reservoir solution containing 5% (*v*/*v*) glycerol and 4 M sodium formate. Crystals required cryo-protection with 20% glycerol before flash freezing in liquid nitrogen and the diffraction experiment.

The final diffraction data were collected at the Helmholz–Zentrum Berlin (Bessy II) on a beamline 14.1 at 100 K. The diffraction images were processed using the XDS program [38] and scaled using the AIMLESS program [39] from the CCP4 program package [40]. The phase problem was solved by molecular replacement using PHASER [41] using monomeric 4PSF structure [42] as a search model. The structure was refined using the REFMAC5 program [43] using 95% of reflections as a working set and 5% or reflections as a test set. Manual corrections to the model were done using the graphic COOT program [44]. The last run of several cycles of refinement was performed using all reflections and anisotropic ADPs. The structure quality was checked using the validation tools implemented in CCP4, MOLPROBITY [45], and the PDB. Final data processing statistics and structure refinement statistics are shown in Table S2.

2.6. Construction of DNA Library of ProBi Scaffold (PatchC)

The earlier design of the DNA library [46] was adopted with small changes regarding restriction enzyme recognition sites and N- and C-terminal tags. Our ribosome display DNA construct contained a T7 promoter, 5′ stem-loop, ribosome binding site, N-terminal Strep-tag, open reading frame encoding the ProBi scaffold library, and C-terminal c-Myc-tag fused to TolA-spacer without stop codon. The library of ProBi scaffold (PatchC) was prepared by the GENEWIZ company using NNK codons technology. We selected ten residues to be randomized. The initial synthetic DNA library was amplified by PCR and then used in the first selection round of ribosome display.

2.7. Selection of Novel Binders by In Vitro Ribosome Display

The previous protocols from Pluckthun's laboratory regarding the in vitro selection by ribosome display [46,47] were adopted and adjusted. All the following steps were first performed with the ProBi-WT scaffold to establish and verify the protocol.

(a) Preparation of the mRNA-ribosome-protein (MRP) complex. The DNA library was amplified by PCR (supplemented by 6% DMSO) with primers T7b and TolAk (Table S1) using Q5 polymerase (New England Biolabs, Ipswich, MA, USA) with an annealing temperature of 66 °C. As a template, either the starting DNA library in the first selection round was used or the sorted DNA library cloned into the pRDVsm vector in the next rounds. The resulted amplicon contained a T7 promoter, 5′ stem-loop, ribosome binding site (RBS), start codon, N-terminal Strep-tag II, scaffold library, C-terminal c-Myc-tag, tolA spacer, and 3′ stem-loop without stop codon. The amplicon was purified from 1% agarose gel and used in the following cell-free protein synthesis step. We utilized RTS 100 *E. coli*

HY Kit (Biotechrabbit) according to manufacturer recommendations to create the MRP complex composed of mRNA, ribosome, and protein scaffold variants.

(b) Ribosome display selection rounds. The Nunc PolySorp plate (Invitrogen, Carlsbad, CA, USA) was coated with anti-His antibody (His-tag Monoclonal Antibody (HIS.H8) by Invitrogen) diluted to a final concentration of 25 μg/mL in bicarbonate coating buffer (pH 9.6). All following steps were undertaken in the TBS buffer, pH 7.4. Blocking of plate and incubation of scaffold variants with proteins were undertaken either by an initial preselection process or an optimized preselection process described below.

(c) Initial preselection process. The plate was blocked by 3% bovine serum albumin (BSA) during all rounds. The MRP complex supplemented by 0.5% BSA and heparin (200 mg/mL) was added to the Preselection well that was coated with a mixture of 3% BSA, interferon-gamma (IFNg, 25 μg/mL), and the 4PSF wild type protein (4PSF-WT, 25 μg/mL). The reaction was incubated for 1 h at 4 °C.

(d) Optimized preselection process. The plate was blocked by either 3% BSA (used in the first and third round) or by 3% dry skimmed cow milk (used in the second and fourth round). The MRP complex supplemented by 0.5% BSA and heparin (200 mg/mL) was added to the Preselection well that was coated with a mixture of 3% BSA, IFNg (25 μg/mL), IL-29 (25 μg/mL), and 4PSF-WT (25 μg/mL). The reaction was incubated for 1 h at 4 °C.

(e) Selection process. The preselected MRP complex was transferred into the Selection well that contained IL-10 (25 μg/mL) as a target molecule and incubated again for 1 h at 4 °C. In each round, the well was washed by different cycles of washes and increasing concentration of Pluronic F-127 in TBS buffer. The conditions were as follows: First round–5 wash cycles with 0.10%, second round–10 wash cycles with 0.15%, third and fourth-round–10 wash cycles with 0.20%. The library complex was incubated with an EB buffer (50 mM Tris-Acetate, 150 mM NaCl, 50 mM EDTA, pH adjusted to 7.5) containing S. cerevisiae RNA (1 mg/mL) and heparin (200 mg/mL) to release the mRNA. The mRNA was purified by the High Pure RNA Isolation Kit (Roche, Basel, Switzerland) according to the manufacturer's instructions and used in the GoScript Reverse Transcription System (Promega, Dane County, WI, USA) with ProBi-cloning-rev primer according to the manufacturer's instruction. The created cDNA was amplified by PCR (supplemented by 6% DMSO) with primers ProBi-cloning-for and ProBi-cloning-rev using Q5 polymerase (New England Biolabs). The resulted amplicon was cloned into the pRDVsm vector using NcoI and BamHI restriction enzymes and T4 ligase (New England Biolabs). The pRDVsm-ProBi was used in the following preselection and selection process. After the fourth round, the amplicon was cloned into the pETsm vector and transformed into BL21(DE3)-Gold (Agilent, Santa Clara, CA, USA) for affinity estimation by enzyme-linked immunosorbent assay (ELISA) and characterization.

2.8. Characterization of Novel Binders

(a) ELISA assay. The Nunc PolySorp plate (Invitrogen) was coated with anti-His antibody (His-tag Monoclonal Antibody (HIS.H8) by Invitrogen) diluted to a final concentration of 5 μg/mL in bicarbonate coating buffer (pH 9.6). All following steps were done in the PBS-P buffer (PBS buffer, pH 7.4, 0.1% Pluronic F-127). For the first evaluation of the binding of 190 variants, the wells were blocked by 1% BSA, then IL-10 (10 μg/mL) was attached as a target molecule, and samples were applied in the form of cell lysates. For initial specificity mapping of 47 variants, the wells were blocked by 3% dry skim milk, 3% BSA, or 1% BSA. The wells with 1% BSA were further incubated with either IL-10 or IL-29 (10 μg/mL), and samples were applied in the form of cell lysates. For affinity estimation of 10 selected variants, the wells were blocked by 1% BSA and then incubated with IL-10 (10 μg/mL) as a target molecule. Then the samples in the form of proteins purified on StrepTactinXT were applied. Finally, the binders were detected by horseradish peroxidase (HRP)-conjugated antibody against the C-terminal c-Myc-tag and TMB-2 as substrate.

(b) Inhibitory assay. Murine RAW264.7 cells were cultured in DMEM medium supplemented with 10% FBS. The mixture containing IL-10 (final concentration of 10 ng/mL)

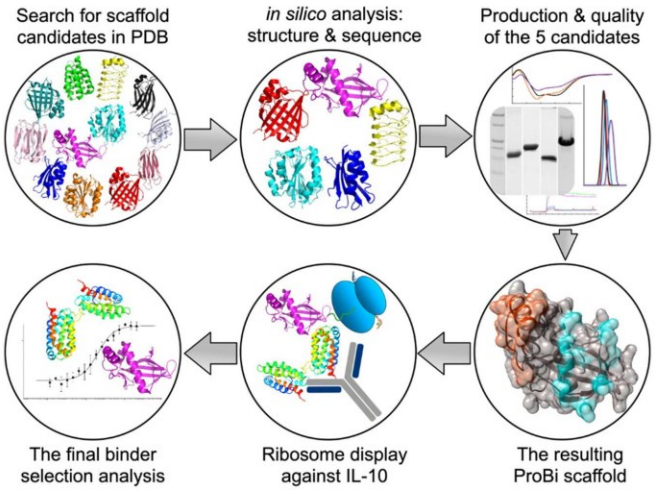
and ProBi variants (concentrations ranging from 160 ng to 100 µg/mL) was added to the cells and incubated for 30 min. The plates with cells were placed on ice, medium discarded, and washed with a cold PBS buffer. The cells were disrupted by a cold RIPA buffer supplemented by protease and phosphatase inhibitors. An amount of 20 µg of total protein was resolved using 10% Tris-Glycine SDS-PAGE gel and transferred to nitrocellulose membrane using Trans-Blot Turbo Transfer System according to the manufacturer's instructions. Membranes were blocked by 5% BSA in TBST buffer and incubated with antibodies against Y705-STAT3 (dilution 1/700) and alpha-tubulin (dilution 1/5000) diluted in 5% BSA in TBST buffer. Blots were washed in TBST buffer and incubated with appropriate HRP-conjugated secondary antibodies diluted in 1% BSA in TBST buffer. Membranes were treated with ECL solution (Merck Millipore, Burlington, MA, USA) and visualized on the Azure c600 instrument.

(c) Affinity measurements. The binding affinity in a solution was measured using the Monolith NT.115 instrument to monitor microscale thermophoresis (MST). The composition of the Assay buffer was 50 mM Tris, pH 8.0, 300 mM NaCl, and 0.1% Pluronic F-127. The IL-10 with C-terminal His-tag was labeled in assay buffer using the Monolith His-tag Labeling Kit RED-tris-NTA kit according to the manufacturer's instructions. The labeled IL-10 was titrated by ProBi variants diluted in an assay buffer and incubated for 10 min at room temperature. The reaction mix was loaded into standard capillaries and MST was monitored using Medium MST power and 60% of LED power. The analysis of interaction was done in MO.Affinity Analysis v2.3 software.

(d) Circular dichroism. The folding of new ProBi variants was measured by Circular dichroism spectra the same way as done with allAla mutants of protein scaffold candidates but with a final concentration of 0.1 mg/mL.

3. Results and Discussion

This work aimed to formulate and test a general procedure for selecting "small engineered non-antibody protein scaffolds" [15,16] suitable for generating artificial recombinant high-affinity binders by methods of targeted in vitro evolution. The systematic procedure for the development and testing of a new stable small protein scaffold presented in this work consists of two main parts comprising the steps graphically represented in Scheme 1 and summarized below:



Scheme 1. The procedure for the development of a new stable small protein scaffold.

(I) Development of a novel scaffold:

(i) Selection of suitable protein candidates from the Protein Data Bank.
(ii) In silico identification of mutable patches on the surfaces of the scaffold candidates.
(iii) Characterization of soluble wild-type variants of the scaffold candidates.
(iv) Testing the stability of the potential scaffolds.
(v) Testing the suitability of the scaffold candidate(s) for in vitro evolution.
(vi) Construction of the final ProBi scaffold.
(II) Application of the ProBi scaffold on the IL-10 target as a model system:
(i) Selection of novel ProBi-based binders by ribosome display.
(ii) Characterization of the novel ProBi-based binders.
(iii) Further development of scaffold selection and improvement of ProBi-based binders.

3.1. Development of a Novel Scaffold

3.1.1. Selection of Suitable Protein Structures from the Protein Data Bank

The first step of the procedure was to search for appropriate structures within the Protein Data Bank (PDB) database [30] performed by the members of the laboratory. The initial search was based on visual inspection of several hundred structures of small (less than 25 kDa) monomeric proteins solved by X-ray crystallography. The search results were limited to proteins known to be expressed in *Escherichia coli*. We finally selected about 100 structures that were further evaluated based on the literature data. Our selected candidates should (i) be new protein scaffolds, (ii) be based on different folds, (iii) preferably contain no or low number of cysteines, (iv) not to be toxic, (v) not to need any cofactor to maintain the structure, and (vi) be easy to purify.

We chose 12 proteins (Table 1) for closer structural examination and a more in-depth literature review. During this step, we excluded 2W4P [48] and 4LKT [49] because of structural similarities with other selected non-human candidates, 1W2I [50] and 4I3B [51], respectively. Furthermore, we eliminated 4MJJ and 4JOX because the manuscripts were not published at that time and we could not check their expression and purification protocols. To the best of our knowledge, none of the eight finally selected proteins had previously been used as a scaffold for directed evolution.

3.1.2. In Silico Identification of Mutable Surface Patches on the Protein Scaffold Candidates

On the surface of the eight candidate scaffolds, we looked for continuous patches consisting of residues that may be mutated without destabilizing the overall structure. To identify such mutable residues, we combined the analysis of protein sequence conservation with in silico saturation mutagenesis. We did not limit our search only for loop segments [28] because protein scaffolds could be mutated also at flat surfaces or combinations of loops and helices [18].

(a) Multiple sequence alignment. Sequences of protein homologs from different organisms for the eight protein scaffold candidates were aligned and the conservation of each amino acid position was calculated.

(b) In silico saturation mutagenesis. To evaluate the mutability of the residues, we performed in silico mutation scanning using the FoldX program [36]. Every amino acid position in each of the eight candidates was substituted by all of the 20 standard amino acid residues, and the corresponding free energy differences ($\Delta\Delta G$) between the mutant and the wild-type structure were calculated. Thus, each position in the protein sequence was characterized by 20 $\Delta\Delta G$ values. We identified spots in which most of the mutations were stabilizing ($\Delta\Delta G < 0$), or only slightly destabilizing ($0 < \Delta\Delta G < 0.5$ kcal/mol). We calculated the mutability score for each position as a percentage of mutations fulfilling these criteria.

(c) Sequence and structural evaluation. For the selection of mutable residues, we combined the evolutionary conservation and mutability scores. Positions with identity conservation <90% and FoldX mutability score >50% were considered mutable. Next, we mapped the mutable residues onto the protein 3D structures. We visually identified mutable positions and selected those constituting compact patches on the protein surface.

We preferred those patches consisting of residues that were not only structurally, but also sequentially close together, for ease of the subsequent library construction. To achieve a high variability of the designed DNA library, we selected 10–12 mutable residues, the numbers typically used in ribosome display experiments [57,58]. In cases where the patches contained less than 8–10 predicted mutable residues, we chose additional neighboring solvent-accessible residues to bring the total number of residues to 10.

Table 1. List of protein scaffold candidates that were selected for closer in silico structural analysis and literature review. The first five candidates were chosen for further experimental characterization.

PDB Code	UniProt	Protein Name	Source Organism	Size (kDa)	Size (aa)	Number of Cysteines	Reference
4PSF	Q9NWS0	PIH1D1 N-terminal domain	*Homo sapiens*	15	138	3	[42]
1N3Y	P20702	Alpha-X beta2 integrin I domain	*Homo sapiens*	22	198	0	[52]
4I3B	P0DM59	Fluorescent protein UnaG wild type	*Anguilla japonica*	15	139	0	[51]
2F3L	B1WVN5	Lumenal Rfr-domain protein	*Cyanothece* sp. 51142	20	184	2	[53]
1W2I	P84142	Acylphosphatase	*Pyrococcus horikoshii*	10	91	0	[50]
4NBO	Q9HD15	Steroid receptor RNA activator protein carboxy-terminal domain	*Homo sapiens*	13	111	2	[54]
3APA	O60844	Human pancreatic secretory protein ZG16p	*Homo sapiens*	16	141	0	[55]
4IGI	E9PWQ3	Collagen VI alpha3 N5 domain	*Mus musculus*	22	203	0	[56]
2W4P	P07311	Human common-type acylphosphatase variant, A99G	*Homo sapiens*	11	99	0	[48]
4LKT	Q01469	Human Epidermal Fatty Acid Binding Protein (FABP5)	*Homo sapiens*	15	138	6	[49]
4MJJ	Q14183	C2A domain of DOC2A	*Homo sapiens*	15	138	5	To be published
4JOX	Q939T0	Cry34Ab1 protein	*Bacillus thuringiensis*	14	123	0	To be published

In five candidate proteins, we found at least one mutable surface patch, in three of them even two independent patches (Figure 1). We excluded three proteins–4NBO [54], 3APA [55], and 4IGI [56], since they did not contain any suitable surface patch.

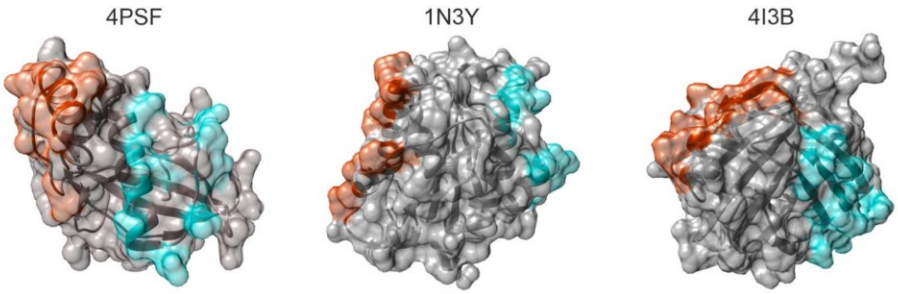


Figure 1. Surface patches of mutable residues identified in the structures of Protein Data Bank (PDB) IDs 4PSF [42], 1N3Y [52], and 4I3B [51]. The proteins are shown as ribbons in semitransparent surface representation. Residues constituting PatchN and PatchC in each protein are colored red and cyan, respectively.

3.1.3. Characterization of Soluble Wild-Type Variants of the Potential Scaffolds

We continued with five small protein scaffold candidates–4PSF [42], 1N3Y [52], 4I3B [51], 1W2I [50], and 2F3L [53]. We analyzed the expression level, solubility, purification simplicity, oligomerization state, secondary structure, and thermal stability of their wild-type (WT) variants.

(a) Protein expression and solubility. We prepared two versions of each scaffold candidate, first with N-terminal His-tag and second with C-terminal His-tag. We tested the expression and solubility of these five proteins in *E. coli* BL21(DE3) at four different temperatures, 37, 30, 25, and 16 °C. We detected high expression levels and excellent solubility of 4PSF, 1N3Y, 4I3B, and 2F3L with the N-terminal His-tag at 16 °C. The solubility varied with temperature. We were not able to express 4PSF, 1N3Y, and 4I3B proteins with the C-terminal His-tag. In the case of 1W2I, the proteins were not expressed in *E. coli* BL21(DE3) strain, only in *E. coli* C41(DE3) and C43(DE3) cells which are usually used for the expression of toxic proteins. Therefore, we decided to exclude the 1W2I protein.

(b) Protein purification. We continued with the purification of the remaining four proteins–4PSF, 1N3Y, 4I3B, and 2F3L, all of them as variants with the N-terminal His-tag. We performed a two-step purification procedure, first on Ni-NTA agarose beads and second by size exclusion chromatography. We isolated the 4PSF, 1N3Y, and 4I3B proteins at very high purity, but the 2F3L showed a minor band of either a homodimer or a non-relevant protein contaminant on SDS-PAGE (Figure 2A). The purified proteins retained solubility without aggregation and degradation at 4 °C in 20 mM sodium phosphate buffer at pH 7.5 and 100 mM NaCl in the range of days and weeks.

(c) Oligomerization. We used dynamic light scattering (DLS) to measure the oligomerization of purified proteins. The 4PSF, 1N3Y, and 4I3B proteins were monodisperse and monomeric based on the intensity-based particle size distribution with Z-averages of 4.6, 5.4, and 4.4 nm in diameter, respectively (Figure 2B). Although the 2F3L protein was monodisperse, the Z-average of 7.8 nm in diameter was higher than expected. It correlates with the impurities appearing on SDS-PAGE (Figure 2A). Since other than monomeric states of a protein would complicate biophysical measurements and binding assays, we decided to exclude 2F3L.

(d) Secondary structure. We checked the folding of the remaining three protein scaffold candidates, 4PSF, 1N3Y, and 4I3B, by circular dichroism (CD) measurements in the ultraviolet (UV) range (Figure 3). Based on the standard analysis of CD spectra using the CDNN software [59], the spectra of all three proteins showed features of regular secondary structures. In case of 4PSF and 1N3Y the combination of alpha-helices and beta-sheets, in case of 4I3B mainly of beta-sheets, agreeing qualitatively with the features of the respective crystal structures.

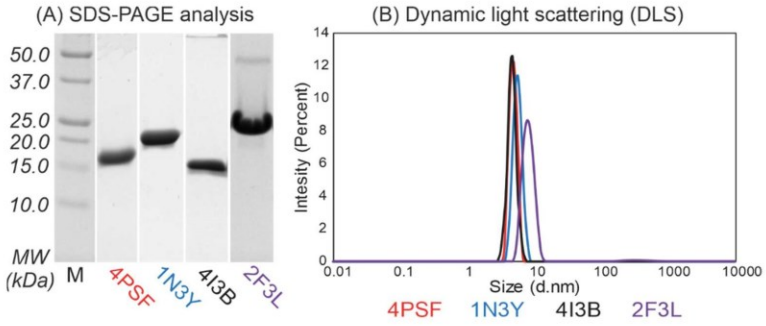


Figure 2. (**A**) Sodium dodecyl sulphate–polyacrylamide gel electrophoresis (SDS-PAGE, 15%) analysis and (**B**) dynamic light scattering (DLS) measurement of the purified protein scaffold candidates with N-terminal His-tags after size exclusion chromatography. The 4PSF, 1N3Y, and 4I3B proteins were isolated to high-purity as monomeric proteins. On the contrary, 2F3L showed minor contaminations around 45 kDa on SDS-PAGE gel and DLS measurement confirmed the presence of bigger particles.

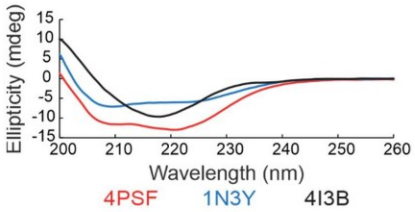


Figure 3. Far-ultraviolet (UV) circular dichroism (CD) spectra of three scaffold candidates. The results indicate that all three proteins acquire the expected secondary structure features. The 4PSF (red) and 1N3Y (blue) showed a combination of alpha-helices and beta-sheets, whereas the 4I3B (black) formed mainly beta-sheets.

(e) Thermal stability. We determined the thermal stability of proteins by measuring their melting temperature (Tm) using a circular dichroism technique. The melting temperatures of the WT variants of 4PSF, 1N3Y, and 4I3B were 75 °C, 57 °C, and 47 °C, respectively.

3.1.4. Testing the Stability of the Potential Scaffolds

Before starting a directed molecular evolution campaign, we experimentally verified the mutability of the predicted surface patches, for which we mutated all selected residues in each surface patch to alanines. The alanine-scanning mutagenesis is a method used to identify the structural role of protein residues [60–62]. The idea was that the protein scaffold candidate should sustain the expression, solubility, and folding upon forming a large relatively hydrophobic surface patch of alanines.

Each of three small protein scaffold candidates, 4PSF, 1N3Y, and 4I3B, has two mutable surface patches, either closer to N-terminus (PatchN) or C-terminus (PatchC) (Figure 1). Therefore, we constructed six allAla mutants. We performed the same characterization of the allAla mutants as for the WT proteins. The number of mutated residues, protein expression, solubility, and melting temperatures are summarized in Table 2. Based on these results, we decided not to continue with the 4I3B candidate because its allAla mutants did not express well or were only poorly soluble. The 1N3Y-allAla mutants were both expressed, but only the PatchN was soluble. The best results were achieved for 4PSF because both its allAla-PatchN and allAla-PatchC variants were expressed and soluble.

Table 2. Properties of three most promising wild-type (WT) protein scaffold candidates and their allAla variants. The melting temperatures (Tm) were measured by circular dichroism.

Protein Scaffold Candidate Variant	Number of Mutated Residues	Expression	Solubility	Tm
4PSF-WT	0	Yes	Very good	75 °C
4PSF-allAla-PatchN	10	Yes	Very good	60 °C
4PSF-allAla-PatchC	10	Yes	Very good	67 °C
1N3Y-WT	0	Yes	Very good	57 °C
1N3Y-allAla-PatchN	10	Yes	Good	51 °C
1N3Y-allAla-PatchC	11	Yes	Not soluble	N/A
4I3B-WT	0	Yes	Good	47 °C
4I3B-allAla-PatchN	12	No	No expression	N/A
4I3B-allAla-PatchC	11	Yes	Poor	N/A

Next, we performed control measurements to check for non-specific binding of the 1N3Y and 4PSF scaffold candidates and their allAla mutants. We measured their binding to the capillaries used for microscale thermophoresis (MST) and to sensor chips for surface plasmon resonance (SPR) (Figure S2). We determined that the 1N3Y-WT and 1N3Y-allAla-PatchN were applicable for MST but not for SPR because they non-specifically bind to a clean surface of the SPR sensor chip. We observed no such effects with any of the 4PSF variants.

Considering all the findings, we chose the 4PSF protein as the best scaffold candidate. The WT protein as well as the allAla mutants were expressed at high yield, soluble and easy to purify, were monodispersed and folded, and possessed appropriate characteristics for MST and SPR affinity measurements. We selected the 4PSF C-terminal surface patch (PatchC) for further work because the 4PSF-PatchC-allAla mutant had the highest melting temperature among the allAla mutants.

3.1.5. Testing the Suitability of the Scaffold Candidates for In Vitro Evolution

First, we constructed a degenerate DNA library for ribosome display. The selected PatchC of the 4PSF scaffold candidate had 10 calculated mutable residues (Figure 4). The random mutagenesis of these 10 positions gave rise to the maximal diversity of more than 1012 variants, which agreed with the potential diversity suitable for ribosome display [57,58]. The DNA library was synthesized by GENEWIZ company by the degenerate NNK codons technology. The same company estimated library correctness as being 78%, which was sufficient for further work.

Next, we performed pilot ribosome display selection and crystallization of variants. We took advantage of directed evolution to train the 4PSF scaffold candidate to bind the IL-10 protein as the target molecule. Pilot experiments were undertaken to establish the ribosome display protocol for the initial preselection process (see section Section 3.2.2 below). After the fourth selection round of ribosome display, we selected eight random colonies for two-step purification and crystallization trials. We used commercial screens for preliminary crystallization trials and then manually optimized the conditions to get better diffracting crystals. We started the crystallization with seven variants (G14, G21, H25, H33, J61, J70, and J93). We observed a crystallization process of six variants, and only the G21 did not result in any crystal form in our tested screens. Four variants created crystals with low diffraction quality (ranging 6 to 8 Å). Crystallization of the J61 variant resulted in crystals diffracting up to a high resolution (1.2 Å). We have solved the structure with the molecular replacement method using the original 4PSF structure [42]. The structure, deposited under the PDB entry ID 7AVC, was similar to the original 4PSF structure with root mean square deviation of 0.93 Å calculated on 129 Cα atoms (Figure S3). The fold remains stable even after nine mutations at the selected positions. Thus, we established the suitability of the 4PSF scaffold candidate for in vitro evolution by ribosome display. Because the affinity

of the J61 variant to IL-10 was only in the micromolar range we continued in the effort to develop higher affinity binders.

Variant	Position																								
	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125	126	127	128	129	130	131	132	133	134
WT	R	E	L	V	I	T	I	A	R	E	G	L	E	D	K	Y	N	L	Q	L	N	P	E	W	R	M
A2	R	G	L	V	I	R	I	A	Q	R	G	I	E	F	K	Y	L	L	A	L	N	P	R	W	I	M
A3	R	P	L	V	I	R	I	A	V	G	G	L	E	R	K	Y	G	L	S	L	P	P	L	W	R	M
B4	R	R	L	V	I	R	I	A	R	T	G	L	E	L	K	Y	S	L	N	L	W	P	P	W	S	M
C4	R	V	L	V	I	L	I	A	L	L	G	L	E	V	K	Y	R	L	A	L	Q	P	V	W	Y	M
C11	R	L	L	V	I	L	I	A	I	V	G	L	E	W	K	Y	P	L	P	L	V	P	L	W	E	M
C12	R	G	L	V	I	E	I	A	P	T	G	L	E	W	K	Y	F	L	L	L	E	P	S	W	C	M
E3	R	G	L	V	I	G	I	A	H	R	G	L	E	S	K	Y	Y	L	R	L	G	P	R	W	W	M
F5	R	R	L	V	I	T	I	A	L	R	G	L	E	L	K	Y	P	L	C	L	R	P	A	W	H	M
G3	R	R	L	V	I	A	I	A	P	N	G	L	E	R	K	Y	T	L	H	L	T	P	T	W	S	M
G6	R	G	L	V	I	F	I	A	V	G	G	L	E	S	K	Y	L	L	D	L	E	P	L	W	H	M

Figure 4. The amino acid sequences of the wild-type (WT) 4PSF protein scaffold and its selected variants with the highest affinity to interleukin-10 and the lowest affinity to interleukin-29, bovine serum albumin (BSA), and skimmed milk. The yellow color represents amino acids of 4PSF-WT protein that were mutated (in green) in scaffold variants. The residue numbering corresponds to the PDB entry 4PSF.

3.1.6. Construction of the Final ProBi Scaffold

The final 4PSF scaffold construct used for directed evolution experiments, termed ProBi-WT hereafter, consists of three sequence segments (Figure 5): (i) the N-terminal Strep-tag to simplify purification; (ii) the 4PSF scaffold segment; (iii) the C-terminal c-Myc-tag. It was added to detect variants expressed properly to the c-Myc tag and exclude fragments emerging because of premature stop codons or frameshifts in the randomized DNA library. The expression level, solubility, and biophysical features of this new ProBi scaffold construct were similar to the original scaffold candidate.

MAWSHPQFEK *SMAQGPGQPG FCIKTNSSEG KVFINICHSP SIPPPADVTE EELLQMLEED AGFRIPMSL GEPHAELDAK GQGCTAYDVA VNSDFYRRMQ NSDFLRELVI TIAREGLEDK YNLQLNPEWR MKNRPFMGS IGSE***QKLISE EDL**

Figure 5. The amino acid sequence of the wild type ProBi-WT protein scaffold. The N-terminal Strep-tag and C-terminal c-Myc-tag are in bold, the 4PSF scaffold segment is in italic. The ten mutable amino acid residues forming the PatchC are highlighted in red.

3.2. Application of the ProBi Scaffold on the Interleukin-10 (IL-10) Target as a Model System

3.2.1. Selection of the Target Protein

Cytokines from the family of interleukin-10 are medically important signaling proteins of native immunity [31–33]. Immunosuppressive effects of the prototypical member of the family, IL-10, which may hinder immunotherapeutic strategies of cancer therapy [32], are not completely understood. In addition, monodispersity, CD spectra, and signaling on the RAW264.7 murine cell line of recombinant IL-10 produced in the laboratory encouraged use of this cytokine as the biologically relevant and molecularly well-defined target for our study.

3.2.2. Selection of Novel ProBi-Based Binders by Ribosome Display

(a) Initial preselection process. We performed four selection rounds of ribosome display, with the following preselection conditions: incubation for an hour with a mixture of BSA, IFNg, and ProBi-WT scaffold; blocking solution 3% BSA for all rounds. After the fourth round, we selected 190 random colonies (together with ProBi-WT) for a small-scale expression in deep-96-well plates and for detection of binding to IL-10 using the ELISA assay. We took advantage of the C-terminal c-Myc-tag to reveal those clones that have the correct open reading frame and bind the IL-10. From these, we selected 47 clones with the highest signal for larger-scale expression in deep-24-well plates. These 47 clones were tested for binding to IL-29, which shares the IL-10R2 receptor with IL-10 [63]. We discovered that most of the clones showed binding to IL-29 as well as to IL-10. Most variants also displayed non-specific binding to BSA and skimmed milk.

(b) Optimized preselection process. To decrease the non-specific binding of ProBi variants to BSA and milk proteins, we optimized the preselection conditions. The initial preselection conditions were extended in three ways. We (1) included IL-29 to the preselection mix; (2) varied the blocking solution between BSA (used in the first and third round) and skimmed milk (used in the second and fourth round); and (3) performed the preselection reaction in three consecutive wells. Otherwise, we completed the ribosome display the same way as in the point a) above. The optimization of the preselection monitored on 190 variants led to a decrease of non-specific binding to the BSA and milk to 50% and 60%, respectively, of the binding to IL-10. For further characterization, we selected 10 variants (Figure 4) with high affinities to IL-10 and low affinities to IL-29, BSA, and milk. We created the phylogenetic tree that is described in Supporting Information (Figure S4).

3.2.3. Characterization of Novel ProBi-Based Binders

(a) Affinity estimation by ELISA assay. We expressed and purified ten selected variants by one step purification on affinity StrepTactinXT resin and performed ELISA measurements to estimate their affinities to IL-10 and BSA (data not shown). We discarded six variants with the highest binding background to BSA and continued with four variants for more detailed characterization.

(b) Affinity measurements by MST. We expressed and purified ProBi-WT and its four variants labeled F5, G3, A2, and G6 by two-step purification on affinity StrepTactinXT resin and size exclusion chromatography. We successfully measured the IL-10 binding affinities of two variants, G3 and F5 (Figure 6). Their dissociation constants (Kd) are shown in Table 3. The F5 variant had the highest affinities, 6 nM in Tris (pH 8.0) buffer, and 15 nM in Hepes (pH 7.5) buffer, respectively. We detected binding of the A2 variant to IL-10 but we were not able to reach the bound (saturation) state of the binding curve. Therefore, we only estimated the affinities to be higher than 1 μM. The G3 variant assembled into a dimer, and the A2 and F5 variants behaved the same way as the ProBi-WT protein. The G6 variant showed the highest tendency to form oligomers (data not shown) and, therefore, it was discarded.

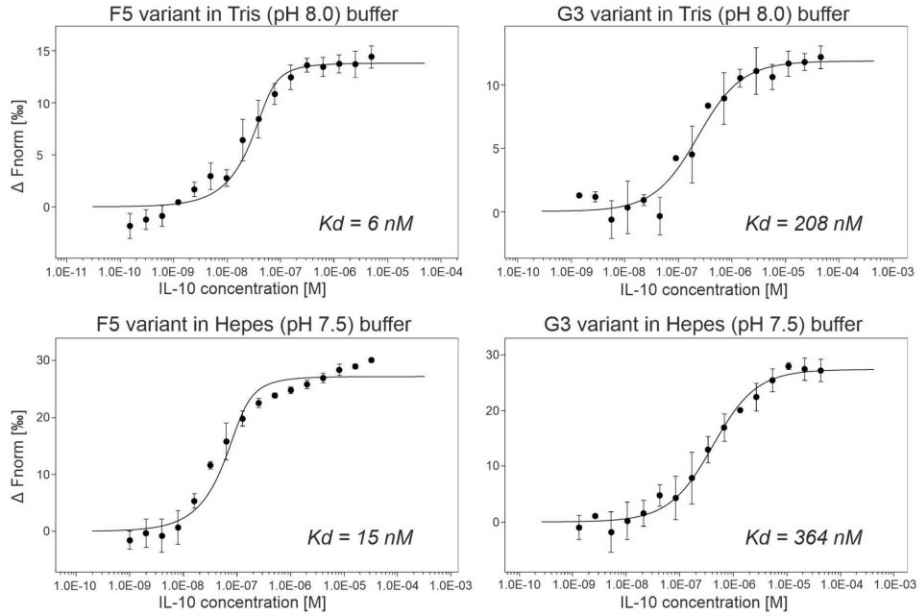


Figure 6. The microscale thermophoresis (MST) curves of two best ProBi IL-10 binders, F5 and G3. The results showed that the change of buffer from Tris to Hepes and pH from 8.0 to 7.5 did not affect the affinities.

Table 3. Variants of the ProBi scaffold sorted by their affinities to IL-10. The table shows affinities, response amplitude during microscale thermophoresis (MST) measurements, number of replicates of the MST measurements, and melting temperatures (Tm) measured by the nanoDSF method. Not measurable indicates that the affinity is so low that it falls beyond the dynamic range of the MST technique.

ProBi Variant (Buffer)	Affinity	Response Amplitude	Replicates	Tm
WT (Tris)	Not measurable	Not measurable	3	66 °C
F5 (Tris)	6 nM	14	4	51 °C
F5 (Hepes)	15 nM	27	3	N/A
G3 (Tris)	208 nM	12	3	58 °C
G3 (Hepes)	364 nM	27	3	N/A
A2 (Tris)	>1 μM	N/A	2	48 °C

To evaluate the affinities of ProBi variants to IL-10, we utilized the microscale thermophoresis technique that utilizes low material consumption [64]. We used a commercial kit to fluorescently label the IL-10 with C-terminal His-tag as a target molecule and titrated it by evolved ProBi variants with the N-terminal Strep-tag. The measured affinities are shown in Table 3. We did not detect binding of IL-10 to the parental wild-type ProBi scaffold in our binding buffer (50 mM Tris-HCl, pH 8.0, 300 mM NaCl, 0.1% Pluronic F-127). We cannot measure the affinities to BSA and milk in our MST setup because they do not possess His-tag for labeling.

(c) Characterization of ProBi variants with the highest affinity. The chromatograms from size exclusion chromatography and SDS-PAGE analysis of purified proteins are shown in Figure S5. We measured the melting temperature (Tm) of the variants (Table 3) finding that Tm was lower than for the WT protein, but still in a range acceptable for practical

purposes. We confirmed the folded structure of the selected ProBi variants with a high content of alpha helices according to their circular dichroism spectra (Figure S6).

(d) Inhibitory assay. We tested potential of the G3 and F5 variants to inhibit the IL-10 signaling pathway by a competitive binding assay on the RAW264.7 cell line, which expresses both IL10R1 and IL10R2 receptors on the cell surface. We used ProBi-WT protein as a negative control. Using these experimental conditions, we observed no inhibition of the IL-10 signaling pathway by either G3 or F5 (Figure S7). We hypothesize that these ProBi variants bind to the surface of IL-10 in such a way that they do not prevent IL-10 binding to the receptors IL10R1 and/or IL10R2 and do not block the signalization.

In order to identify truly inhibitory binders with a potential medical use, we need to include more ProBi variants, correlate their activity with the effect of neutralizing anti-IL-10 antibodies, test the inhibition on more cell types (e.g., THP-1 or U937), and possibly monitor signalization by outcomes other than STAT3 phosphorylation. The ProBi IL-10 binders developed as a proof of principle will have to be engineered further in order to inhibit the signaling pathway of IL-10 cytokine.

3.2.4. Further Development of Scaffold Selection and Improvement of Probi-Based Binders

We are aware that the procedure of scaffold selection can be optimized, and here we discuss possible future modifications of the present protocol.

(a) The first step of our procedure, the visual screening of the PDB, can be performed more systematically. A possible way would be to automate selection of monomeric small proteins and the following identification of mutable patches on their surfaces by in silico screening by employing the FoldX [36] or Rosetta [65] programs. We can also include proteins that were produced not only in *E. coli* expression systems because the expression level and solubility of the tested scaffolds can be tested in a high-throughput format.

(b) The proposed workflow will be further tested by development of our second candidate scaffold, protein alpha-X beta2 integrin I domain (UniProt ID P20702, structure of PDB ID 1N3Y [52]) in the near future.

(c) The selectivity of the best ProBi binders to other IL-10 family cytokines was tested only by the ELISA pre-screening against IL-29 (Section 3.2.2). Further development of the binders would undoubtedly require thorough checking of their cross-reactivity to more cytokines and other proteins.

(d) A substantial challenge for possible medical use of the binders is their immunogenicity [24]. We do not expect the ProBi-based binders to be immunogenic as the scaffold is derived from the human PIH1 domain-containing protein 1 (UniProt ID Q9NWS0, structure of PDB ID 4PSF [42]). However, potentially immunogenic scaffolds could be repurposed for molecular imaging [16].

(e) The directed evolution methodologies [66,67] offer avenues to development of proteins with new properties. Over the years, several display techniques [68,69] have evolved and they offer alternative and complementary ways to the selection of optimal protein molecules for the task. In this work, we make use only one of the display techniques, ribosome display. Because the yeast display offers certain advantages compared to ribosome display, we designed the final ProBi scaffold construct to be directly used by both methods. Therefore, we included the C-terminal c-Myc-tag that is widely used for high-throughput detection in Fluorescence-activated cell sorting (FACS).

(f) The protein scaffolds as well as antibodies typically bind just one interaction partner. Two independently mutable patches on one protein scaffold molecule could open a way to train binders to interact simultaneously with two partners. Therefore, we aimed at scaffolds with two mutable patches. Such a feature would mimic a function of natural proteins, such as cytokines and/or other proteins signaling through simultaneous binding to two receptors i.e., via formation of a ternary complex. The protein scaffolds can also work as "synthekines" to engineer non-natural receptor heterodimers that could activate new unexplored cellular responses [70].

4. Conclusions

Protein scaffolds represent a great engineering tool that could complement functions of more commonly used antibodies in high-affinity binding of biomolecular targets. We think that there is no "one-scaffold-fits-all" and that the development of new scaffolds tailored to specific functions is crucial for real-life applications, e.g., regulation of signaling pathways in vivo, biologics, and molecular imaging.

Visual inspection of several hundred structures from the PDB, in silico investigation of about 100 selected candidates, and finally experimental examination of 12 of them, led to selection of two potentially useful scaffold proteins. The first candidate was based on the human PIH1 domain-containing protein 1 (UniProt ID Q9NWS0, structure of PDB ID 4PSF [42]) and the second one on the alpha-X beta2 integrin I domain (UniProt ID P20702, structure of PDB ID 1N3Y [52]). The latter of the two candidates is going to be developed further in the near future.

We preferred the 4PSF protein structure because it contained two surface patches amenable for independent mutagenesis. Thus, we modified the 4PSF protein construct into the protein scaffold construct called ProBi that was ready to be used in both ribosome and yeast display technology. For the purpose of this work, we focused on one of the patches and utilized the ribosome display to evolve the ProBi scaffold into the binders of interleukin-10 with nanomolar affinity. In the future, we plan to include more scaffold candidates and concentrate on their other features, such as signaling inhibition, immunogenicity, and selectivity.

In this work, we present a proof of concept methodology to identify protein structures that could be converted into a new protein scaffold constructs. We experimentally proved that at least one of them could be adapted into binders of a medically important target with nanomolar affinity by methods of directed evolution.

Supplementary Materials: The following are available online at https://www.mdpi.com/1999-4915/13/2/190/s1, Table S1: List of primers, Table S2: Data processing statistics of PDB entry 7AVC. Figure S1: DNA sequences of internal arrangement within (A) the pRDVsm and (B) pETsm vectors. Figure S2: Testing of non-specific binding of scaffold candidates. Figure S3: Structural alignment. Figure S4: The phylogenetic tree of ten ProBi scaffold variants. Figure S5: SDS-PAGE (15%) analysis of purified ProBi variants. Figure S6: Circular dichroism spectra of (A) ProBi scaffold. Figure S7: Inhibition of the IL-10 signaling pathway estimated by a competitive binding assay.

Author Contributions: Conceptualization, B.S. and P.M.; methodology, B.S., P.M. and J.Z.; computational analysis, L.B., J.Č. and B.S.; molecular biology experiments, P.N.P., M.H., L.K. and P.M.; biophysical experiments P.N.P., M.H., T.C., G.F., P.K., Š.H. and J.P.; data analysis, M.H., P.N.P., Š.H., L.K., P.K., J.P. and J.Z.; writing—original draft preparation, B.S., P.M. and L.B.; writing—review and editing, B.S., P.M., P.N.P. and M.H.; visualization, P.M., M.H. and P.N.P.; supervision, B.S.; project administration, B.S. and P.M.; funding acquisition, B.S.; P.N.P. and M.H. contributed equally to the manuscript preparation. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Czech Science Foundation, grant 16-20507S, Czech Academy of Sciences, grant RVO 86652036, Ministry of Education, Youth, and Sports of the Czech Republic, grant LM2015043, LM2018127.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The structural data presented in this study are openly available in the Protein Data Bank, identification number 7AVC. All other data presented in this study are available on request from the corresponding authors.

Acknowledgments: We thank Marketa Janovska for her invaluable technical assistance.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bryson, J.W.; Desjarlais, J.R.; Handel, T.M.; DeGrado, W.F. From coiled coils to small globular proteins: Design of a native-like three-helix bundle. *Protein Sci.* **1998**, *7*, 1404–1414. [CrossRef] [PubMed]
2. Khoury, G.A.; Smadbeck, J.; Kieslich, C.A.; Floudas, C.A. Protein folding and de novo protein design for biotechnological applications. *Trends Biotechnol.* **2014**, *32*, 99–109. [CrossRef] [PubMed]
3. Burnside, D.; Schoenrock, A.; Moteshareie, H.; Hooshyar, M.; Basra, P.; Hajikarimlou, M.; Dick, K.; Barnes, B.; Kazmirchuk, T.; Jessulat, M.; et al. In Silico Engineering of Synthetic Binding Proteins from Random Amino Acid Sequences. *iScience* **2019**, *11*, 375–387. [CrossRef] [PubMed]
4. Kortemme, T.; Baker, D. Computational design of protein-protein interactions. *Curr. Opin. Chem. Biol.* **2004**, *8*, 91–97. [CrossRef] [PubMed]
5. Mikulecky, P.; Cerny, J.; Biedermannova, L.; Petrokova, H.; Kuchar, M.; Vondrasek, J.; Maly, P.; Sebo, P.; Schneider, B. Increasing affinity of interferon-gamma receptor 1 to interferon-gamma by computer-aided design. *BioMed Res. Int.* **2013**, *2013*, 752514. [CrossRef]
6. Smith, G.P.; Petrenko, V.A. Phage Display. *Chem. Rev.* **1997**, *97*, 391–410. [CrossRef]
7. Hosse, R.J.; Rothe, A.; Power, B.E. A new generation of protein display scaffolds for molecular recognition. *Protein Sci.* **2006**, *15*, 14–27. [CrossRef]
8. Brender, J.R.; Shultis, D.; Khattak, N.A.; Zhang, Y. An Evolution-Based Approach to De Novo Protein Design. *Methods Mol. Biol.* **2017**, *1529*, 243–264.
9. Goldenzweig, A.; Goldsmith, M.; Hill, S.E.; Gertman, O.; Laurino, P.; Ashani, Y.; Dym, O.; Unger, T.; Albeck, S.; Prilusky, J.; et al. Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability. *Mol. Cell* **2016**, *63*, 337–346. [CrossRef]
10. Musil, M.; Stourac, J.; Bendl, J.; Brezovsky, J.; Prokop, Z.; Zendulka, J.; Martinek, T.; Bednar, D.; Damborsky, J. FireProt: Web server for automated design of thermostable proteins. *Nucleic Acids Res.* **2017**, *45*, W393–W399. [CrossRef]
11. Zahradnik, J.; Kolarova, L.; Peleg, Y.; Kolenko, P.; Svidenska, S.; Charnavets, T.; Unger, T.; Sussman, J.L.; Schneider, B. Flexible regions govern promiscuous binding of IL-24 to receptors IL-20R1 and IL-22R1. *FEBS J.* **2019**, *286*, 3858–3873. [CrossRef] [PubMed]
12. Banta, S.; Dooley, K.; Shur, O. Replacing antibodies: Engineering new binding proteins. *Annu. Rev. Biomed. Eng.* **2013**, *15*, 93–113. [CrossRef]
13. Yu, X.; Yang, Y.P.; Dikici, E.; Deo, S.K.; Daunert, S. Beyond Antibodies as Binding Partners: The Role of Antibody Mimetics in Bioanalysis. *Annu. Rev. Anal. Chem.* **2017**, *10*, 293–320. [CrossRef] [PubMed]
14. Skerra, A. Engineered protein scaffolds for molecular recognition. *J. Mol. Recognit.* **2000**, *13*, 167–187. [CrossRef]
15. Hey, T.; Fiedler, E.; Rudolph, R.; Fiedler, M. Artificial, non-antibody binding proteins for pharmaceutical and industrial applications. *Trends Biotechnol.* **2005**, *23*, 514–522. [CrossRef]
16. Simeon, R.; Chen, Z. In vitro-engineered non-antibody protein therapeutics. *Protein Cell* **2018**, *9*, 3–14. [CrossRef]
17. Nord, K.; Gunneriusson, E.; Ringdahl, J.; Stahl, S.; Uhlen, M.; Nygren, P.A. Binding proteins selected from combinatorial libraries of an alpha-helical bacterial receptor domain. *Nat. Biotechnol.* **1997**, *15*, 772–777. [CrossRef]
18. Binz, H.K.; Amstutz, P.; Pluckthun, A. Engineering novel binding proteins from nonimmunoglobulin domains. *Nat. Biotechnol.* **2005**, *23*, 1257–1268. [CrossRef]
19. Skerra, A. Alternative non-antibody scaffolds for molecular recognition. *Curr. Opin. Biotechnol.* **2007**, *18*, 295–304. [CrossRef]
20. Barinka, C.; Ptacek, J.; Richter, A.; Novakova, Z.; Morath, V.; Skerra, A. Selection and characterization of Anticalins targeting human prostate-specific membrane antigen (PSMA). *Protein Eng. Des. Select.* **2016**, *29*, 105–115. [CrossRef]
21. Bedford, R.; Tiede, C.; Hughes, R.; Curd, A.; McPherson, M.J.; Peckham, M.; Tomlinson, D.C. Alternative reagents to antibodies in imaging applications. *Biophys. Rev.* **2017**, *9*, 299–308. [CrossRef] [PubMed]
22. Rothe, C.; Skerra, A. Anticalin((R)) Proteins as Therapeutic Agents in Human Diseases. *BioDrugs* **2018**, *32*, 233–243. [CrossRef] [PubMed]
23. Kosztyu, P.; Kuchar, M.; Cerny, J.; Barkocziova, L.; Maly, M.; Petrokova, H.; Czernekova, L.; Liskova, V.; Raskova Kafkova, L.; Knotigova, P.; et al. Proteins mimicking epitope of HIV-1 virus neutralizing antibody induce virus-neutralizing sera in mice. *EBioMedicine* **2019**, *47*, 247–256. [CrossRef] [PubMed]
24. Vazquez-Lombardi, R.; Phan, T.G.; Zimmermann, C.; Lowe, D.; Jermutus, L.; Christ, D. Challenges and opportunities for non-antibody scaffold drugs. *Drug Discov. Today* **2015**, *20*, 1271–1283. [CrossRef] [PubMed]
25. Azhar, A.; Ahmad, E.; Zia, Q.; Rauf, M.A.; Owais, M.; Ashraf, G.M. Recent advances in the development of novel protein scaffolds based therapeutics. *Int. J. Biol. Macromol.* **2017**, *102*, 630–641. [CrossRef]
26. Ahmad, J.N.; Li, J.; Biedermannova, L.; Kuchar, M.; Sipova, H.; Semeradtova, A.; Cerny, J.; Petrokova, H.; Mikulecky, P.; Polinek, J.; et al. Novel high-affinity binders of human interferon gamma derived from albumin-binding domain of protein G. *Proteins* **2012**, *80*, 774–789. [CrossRef]
27. Kuchar, M.; Vankova, L.; Petrokova, H.; Cerny, J.; Osicka, R.; Pelak, O.; Sipova, H.; Schneider, B.; Homola, J.; Sebo, P.; et al. Human interleukin-23 receptor antagonists derived from an albumin-binding domain scaffold inhibit IL-23-dependent ex vivo expansion of IL-17-producing T-cells. *Proteins* **2014**, *82*, 975–989. [CrossRef]

28. Kruziki, M.A.; Bhatnagar, S.; Woldring, D.R.; Duong, V.T.; Hackel, B.J. A 45-Amino-Acid Scaffold Mined from the PDB for High-Affinity Ligand Engineering. *Chem. Biol.* **2015**, *22*, 946–956. [CrossRef]
29. Alsultan, A.M.; Chin, D.Y.; Howard, C.B.; de Bakker, C.J.; Jones, M.L.; Mahler, S.M. Beyond Antibodies: Development of a Novel Protein Scaffold Based on Human Chaperonin 10. *Sci. Rep.* **2016**, *5*, 37348. [CrossRef]
30. Berman, H.M.; Battistuz, T.; Bhat, T.N.; Bluhm, W.F.; Bourne, P.E.; Burkhardt, K.; Feng, Z.; Gilliland, G.L.; Iype, L.; Jain, S.; et al. The Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.* **2002**, *58 Pt 6*, 899–907. [CrossRef]
31. El Kasmi, K.C.; Smith, A.M.; Williams, L.; Neale, G.; Panopoulos, A.D.; Watowich, S.S.; Hacker, H.; Foxwell, B.M.; Murray, P.J. Cutting edge: A transcriptional repressor and corepressor induced by the STAT3-regulated anti-inflammatory signaling pathway. *J. Immunol.* **2007**, *179*, 7215–7219. [CrossRef] [PubMed]
32. Asadullah, K.; Sterry, W.; Volk, H.D. Interleukin-10 therapy—Review of a new approach. *Pharmacol. Rev.* **2003**, *55*, 241–269. [CrossRef] [PubMed]
33. Liu, Y.; de Waal Malefyt, R.; Briere, F.; Parham, C.; Bridon, J.M.; Banchereau, J.; Moore, K.W.; Xu, J. The EBV IL-10 homologue is a selective agonist with impaired binding to the IL-10 receptor. *J. Immunol.* **1997**, *158*, 604–613. [PubMed]
34. Larkin, M.A.; Blackshields, G.; Brown, N.P.; Chenna, R.; McGettigan, P.A.; McWilliam, H.; Valentin, F.; Wallace, I.M.; Wilm, A.; Lopez, R.; et al. Clustal W and Clustal X version 2.0. *Bioinformatics* **2007**, *23*, 2947–2948. [CrossRef] [PubMed]
35. Okonechnikov, K.; Golosova, O.; Fursov, M.; Ugene Team. Unipro UGENE: A unified bioinformatics toolkit. *Bioinformatics* **2012**, *28*, 1166–1167. [CrossRef] [PubMed]
36. Schymkowitz, J.; Borg, J.; Stricher, F.; Nys, R.; Rousseau, F.; Serrano, L. The FoldX web server: An online force field. *Nucleic Acids Res.* **2005**, *33*, W382–W388. [CrossRef]
37. Mikulecky, P.; Zahradnik, J.; Kolenko, P.; Cerny, J.; Charnavets, T.; Kolarova, L.; Necasova, I.; Pham, P.N.; Schneider, B. Crystal structure of human interferon-gamma receptor 2 reveals the structural basis for receptor specificity. *Acta Crystallogr. D Struct. Biol.* **2016**, *72 Pt 9*, 1017–1025. [CrossRef]
38. Kabsch, W. Xds. *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66 Pt 2*, 125–132. [CrossRef]
39. Evans, P.R.; Murshudov, G.N. How good are my data and what is the resolution? *Acta Crystallogr. D Biol. Crystallogr.* **2013**, *69 Pt 7*, 1204–1214. [CrossRef]
40. Winn, M.D.; Ballard, C.C.; Cowtan, K.D.; Dodson, E.J.; Emsley, P.; Evans, P.R.; Keegan, R.M.; Krissinel, E.B.; Leslie, A.G.; McCoy, A.; et al. Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.* **2011**, *67 Pt 4*, 235–242. [CrossRef]
41. McCoy, A.J.; Grosse-Kunstleve, R.W.; Adams, P.D.; Winn, M.D.; Storoni, L.C.; Read, R.J. Phaser crystallographic software. *J. Appl. Crystallogr.* **2007**, *40 Pt 4*, 658–674. [CrossRef]
42. Horejsi, Z.; Stach, L.; Flower, T.G.; Joshi, D.; Flynn, H.; Skehel, J.M.; O'Reilly, N.J.; Ogrodowicz, R.W.; Smerdon, S.J.; Boulton, S.J. Phosphorylation-dependent PIH1D1 interactions define substrate specificity of the R2TP cochaperone complex. *Cell Rep.* **2014**, *7*, 19–26. [CrossRef] [PubMed]
43. Murshudov, G.N.; Skubak, P.; Lebedev, A.A.; Pannu, N.S.; Steiner, R.A.; Nicholls, R.A.; Winn, M.D.; Long, F.; Vagin, A.A. REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D Biol. Crystallogr.* **2011**, *67 Pt 4*, 355–367. [CrossRef]
44. Emsley, P.; Lohkamp, B.; Scott, W.G.; Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66 Pt 4*, 486–501. [CrossRef]
45. Williams, C.J.; Headd, J.J.; Moriarty, N.W.; Prisant, M.G.; Videau, L.L.; Deis, L.N.; Verma, V.; Keedy, D.A.; Hintze, B.J.; Chen, V.B.; et al. MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci.* **2018**, *27*, 293–315. [CrossRef] [PubMed]
46. Dreier, B.; Pluckthun, A. Ribosome display: A technology for selecting and evolving proteins from large libraries. *Methods Mol. Biol.* **2011**, *687*, 283–306. [PubMed]
47. Zahnd, C.; Amstutz, P.; Pluckthun, A. Ribosome display: Selecting and evolving proteins in vitro that specifically bind to a target. *Nat. Methods* **2007**, *4*, 269–279. [CrossRef]
48. Lam, S.Y.; Yeung, R.C.; Yu, T.H.; Sze, K.H.; Wong, K.B. A rigidifying salt-bridge favors the activity of thermophilic enzyme at high temperatures at the expense of low-temperature activity. *PLoS Biol.* **2011**, *9*, e1001027. [CrossRef]
49. Armstrong, E.H.; Goswami, D.; Griffin, P.R.; Noy, N.; Ortlund, E.A. Structural basis for ligand regulation of the fatty acid-binding protein 5, peroxisome proliferator-activated receptor beta/delta (FABP5-PPARbeta/delta) signaling pathway. *J. Biol. Chem.* **2014**, *289*, 14941–14954. [CrossRef]
50. Cheung, Y.Y.; Lam, S.Y.; Chu, W.K.; Allen, M.D.; Bycroft, M.; Wong, K.B. Crystal structure of a hyperthermophilic archaeal acylphosphatase from Pyrococcus horikoshii—Structural insights into enzymatic catalysis, thermostability, and dimerization. *Biochemistry* **2005**, *44*, 4601–4611. [CrossRef]
51. Kumagai, A.; Ando, R.; Miyatake, H.; Greimel, P.; Kobayashi, T.; Hirabayashi, Y.; Shimogori, T.; Miyawaki, A. A bilirubin-inducible fluorescent protein from eel muscle. *Cell* **2013**, *153*, 1602–1611. [CrossRef] [PubMed]
52. Vorup-Jensen, T.; Ostermeier, C.; Shimaoka, M.; Hommel, U.; Springer, T.A. Structure and allosteric regulation of the alpha X beta 2 integrin I domain. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 1873–1878. [CrossRef] [PubMed]

53. Buchko, G.W.; Ni, S.; Robinson, H.; Welsh, E.A.; Pakrasi, H.B.; Kennedy, M.A. Characterization of two potentially universal turn motifs that shape the repeated five-residues fold—Crystal structure of a lumenal pentapeptide repeat protein from Cyanothece 51142. *Protein Sci.* **2006**, *15*, 2579–2595. [CrossRef] [PubMed]
54. McKay, D.B.; Xi, L.; Barthel, K.K.B.; Cech, T.R. Structure and function of steroid receptor RNA activator protein, the proposed partner of SRA noncoding RNA. *J. Mol. Biol.* **2014**, *426*, 1766–1785. [CrossRef] [PubMed]
55. Kanagawa, M.; Satoh, T.; Ikeda, A.; Nakano, Y.; Yagi, H.; Kato, K.; Kojima-Aikawa, K.; Yamaguchi, Y. Crystal structures of human secretory proteins ZG16p and ZG16b reveal a Jacalin-related beta-prism fold. *Biochem. Biophys. Res. Commun.* **2011**, *404*, 201–205. [CrossRef]
56. Becker, A.K.; Mikolajek, H.; Paulsson, M.; Wagener, R.; Werner, J.M. A structure of a collagen VI VWA domain displays N and C termini at opposite sides of the protein. *Structure* **2014**, *22*, 199–208. [CrossRef]
57. Groves, M.A.; Osbourn, J.K. Applications of ribosome display to antibody drug discovery. *Expert Opin. Biol. Ther.* **2005**, *5*, 125–135. [CrossRef]
58. Pluckthun, A. Ribosome display: A perspective. *Methods Mol. Biol.* **2012**, *805*, 3–28.
59. Bohm, G.; Muhr, R.; Jaenicke, R. Quantitative analysis of protein far UV circular dichroism spectra by neural networks. *Protein Eng.* **1992**, *5*, 191–195. [CrossRef] [PubMed]
60. Lefevre, F.; Remy, M.H.; Masson, J.M. Alanine-stretch scanning mutagenesis: A simple and efficient method to probe protein structure and function. *Nucleic Acids Res.* **1997**, *25*, 447–448. [CrossRef] [PubMed]
61. Weiss, G.A.; Watanabe, C.K.; Zhong, A.; Goddard, A.; Sidhu, S.S. Rapid mapping of protein functional epitopes by combinatorial alanine scanning. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 8950–8954. [CrossRef] [PubMed]
62. Morrison, K.L.; Weiss, G.A. Combinatorial alanine-scanning. *Curr. Opin. Chem. Biol.* **2001**, *5*, 302–307. [CrossRef]
63. Miknis, Z.J.; Magracheva, E.; Li, W.; Zdanov, A.; Kotenko, S.V.; Wlodawer, A. Crystal structure of human interferon-lambda1 in complex with its high-affinity receptor interferon-lambdaR1. *J. Mol. Biol.* **2010**, *404*, 650–664. [CrossRef] [PubMed]
64. Jerabek-Willemsen, M.; Wienken, C.J.; Braun, D.; Baaske, P.; Duhr, S. Molecular interaction studies using microscale thermophoresis. *Assay Drug Dev. Technol.* **2011**, *9*, 342–353. [CrossRef]
65. Moretti, R.; Lyskov, S.; Das, R.; Meiler, J.; Gray, J.J. Web-accessible molecular modeling with Rosetta: The Rosetta Online Server that Includes Everyone (ROSIE). *Protein Sci.* **2018**, *27*, 259–268. [CrossRef]
66. Hibbert, E.G.; Dalby, P.A. Directed evolution strategies for improved enzymatic performance. *Microb. Cell Factories* **2005**, *4*, 29. [CrossRef]
67. Bloom, J.D.; Arnold, F.H. In the light of directed evolution: Pathways of adaptive protein evolution. *Proc. Natl. Acad. Sci. USA* **2009**, *106* (Suppl. 1), 9995–10000. [CrossRef]
68. Packer, M.S.; Liu, D.R. Methods for the directed evolution of proteins. *Nat. Rev. Genet.* **2015**, *16*, 379–394. [CrossRef]
69. Xiao, H.; Bao, Z.; Zhao, H. High Throughput Screening and Selection Methods for Directed Enzyme Evolution. *Ind. Eng. Chem. Res.* **2015**, *54*, 4011–4020. [CrossRef]
70. Moraga, I.; Spangler, J.B.; Mendoza, J.L.; Gakovic, M.; Wehrman, T.S.; Krutzik, P.; Garcia, K.C. Synthekines are surrogate cytokine and growth factor agonists that compel signaling through non-natural receptor dimers. *eLife* **2017**, *6*, e22882. [CrossRef]

Check for updates

The FEBS Journal

FEBS PRESS science publishing by scientists

Increasing recombinant protein production in *E. coli* via FACS-based selection of N-terminal coding DNA libraries

Štěpán Herynek[1], Jakub Svoboda[1], Maroš Huličiak[1], Yoav Peleg[2], Ľubica Škultétyová[1], Pavel Mikulecký[1] and Bohdan Schneider[1]

1 Institute of Biotechnology, Czech Academy of Sciences, BIOCEV, Prague, Czech Republic
2 Structural Proteomics Unit (SPU), Department of Life Sciences Core Facilities (LSCF), Weizmann Institute of Science, Rehovot, Israel

Keywords
directed evolution; DNA libraries; fluorescence-activated cell sorting (FACS); N-terminal sequences; protein expression optimization

Correspondence
B. Schneider, Institute of Biotechnology, Czech Academy of Sciences, BIOCEV, Průmyslová 595, Prague West, Prague 25250, Czech Republic
Tel: +420 728 303 566
E-mail: bohdan.schneider@ibt.cas.cz

(Received 18 July 2024, revised 28 August 2024, accepted 25 November 2024)

doi:10.1111/febs.17376

Here, we present a previously undescribed approach to modify N-terminal sequences of recombinant proteins to increase their production yield in *Escherichia coli*. Prior research has demonstrated that the nucleotides immediately following the start codon can significantly influence protein expression. However, the impact of these sequences is construct-specific and is not universally applicable to all proteins. Most of the previous research has been limited to selecting from a few rationally designed sequences. In contrast, we used a directed evolution-based methodology, screening large numbers of diversified sequences derived from DNA libraries coding for the N-termini of investigated proteins. To facilitate the identification of cells with increased expression of the target construct, we cloned a GFP gene at the C-terminus of the expressed genes and used fluorescent activated cell sorting (FACS) to separate cells based on their fluorescence. By following this systematic workflow, we successfully elevated the yield of soluble recombinant proteins of multiple constructs up to over 30-fold.

Introduction

Bacteria *Escherichia coli* (*E. coli*) is frequently utilized for producing recombinant proteins. This expression system is preferred due to well-established workflow, simple cultivation and transformation processes, rapid growth rate, and overall cost-effectiveness [1–4]. Nevertheless, recombinant protein production in *E. coli* remains a challenge and requires numerous optimization strategies. Factors such as vector type, promoter, *E. coli* strain choice, inducer concentration, temperature, and media composition significantly influence yield [5].

It has been reported multiple times that the composition of the first few codons following a start codon plays a critical role in translation regulation [6–15]. However, it is not clear whether the control mechanism takes place on the level of nucleic acid sequence-induced structure, codon usage, or folding of nascent peptide.

Translation rate can be influenced by N-terminal amino acids [16] and it has been identified that the crucial molecular component is the second amino acid [6]. The authors have utilized a pET vector using different *E. coli* strains and mutated the second amino acid position of the extracellular domain of Igα protein to all 20 natural amino acid residues. The data indicated that the amino acid at the second position had a 10-fold impact on the yield of recombinant protein expression [6]. The importance of the second amino acid position has been found crucial before,

Abbreviations
CN, natural codons; CO, optimized codons; ERFE, erythroferrone; FACS, fluorescence-activated cell sorting; GFP, green fluorescent protein; GST, glutathione S-transferase; IFNb, interferon beta; IL9, interleukin 9; MCS, multiple cloning site; RF, Restriction-Free (cloning); SOL, soluble fraction; TrxA, thioredoxin A.

The FEBS Journal (2024) © 2024 The Author(s). *The FEBS Journal* published by John Wiley & Sons Ltd on behalf of Federation of European Biochemical Societies.
This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

though authors attributed the effect to the nucleotide triplet rather than the amino acid [17].

The protein production has been successfully increased by adding a small number of amino acids common to the N termini of highly expressed proteins. The N-terminal sequence MSKIK was designed by this approach and proven to enhance the protein yield although the mechanism underlying this phenomenon has not been identified [9]. It is worth noting that this sequence contained a relatively rare ATA codon (coding for isoleucine) at its N-terminal part. Changing the nucleotide sequence without altering the amino acid sequence did not affect the recombinant protein yield and authors also reported that the mRNA levels coding the protein tagged with the MSKIK sequence were 4–5 times lower than those of equivalent untagged proteins, suggesting that the mRNA level is not the key factor [9]. Six years later the same authors proposed that their MSKIK N-terminal peptide prevents or releases ribosomal stalling resulting in a higher protein yield, while also ruling out the influence of codon usage [18].

The speed of translation initiation is also influenced by the mRNA composition and structure. Structural organization of the ribosome binding site is especially important and important is also beginning of the open reading frame [14,19–22]. Indeed, bioinformatic analysis revealed a trend across various eubacteria wherein the content of mRNA secondary structures decreased towards the 5′ end of analyzed genes [23]. The stability of mRNA structures is usually estimated using nearest-neighbor minimum-free energy models. These calculations performed on a sequence near the Shine-Dalgarno sequence have revealed that a more stable mRNA secondary structure decreases the protein expression [24,25]. Subsequently, a bioinformatics tool called TISIGNER has been developed to minimize the likelihood of mRNA secondary structure formation. It designs optimal initial codons of the gene while keeping the same amino acid sequence [26]. An alternative methodology has been employed modifying the amino acid sequence to achieve the optimal codon utilization [17]. To systematically investigate this, all 64 codons were introduced after the start codon of modified *lacZ* gene, and their impact on recombinant protein yield was analyzed, showing a 15-fold disparity in yield between the lowest and the highest amount of the produced recombinant protein [17]. As expected, the outcome of several studies suggests a connection between translation efficiency and mRNA disorder near the start codon in *E. coli* [20], but also in other organisms including other bacteria, archaea, fungi, plants, insects, and fishes [27] and even in dsDNA of viruses [28].

Codon usage bias is another factor influencing the translation process [29,30]. This phenomenon has also been addressed while researching the N-terminal sequences. Some findings indicate that the point of interest should be the relative frequency of codons. This is consistent with the previous results showing that rare codons such as AGG, the least used codon in *E. coli*, appear more frequently among the first 25 codons [11,31], even though the overall codon usage usually correlates with the abundance of tRNAs [32,33]. The ability of the rare codon placed at the beginning of the gene to increase the yield has been experimentally proven using different proteins such as lacZ and streptokinase, both of which have relatively high amounts of rare codons in their sequences (about 30%). The expression of AGG triplet codon encoding Arginine at various positions within a gene has been evaluated. Placing the codon at the second position resulted in increased expression. However, when it was placed farther along the gene, the yield decreased, and expression was stalled [11,34]. This again emphasizes the importance of the position of the second amino acid in sequences [6]. In contrast, more frequently used Arginine codons at the second and at the third positions lowered the expression [34]. This is in concordance with the previous studies, which show that the placement of rare codons (such as AGG or AGA codons) has a crucial effect on a yield of a produced protein [31,35].

The varying levels of expression yields might be influenced by the number of mRNA copies, suggesting that optimizing protein production may be linked to the level of DNA transcription. However, this is not a universal mechanism. For example, in the case of ovine growth hormone, higher production levels do not correlate with increased mRNA levels compared to the wild-type variant [36]. In another instance, the mRNA levels for an anti-*E. coli* O157 fragment of antigen binding fused with a leucine zipper were reported to be even lower [9].

The above cited and other studies demonstrate that in specific cases the N-terminal sequence of to-be-produced proteins can be optimized for yield improvement. However, they also demonstrate that to formulate the ultimate and generally applicable rules how to optimize the yield or explain why the optimization has been fulfilled is not easy [37]. From the point of the final goal, i.e., increasing the protein production, it is not critical if the increased yield has been achieved by optimized codon composition at the level

The FEBS Journal (2024) © 2024 The Author(s). *The FEBS Journal* published by John Wiley & Sons Ltd on behalf of Federation of European Biochemical Societies.

17424658, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1111/febs.17376 by Cochrane Czech Republic, Wiley Online Library on [02/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

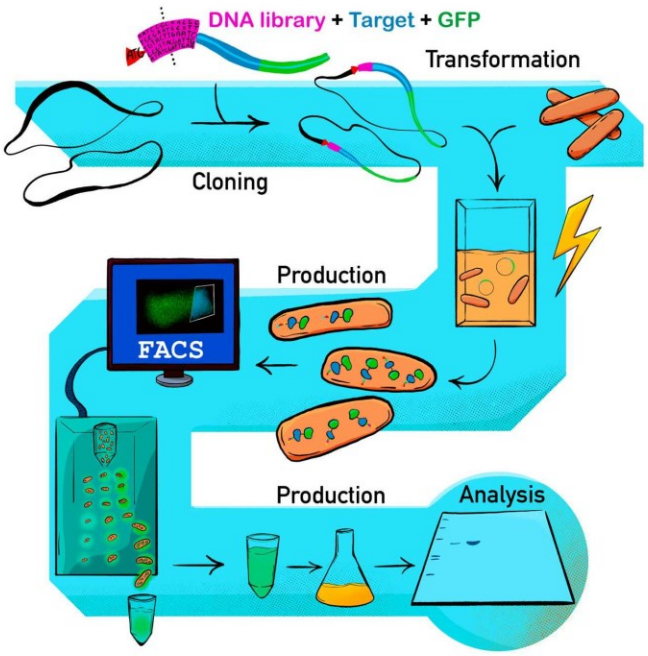


Fig. 1. Schema of the protocol for enhancing protein yields in *E. coli*. The main steps of the protocol are: *Cloning* of the production vector, which includes inserting the target gene (in blue) into the plasmid along with a DNA library coding the randomized N-terminus (in purple) and the GFP gene (in green) fused to the C-terminus of the construct. *Transformation* into electrocompetent cells and protein *production* followed by *FACS selection* of the clones overproducing target protein monitored by higher fluorescence levels. Sorted cells are used for protein *production* with the library-derived sequences. The protein production is *analyzed* by SDS/PAGE.

of the transcription, initial phases of the translation, or behavior of a group of few N-terminal amino acids of the nascent protein. To directly address the issue of protein production we decided to systematically manipulate three to five amino acid residues at the N termini of several proteins. We utilized randomized DNA libraries, which we previously used with success to increase stability and binding of a small protein binders [38,39]. To separate cells showing higher protein production, we used fluorescent activated cell sorting (FACS) and summarized our findings into an easy-to-follow protocol.

Results and discussion

Workflow overview

We present a new method to enhance the production of various recombinant proteins in *E. coli* using genetic engineering techniques. To accomplish this, we constructed DNA libraries that randomly modified the N-terminal sequences of the target proteins. These libraries were fused with the genes encoding the target proteins and a green fluorescent protein (GFP) marker at the C terminus. We used cell sorting to assess the effectiveness of these modified constructs. Based on our successful results, we established a streamlined protocol for this process, as depicted in Fig. 1 and described in detail below.

Experimental design and gating strategy

Selection of the reporter protein molecule and vector for FACS

We used the cycle 3 GFP as a reporter molecule for all our constructs because it includes several mutations resulting in increased fluorescence compared to the wild-type GFP. These mutations also facilitate proper maturation of the protein at 37 °C and the molecule shows lower tendency to aggregate [40,41]. We commonly refer to this molecule as "GFP" in our study.

N-terminal sequence optimization by FACS selection

To verify that changing the N-terminal nascent protein sequence affects the protein yield, we used *E. coli* BL21-Gold (DE3) cells transformed either with pET22b vector without GFP (negative control), pET22b with

17424658, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1111/febs.17376 by Cochrane Czech Republic, Wiley Online Library on [02/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

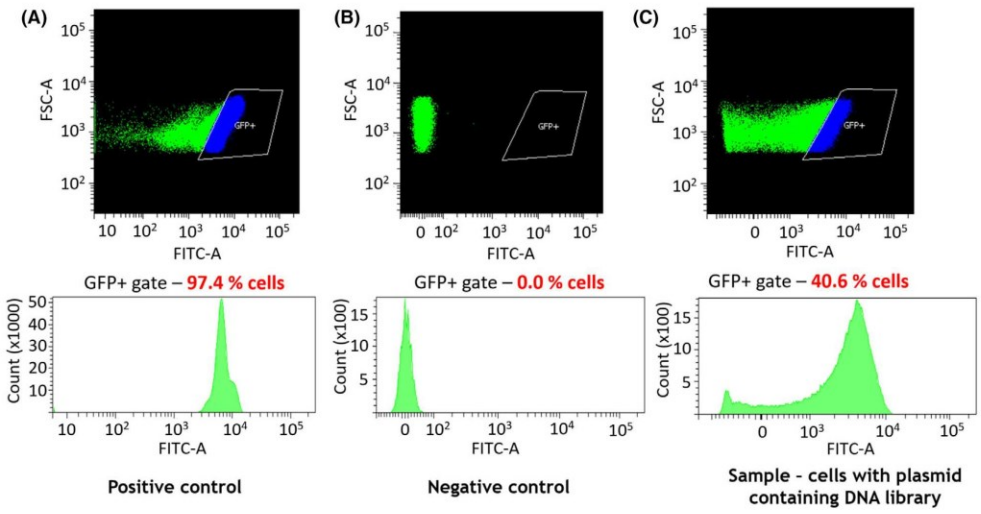


Fig. 2. Initial experimental setup. (A) Positive control showing 97.4% of the singlet cells in the GFP+ gate, which was formed based on this sample. (B) Negative control (pET22b) without GFP, showing no fluorescence and 0.0% of the shown singlet cells in the GFP+ gate. (C) Sample containing the library showing 40.6% of the singlet cells in the GFP+ gate.

GFP gene (positive control) or N-terminal library fused to GFP molecule in the pET22b vector library containing GFP sample (library containing GFP sample). Productions were carried out overnight at 18 °C. Results are shown in Fig. 2.

Based on the positive control, we designed the GFP+ gate, which captured the major population containing 97.4% of the shown singlet cells. We kept the same gate for other samples, showing that the negative control (pET22b sample) contained 0.0% of the shown singlet cells, which formed a separate population exhibiting no fluorescence. The library containing the sample did not form a uniform population regarding the fluorescence level. In this case, the GFP+ gate contained 40.6% of the shown singlet cells, as expected.

This initial experiment had shown that adding the library to the N terminus of constructs affected the yield, since we had observed a much wider distribution of the FITC-A median values across the spectrum, compared to samples without any library.

Enhancing expression yield by selecting the best N-terminal sequences by FACS

In the following paragraphs, we will show results of enhancing yield of two proteins, glutathione S-transferase (GST) and interferon beta (IFNb), all results are summarized in Fig. 7. *Escherichia coli* cells were transformed by pQIK plasmid containing GST gene and used for protein production. Cells were harvested and sorted based on their GFP fluorescence as shown in Fig. 3. Collected fractions of cells (labeled as P2 and P3 during the sorting) were analyzed by SDS/PAGE shown in Fig. 4. This experiment was performed on four samples in total: GST combined with 2 versions of the DNA library (representing four or five amino acids). We utilized two *E. coli* strains: BL21-Gold (DE3), a B strain derivative known for its high transformation efficiency and deficiency in endonuclease I (*endA*), making it ideal for protein production, and XL1-Blue (XL-1), a K-12 strain traditionally used for cloning. While XL-1 also lacks endonuclease I, it is unsuitable for T7 promoter-based protein expression and has a slower growth rate compared to BL21-Gold (DE3). To illustrate the sorting procedure in Fig. 3, we used only one of the samples – GST with the five-amino acid library in *E. coli* BL21-Gold (DE3) cells. During the FACS experiments, we consistently sorted 50 000 cells for each sample to compare their production levels. This method allowed us to directly compare the yield between 50 000 sorted cells plated onto solid LB-agar medium and used for protein production and 50 000 cells carrying the control vector, which were treated the same way.

The FEBS Journal (2024) © 2024 The Author(s). *The FEBS Journal* published by John Wiley & Sons Ltd on behalf of Federation of European Biochemical Societies.

17424658, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1111/febs.17376 by Cochrane Czech Republic, Wiley Online Library on [02/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

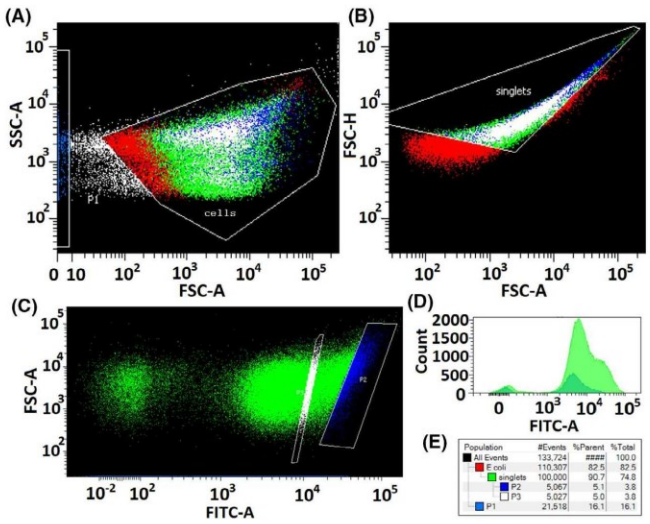
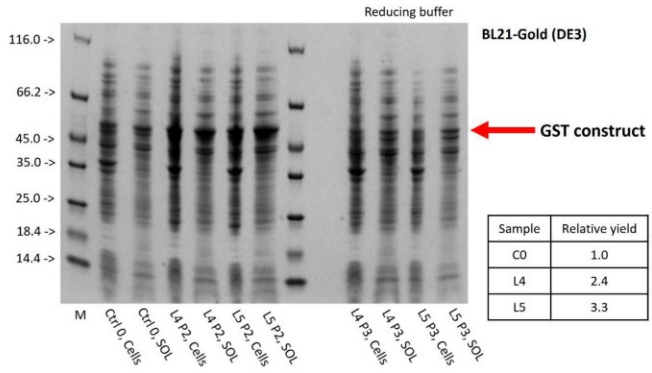


Fig. 3. Sorting procedure of the *E. coli* BL21-Gold (DE3) cells containing pQIK plasmid with GST gene with five-amino acid library. (A) Gating of live cells. We cut out the smallest particles running through the cytometer (hits in the left part of the dot plot). (B) Gating of the singlet cells. The rectangular gate filters out events unlikely to represent singlet cells. As bacteria are near the instrument's detection limit due to their size, this gate excludes extremes to improve singlet cell analysis accuracy. (C) Gating based on the fluorescence level of analyzed cells (gate P2 shown in dark blue and gate P3 shown in white). (D) Histogram of the fluorescence level values of the dot plot shown in (C). (E) Table showing population percentage in all gates. Reason why there are multiple gates for the cell sorting (P2 and P3) is explained in detail in 'Preventing the GFP molecule cleavage'.



Sample	Relative yield
C0	1.0
L4	2.4
L5	3.3

Fig. 4. SDS/PAGE analysis of the GST produced by sorted *E. coli* BL21-Gold (DE3). Expected size of the construct is 53.0 kDa (plus four or five amino acids derived from the N-terminal library). Control samples without any library are labeled Ctrl0, libraries containing samples are labeled L4 and L5 for four- and five-amino acid libraries, respectively. Different samples based on their gating strategy are labeled P2 or P3 as labeled in Fig. 3. Samples labeled Cells contain disintegrated cells after the protein production, samples labeled SOL contain soluble fraction only, which was used to calculate the yield of the protein. The protein yield is summarized in the table on the right showing that we were able to get 2.4 times higher yield out of the sorted cells in the P2 gate with the four-amino acid library and 3.3 times higher yield while working with the five-amino acid library. Molecular weight markers (M, kDa) are indicated on the left.

Sorted cells were plated onto solid agar medium containing carbenicillin and grown overnight. The plate was washed by LB medium, and the protein production procedure was performed. Protein yield including a solubility analysis was analyzed by SDS/PAGE, as shown in Fig. 4. The enrichment was more significant while using *E. coli* BL21-Gold (DE3) strain compared to XL-1 strain, which is why we show *E. coli* BL21-Gold (DE3) data only. Using the BL21-Gold (DE3) cell strain, we were able to increase the yield of the GST fusion 3.3-fold, while the XL-1 cell strain parallel experiment delivered only 1.8-fold enrichment.

Another example of successful protein enrichment through the N-terminal sequence manipulation beside

17424658, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1111/febs.17376 by Cochrane Czech Republic, Wiley Online Library on [02/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

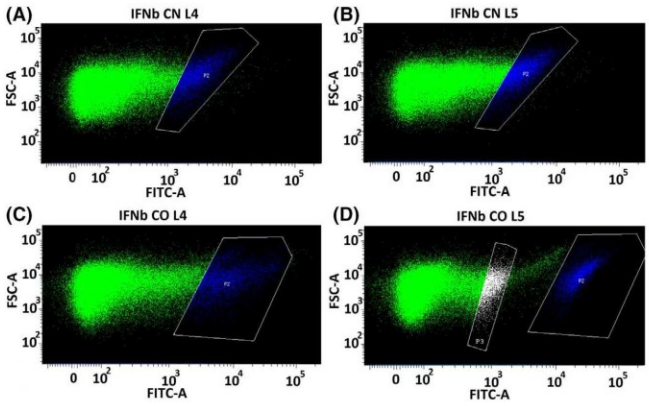
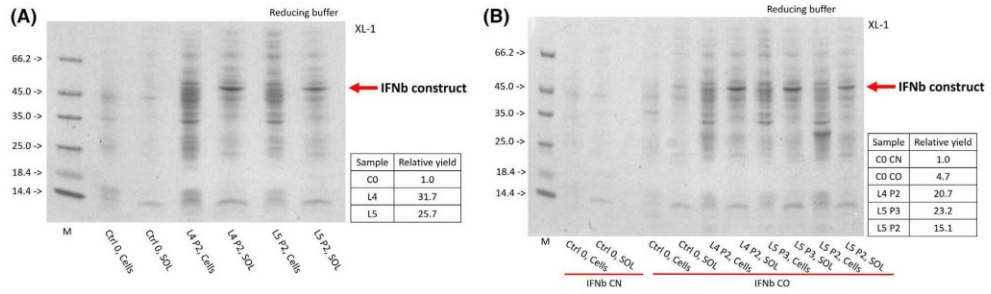


Fig. 5. Gate design during IFNb sorting. IFNb samples (top part (A) and (B)) with natural codons are labeled CN, optimized codon samples (lower part (C) and (D)) are labeled CO. Samples with four-amino acid library (on the left) are labeled L4 and five-amino acid library samples (on the right) are labeled L5. In the case of IFNb CO L5, we designed the P3 gate as well, because the population was not uniform (this strategy is described in more detail in the next chapter Preventing the GFP molecule cleavage). Reason why there are multiple gates for the cell sorting (P2 and P3) is explained in detail in chapter Preventing the GFP molecule cleavage.



Sample	Relative yield
C0	1.0
L4	31.7
L5	25.7

Sample	Relative yield
C0 CN	1.0
C0 CO	4.7
L4 P2	20.7
L5 P3	23.2
L5 P2	15.1

Fig. 6. SDS/PAGE analysis of IFNb yield enrichment. The expected size of the IFNb construct is 47.7 kDa (plus four or five amino acids delivered from the library). Control samples without any library are labeled Ctrl0, libraries containing samples are labeled L4 and L5 for four- and five-amino acid libraries, respectively. Different samples based on their gating strategy are labeled P2 or P3 as labeled in Fig. 5. Samples labeled Cells contain disintegrated cells after the protein production, samples labeled SOL are soluble fraction only, which was used to calculate the yield of the protein. (A) IFNb with natural codons (CN). (B) IFNb with optimized codons (CO) compared to the natural codon variant (first two lanes containing samples). Tables on the right side of each gel shows how many times the yield increased compared to a control with no library (C0). The yield increases are calculated from the soluble fractions. In this case the results are shown as relative multiples compared to a variant with natural codons; optimizing the codons already increased the product yield almost five-fold. Molecular weight markers (M, kDa) are indicated on the left.

the GST construct is IFNb. We used two versions of the gene: with natural codons (labeled CN) and with optimized codons (labeled CO), optimized using the GenScript codon optimization tool. Both versions were cloned into the pQIK vector and transformed into BL21-Gold (DE3) and XL-1 *E. coli* cell strains, produced, and the cells were sorted as described earlier. The entire experiment was conducted similarly to the previous one with the GST construct. The IFNb results are shown in Fig. 5 (cell sorting) and Fig. 6 (-SDS/PAGE analysis). Enrichment in BL21-Gold (DE3) was insignificant (data not shown). However, in the XL-1 cell strain, codon optimization alone increased IFNb yield over 4-fold, and our approach further enhanced production up to over 30-fold. Results of optimization for both IFNb and GST recombinant protein constructs are summarized in Fig. 7 along with other constructs that we tried to optimize using our N-terminal sequence manipulation.

Lanes in all gels, such as the lanes in Figs 6A,B and 10C, respectively, are produced by the same volume of the cell suspension grown from the 50 000 FACS sorted cells. The same number of cells, 50 000, were obtained for both cell cultures, those containing plasmids with variable N termini and cells carrying the control plasmid. Both cell samples were grown under

The FEBS Journal (2024) © 2024 The Author(s). *The FEBS Journal* published by John Wiley & Sons Ltd on behalf of Federation of European Biochemical Societies.

17424658, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1111/febs.17376 by Cochrane Czech Republic, Wiley Online Library on [02/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

17424658, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1111/febs.17376 by Cochrane Czech Republic, Wiley Online Library on [02/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Target protein	Library size (amino acids)	*E. coli* strain used	relative yield
TrxA	4	XL-1	10.7
	5		15.8
	4	BL21-Gold (DE3)	1.7
	5		4.3
EL222	4	XL-1	6.6
	5		7.1
	4	BL21-Gold (DE3)	not increased
	5		not increased
GST	4	XL-1	1.7
	5		1.8
	4	BL21-Gold (DE3)	2.4
	5		3.4
IFNb (CN)	4	XL-1	31.7
	5		25.7
	4	BL21-Gold (DE3)	not increased
	5		not increased
IFNb (CO)	4	XL-1	4.4
	5		4.9
	4	BL21-Gold (DE3)	not increased
	5		not increased

Fig. 7. Summary of proteins that we enriched using the presented protocol. Each construct was linked to two versions of DNA libraries coding four or five amino acids (2nd column) and produced in two different *E. coli* strains, XL-1 or BL21-Gold (DE3) (3rd column). Relative yields compared to a construct without any library are in the last column. The proteins of interest (1st column) were thioredoxin A (TrxA) (UniProt: P0AA25), EL222 (UniProt Q2NB98 residues 17–225), GST (UniProt P08515 residues 1–217) and interferon beta in two variants (CN = codon natural; CO = codon optimized for production in *E. coli* – Uniprot P01574 residues 23–187).

the same conditions and the equal volumes were used for the SDS/PAGE analysis. Since the goal of this process is to increase the yield of the target protein, and we began with the starting cultures containing the same number of cells, the result suggests that in some cases the sorting process itself enhances the overall efficiency of the presented protocol by sorting cells showing higher viability as well.

Preventing the GFP molecule cleavage

In some cases, we experienced a spontaneous cleavage of the GFP molecule from the rest of the construct, which usually results in higher fluorescence of the cells during the cell sorting. Whenever the cleavage occurs, the cell population on the dot plot during sorting appears to lack uniformity, serving as a cautionary signal in the sorting process, as shown in Fig. 8. First time we noticed this problem was while trying to enhance production of human interleukin 9 (IL9).

To prevent selection of clones with cleaved GFP, which occurs even during production of the recombinant protein after sorting, we came up with a solution, which utilizes a different gating strategy. Precise gate design is crucial to separation of these cells. When fluorescence levels appear non-uniform (as illustrated in Fig. 8D), it is important to collect not only the highest fluorescence population but also the second highest. An illustrative example of gate design is presented in section B of Fig. 8. We demonstrate this phenomenon on thioredoxin A (TrxA) with a five-amino acid library. During this experiment the production was carried out at 37 °C and took 3 h. The production was repeated using sorted cells under the same conditions. Results are shown in Fig. 8.

In one of the vectors, we used (pKIK, described in more detail further in the text), there is an HRV3C protease cleavage site to cleave the GFP after the selection. To avoid the unwanted cleavage, we replaced the HRV3C cleavage site with a tetrapeptide GGSG linker and dodecamer (GGSG)3 linker. We also created one version without any linker between the IL9 gene and the GFP. These 3 new constructs were expressed, and their expression compared to the performance of the original vector. The ratio of the cleaved and not cleaved versions was compared by a western blot using anti-His antibody conjugated with a peroxidase, revealing improved coherence of all three new constructs (data not shown) so that the cleavage was prevented. However, the overall expression of IL9 was too low even after the cell sorting, likely due to inability of the bacterial expression system to accommodate the used IL9 constructs. We therefore decided not to further study this protein in this project.

Four- and five-amino acid library is sufficient

Based on the current state of knowledge, we designed three different library sizes – 9, 12, and 15 nucleotides coding for three, four, and five N-terminal amino

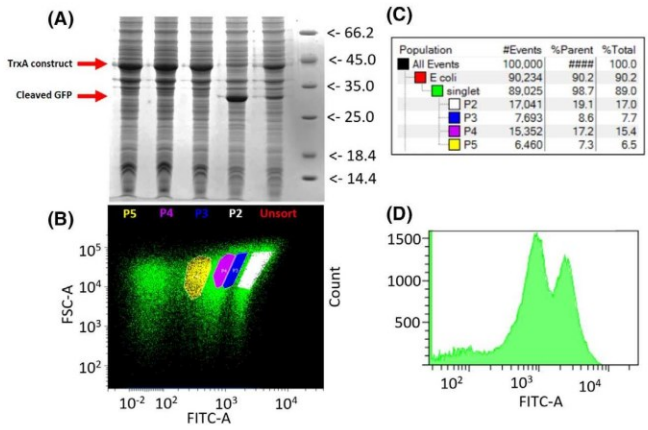


Fig. 8. Identification of GFP cleavage in cell-sorting experiments. (A) SDS/PAGE analysis of the produced constructs. Expected size of the whole construct (TrxA + GFP) is 40.6 kDa. Expected size of the cleaved GFP only is 27.5 kDa. In the P2 sample, the major band corresponds to cleaved GFP, while in sample P3, P4 and P5 the major band represents the whole TrxA-GFP construct. Molecular weight markers (kDa) are indicated on the right. (B) Flow cytometry showing the TrxA-GFP construct with the five-amino acid library. Four different gates (P2–P5) were designed to obtain cells from the whole fluorescent spectrum: P2 (white) for the highest fluorescent population, P3 (blue) for the top of the second highest fluorescent population, P4 (magenta) for the middle part of the same population, and P5 (yellow) for the lowest part. It turned out that the P2 population represented cells with cleaved GFP inside. (C) Table showing population percentage in all gates. (D) Histogram of the fluorescence level values of the dot plot shown in (B).

acids. We selected these sizes with the consideration that utilizing a shorter library would likely result in a better protein product for subsequent experiments. These library versions were cloned on the N terminus of the GFP gene produced in an "empty" pQIK vector. These vectors were transformed into both XL-1 and BL21-Gold (DE3) *E. coli* cell strains alongside a control vector lacking any library insertion.

Following the production, cell cultures underwent FACS-based sorting procedures. Sorted cells were harvested and used for a protein expression afterwards to analyze the difference in the construct expression in sorted cells and the control sample. SDS/PAGE analysis and differences in the yield calculated using ImageJ [42] are shown in Fig. 9.

Following the outcomes of this experiment, the decision was made to progress with the 12 and 15 nucleotide libraries because the smallest version containing 9 nucleotides yielded the least favorable results.

It is also worth noting that yield of proteins that are easily expressed is hard to optimize using this method since their original N terminus is more likely to be closer to the optimal composition. We stumbled over this limitation while failing to significantly increase the yield of easy-to-produce photoactive yellow protein and GFP by manipulating N-terminal parts of their genes.

One round of cell sorting is enough

To quantify the impact of the number of rounds of cell sorting on the yield of recombinant protein production, we ran two rounds of the cell sorting with two different proteins (TrxA and EL222) in three different variants: no library, library coding four amino acids and library coding five amino acids.

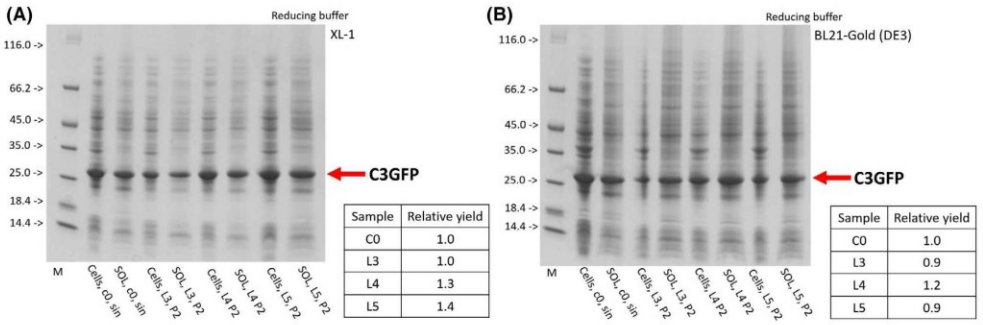
Sorted cells were plated onto solid agar medium and used for protein production the next day. 0.5 mL of the suspension was taken from each sample and used for the second round of the cell sorting. The rest of the cell suspensions were harvested, and the solubility analysis was performed. Protein yields were analyzed by SDS/PAGE (Fig. 10). The same analytical procedure was then repeated with the cells sorted during the second round. The results, summarized in Fig. 10 (SDS/PAGE), and Fig. 11 (the yield analysis), showed that adding another round of the cell sorting was not improving the overall protein yield.

The P2 gate should aim for the top 5% of the cells with the highest fluorescence and 50 000 cells

Experiments we conducted to find out the correct size of the sorting gate showed that the gate should contain approximately 5% of the singlet cells. If the

The FEBS Journal (2024) © 2024 The Author(s). *The FEBS Journal* published by John Wiley & Sons Ltd on behalf of Federation of European Biochemical Societies.

17424658, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1111/febs.17376 by Cochrane Czech Republic, Wiley Online Library on [02/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License



Sample	Relative yield
C0	1.0
L3	1.0
L4	1.3
L5	1.4

Sample	Relative yield
C0	1.0
L3	0.9
L4	1.2
L5	0.9

Fig. 9. SDS/PAGE analysis of the produced GFP constructs. Digested cells (labeled "Cells") and soluble fractions (labeled "SOL") of desired constructs. Expected size of the GFP construct is 27.5 kDa (plus four or five amino acids derived from the N-terminal library). Molecular weight markers (kDa) are indicated on the right. Tables in A and B show the relative yield calculated from the soluble fractions using ImageJ. (A) The GFP construct produced in XL-1 cells. (B) GFP construct produced in BL21-Gold (DE3) cells.

percentage was lower, meaning that the gate was aiming for more strict selection of the cells, the overall yield was not higher and the samples showed more contamination. During cell sorting, we sorted 50 000 cells, which was optimal for covering a 9 cm Petri dish while still allowing for single colony selection if needed.

Solubility of optimized protein variants

Another outcome of this experiment is a notable improvement of the solubility of expressed variants, which is often overlooked in the literature. The solubility improvement is demonstrated in Fig. 10C for TrxA protein. TrxA is present in the Ctrl0 sample (lane labeled Ctrl0, Cells), but there is a very low amount of the construct in the soluble fraction (lane labeled Ctrl0, SOL). After cell sorting, it is evident that the ratio of the protein in the digested cells and soluble fraction has shifted towards the soluble fraction. See lanes containing cells from the P2 gate. For example, in lanes L4 P2, Cells and L4 P2, SOL, there is a significantly higher amount of the product in the L4 P2, SOL lane compared to the C0 lanes.

The N-terminal sequences cannot be used universally

N-terminal sequences used to benchmark success of the yield optimization

To investigate whether the same N-terminal sequence can be used to optimize yield of multiple proteins we designed a set of 10 sequences that can be placed after the start codon of produced proteins. These sequences are taken from literature, copied from common vectors, and from well-expressed genes as indicated in Table 1. These sequences serve for benchmarking the success of the yield optimization.

Vector selection for testing the benchmarking sequences

We designed two different vectors called pKIK and pQIK. The first expression vector, pKIK, is based on commercially available vector pET22 and the cassette contains either a DNA library or a benchmark sequence, cloning site with *NotI* and *AscI* restriction enzyme sites, followed by a HRV3C protease cleavage site to cleave the following GFP with C-terminal His-Tag. The second vector, pQIK, based on Qiagene's pQE30, contains a simpler expression cassette without multiple cloning sites and it also lacks the HRV3C cleavage site. Cloning procedures were designed in a way, which does not require any restriction enzyme sites, so there is a library or a benchmark sequence after the start codon, followed by C-terminal His-Tagged GFP. See Data S1 (Plasmid_sequences) for detailed description of vector sequences.

Originally, we started this project with the pKIK vector, and we decided to implement a different plasmid while trying to prevent the GFP cleavage (discussed above). This is also one of the reasons why the expression cassette of the newer pQIK vector is simpler compared to the original pKIK vector. Vector of choice for the FACS experiments was the pQIK vector for its simpler composition and ability to be used in more different *E. coli* cell strains. The expression cassettes of both vectors are shown in Fig. 12. Main

17424658, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1111/febs.17376 by Cochrane Czech Republic, Wiley Online Library on [02/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

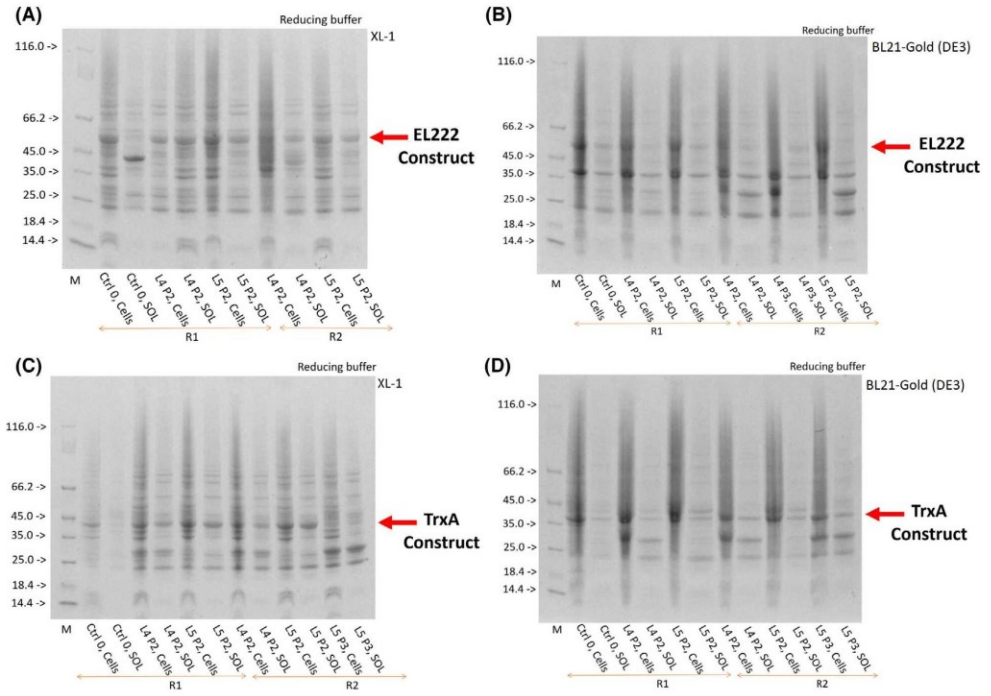


Fig. 10. SDS/PAGE analysis of the produced constructs. Digested cells (labeled "Cells") and soluble fractions (labeled "SOL") of desired constructs. Expected size of the cleaved GFP is 27.5 kDa. (A) EL222 produced in XL-1 cells, expected size of the whole construct (EL222 + GFP) is 50.8 kDa (plus four or five amino acids derived from the N-terminal library). (B) EL222 produced in BL21-Gold (DE3) cells, expected size of the whole construct (EL222 + GFP) is 50.8 kDa (plus four or five amino acids derived from the N-terminal library). (C) TrxA produced in XL-1 cells, expected size of the whole construct (TrxA+GFP) is 39.5 kDa (plus four or five amino acids derived from the N-terminal library). (D) TrxA produced in BL21-Gold (DE3) cells, expected size of the whole construct (TrxA+GFP) is 39.5 kDa (plus four or five amino acids derived from the N-terminal library). Molecular weight markers (M, kDa) are indicated on the left. R1 and R2 refer to rounds 1 and 2 of selection, respectively.

differences between pKIK and pQIK vectors are also shown in Table 2.

To test the influence of the benchmark sequences on the recombinant protein yield, we performed protein production at 37 °C for 3 h and harvested cells were used for a FACS experiment. The comparison of fluorescence among all samples was based on FITC-A median values (Fig. 13).

We decided not to report specific sequences isolated from sorted cells that exhibited higher protein yields. These sequences are highly construct-dependent and are not effective for all constructs. Therefore, their general use could potentially mislead researchers by suggesting a universal solution to this complex problem, which involves multiple molecular mechanisms.

Conclusions

This study presents a novel approach for enhancing recombinant protein yield in *E. coli* expression systems. Utilizing a randomized N-terminal library and fluorescence-activated cell sorting (FACS) selection, we achieved up to 30-fold yield increases for various recombinant proteins (Fig. 7). This approach allows for high-throughput screening of tens of thousands of cells per second containing target protein variants. Significantly, the protocol prioritizes the soluble fraction of produced proteins leading to products more suitable for downstream applications compared to methods relying solely on rational design.

Our findings highlight the potential of N-terminal optimization for improving protein expression.

The FEBS Journal (2024) © 2024 The Author(s). *The FEBS Journal* published by John Wiley & Sons Ltd on behalf of Federation of European Biochemical Societies.

17424658, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1111/febs.17376 by Cochrane Czech Republic, Wiley Online Library on [02/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Color

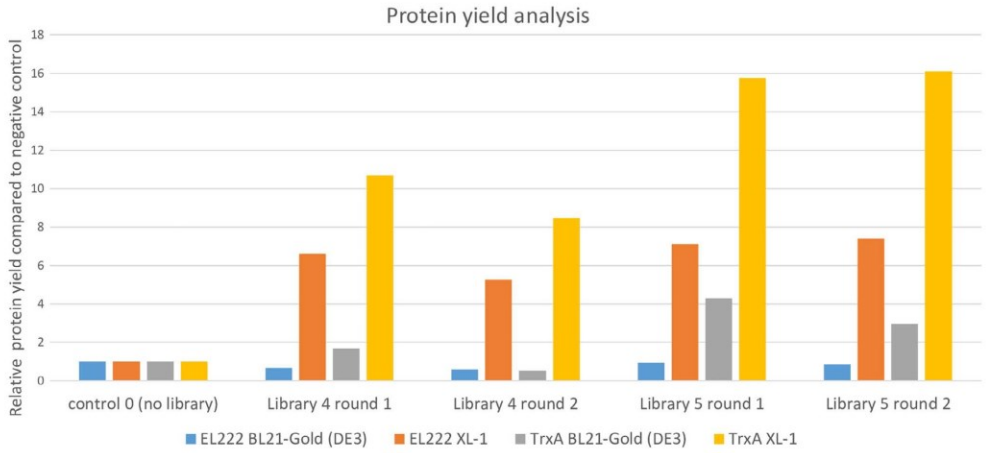


Fig. 11. Chart comparing relative yields of produced proteins (TrxA and EL222) in both BL21-Gold (DE3) and XL-1 *E. coli* cells. The yield was calculated from the soluble fraction on the SDS/PAGE analysis (Fig. 10). Both versions of the library are shown, library 4 referring to a library version containing 12 nucleotides and library 5 referring to a 15 nucleotides library. The set of expressions was performed once prioritizing the exploration of diverse constructs, cell types, and library versions.

Table 1. Benchmark sequences: composition and sources.

Name	Residues	Sequence	Source of the sequence
S01	MGSS	ATGGGCAGCAGC	pET28 vector
S02	MRGS	ATGAGAGGATCG	pQE30 vector
S03	MSKIK	ATGTCTAAAATAAAA	[9,18]
S04	MSKIK	ATGTCTAAAATTAAA	Inspired by [9] but changed according to [12]
S05	MGSDKI	ATGGGTTCAGACAAAATT	ThioHis peptide (M – G – TRX protein)
S06	MSDKI	ATGAGCGATAAAATT	Trx protein, pETM20 based on Addgene plasmid # 176223; http://n2t.net/addgene:176223; RRID: Addgene_176 223; UniProt: P0AA25; *E. coli*
S07	MKIEE	ATGAAAATCGAAGAA	MBP protein, pETM41 based on Addgene plasmid # 38334; http://n2t.net/addgene:38334; RRID: Addgene_38334
S08	MSKEK	ATGTCTAAAGAAAAA	EF-Tu protein; UniProt: P0CE47; AAA50993.1
S09	MKKIA	ATGAAAAAGATTGCA	From rpiB protein (Ribose-5-phosphate isomerase B); UniProt: P37351; *E. coli*; CAA57688.1
S10	MVKIH	ATGGTGAAGATTCAT	[12]

However, finding a universally optimized N-terminal sequence is unlikely and each protein requires its own optimization process, which can be time-consuming and may not always be successful. The method is not suitable for all proteins, particularly those inherently toxic to *E. coli*, those with structures incompatible with bacterial expression. Indeed, we observed limited yield increase for proteins erythroferrone, interleukin 9, and certain thioredoxin A constructs.

In conclusion, this protocol offers a valuable tool for increasing the production of proteins with low yields. This optimization process can also help to get a better ratio of soluble versus insoluble fraction of a protein of interest.

Materials and methods

Cloning

We used several methods of plasmid cloning. The first one was a homologous recombination-based commercial In-Fusion cloning technique (TaKaRa Bio, San Jose,

17424658, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1111/febs.17376 by Cochrane Czech Republic, Wiley Online Library on [02/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

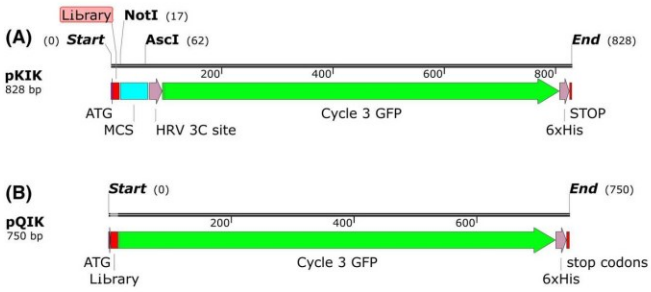


Fig. 12. Scheme of the expression cassettes. (A) pKIK vector set. Protein of interest can be cloned using *NotI* and *AscI* sites into the multiple cloning site (MCS) (B) pQIK vector set. Proteins of interest can be cloned using In-Fusion Restriction-Free cloning techniques between the N-terminal library and the GFP gene.

Table 2. pKIK and pQIK vector set main differences.

Vector pKIK	Vector pQIK
Based on pET22	Based on pQE30
T7 promoter (strong)	T5 promoter (weak)
Restricted *E. coli* strains usage (limited to DE3 cell strains)	Suitable for more *E. coli* strains (TOP10, XL-1, BL21-Gold (DE3)...)
Lac repressor LaqI	No Lac repressor (leaky system)
Production-focused expression cassette design	Clean proof of concept-focused expression cassette

CA, USA). The second method we used was a Restriction-Free (RF) cloning (protocol described elsewhere [43]). Genes coding the examined proteins were ordered in a form of lyophilized DNA string (Thermo Fisher, Waltham, MA, USA). Primers were designed according to TaKaRa Bio Infusion reaction protocol [44] and synthesized (Generi Biotech, Hradec Králové, Czech Republic). Primers for the RF cloning and In-Fusion differed only in their length. Primers for infusion were shorter, having 18–25 bp annealing parts and overhangs, while primers for RF cloning were approximately 31–38 bp long in both regions. DNA libraries were placed directly in the primers, so the library complexity would not get lower during the PCR reaction and primers were designed with a preference to have a G or C nucleotide on their ends, while the melting temperature for both forward and reverse primers was matched. Typical set of primers used in this case for RF cloning the GST gene into the pQIK vector looked like this:

GST-Q-4-FW

5′ *CACAGAATTCATTAAAGAGGAGAAATTAACTATG* NNNNNNNNNNNN**ATGTCCCCTATACTAGGTTA TTGGAAAATTAAGGGC** 3′

GST-Q-REV

5′ **GTGAAAAGTTCTTCTCCTTTGCTAGC***TGGAGGA TGGTCGCCACCACC* 3′

Primers for In-Fusion cloning of the EL222 gene into the pQIK vector looked like this:

EL2-Q-4-FW

5′ *CACAGAATTCATTAAAGAGGAGAAATTAACTATG* NNNNNNNNNNNN**GGGGCAGACGACACACG** 3′

EL2-Q-REV

5′ **GTGAAAAGTTCTTCTCCTTTGCTAGC***GATTCCGG CTTCGACGGC* 3′

where the library (in this case 12 random nucleotides coding four random amino acids) is labeled by letter "N", the annealing part of the primer for the insert is shown in bold and the italic part shows overhang for the infusion cloning, which is complementary to the pQIK vector.

Running an overlap extension PCR with the GFP gene and inserting the whole construct (instead of just inserting the protein of interest) usually delivered higher efficiency, so this extra step was included to obtain samples with higher library complexity.

To replace, delete or add a library or a benchmark sequence in an existing sample, we utilized a KLD cloning technique (New England Biolabs, Ipswich, MA, USA), following the manufacturer's protocol [45]. Typical set of primers (in this case for replacing the benchmark sequences) looked like this:

QIK_T-c1_KLD_FW

5′ *GGCAGCAGC***ATGAGCGATAAAATTATTCACCTG AC** 3′

SH-pQ_KLD_REV

5′ **AGTTAATTTCTCCTCTTTAATGAATTCTGTG** 3′

where the annealing part of the primer is shown in **bold** and the *italic* part shows the new sequence replacement/insertion (in this case benchmark sequence S01 without the start codon, which is already present in the vector).

Traditional cloning based on restriction enzymes cleavage and T4 DNA ligase (New England Biolabs, Ipswich, MA, USA) was used just for a minor part of the samples without DNA library.

The FEBS Journal (2024) © 2024 The Author(s). *The FEBS Journal* published by John Wiley & Sons Ltd on behalf of Federation of European Biochemical Societies.

17424658, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1111/febs.17376 by Cochrane Czech Republic, Wiley Online Library on [02/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Name	TrxA pQIK, XL-1	TrxA, pQIK, BL21-Gold (DE3)	TrxA, pKIK, BL21-Gold (DE3)	GST, pKIK, BL21-Gold (DE3)	only GFP, pKIK, BL21-Gold (DE3)	ERFE, pKIK, BL21-Gold (DE3)
S01	1,165	790	1,029	832	2,838	5
S02	626	902	1,023	698	2,540	2
S03	439	1,040	736	462	2,565	4
S04	1,339	473	733	490	2,440	5
S05	1,243	339	1,070	668	2,845	4
S06	1,499	823	900	495	2,577	5
S07	1,585	663	741	352	2,614	4
S08	1,856	428	1,040	544	2,890	4
S09	1,437	816	625	297	2,300	5
S10	605	480	328	248	1,241	4

Fig. 13. FITC-A levels of fluorescence measurements of the benchmark samples, their sequences are listed in Table 1. Each value represents the median calculated from measurements of hundred thousand analyzed singlet cells, measured once; the highest values are displayed in dark green, decreasing to white towards lower values for each data set. The last column shows erythroferrone (ERFE) constructs that we were not able to produce. Changing the composition of the N-terminus did not boost the production of this construct.

Primers used for the project are listed in a separate table, see Data S2 (-Primers).

Cloned vectors were transformed into electrocompetent *E. coli* TOP10 cells (Thermo Fisher, Waltham, MA, USA) and plated onto solid LB-agar plates containing carbenicillin (50 $\mu g \cdot mL^{-1}$). We selected carbenicillin over ampicillin for its better stability and to minimize the occurrence of satellite colonies.

For benchmark samples and samples without any library single colonies were picked and plasmid DNA was isolated using E.Z.N.A. Plasmid Mini Kit I (Omega Bio-tek, Norcross, GA, USA). For samples containing libraries, plates were washed using 2.4 mL of PBS to resuspend colonies. 2 mL of the suspension were recovered, and plasmid DNA was isolated using the same kit. Isolated plasmids were sequenced (Eurofins Ebersberg, Germany).

Protein expression

Protein expression

This method of protein production was used only for the samples without DNA library. Single colony was picked from a solid agar LB medium containing carbenicillin (50 $\mu g \cdot mL^{-1}$), transferred into fresh LB medium containing carbenicillin (50 $\mu g \cdot mL^{-1}$) and incubated using Duetz microtiter plate system microplates (24-deepwell) at 220 RPM at 37 °C overnight. Overnight culture was then used to inoculate fresh LB carbenicillin medium in 1:100 ratio. This cell suspension was then incubated under the same conditions as the ON culture. When OD_{600} reached 0.6–0.8, induction was performed by addition of IPTG to a final concentration of 1 mM. The cell suspension was then incubated either at 37 °C for 3 h or at 18 °C overnight. The production suspensions were harvested using centrifugation at 5000 ***g*** for 10 min.

For libraries containing samples, the procedure was the same with these exceptions. Instead of picking a single colony from the plate, the plate was washed using 2.4 mL of production media and this mixture was combined with 8 mL of fresh LB media containing carbenicillin, creating a suspension, which was used directly for protein production without overnight culture. Induction for library containing samples was performed by addition of IPTG to a final concentration 1 mM after 60 min of incubation regardless of the OD of the cell culture.

SDS/PAGE and solubility analysis

SDS/PAGE analysis was done using 4–20% Mini-PROTEAN® TGX™ Precast Gels (BioRad, Hercules, CA, USA). After running the electrophoresis, gels were dyed using the Coomassie Brilliant Blue R-250 Staining Solutions Kit (BioRad) and captured using the Transwhite function of the Azure 300 (Azure Bio Systems, Montréal, Canada). ImageJ software was used to analyze the protein yields using the analyze gel function [42]. To show only the soluble fraction of proteins from the harvested cells, we used a B-PER Complete Bacterial Protein Extraction Reagent (Thermo Fisher). Harvested cells were resuspended in this solution and incubated at RT for 30 min. Then the mixture was centrifuged at 20 000 ***g*** and 4 °C for 20 min to separate the soluble fraction from the bacterial debris. Supernatant (soluble fraction) was taken to a fresh tube for further analysis.

Flow cytometry analysis

Flow cytometry experiments were performed by staff of the Imaging Methods Core Facility at BIOCEV using FACSAria™ Fusion SORP (BD Biosciences, New Jersey, USA). Briefly, 0.5 mL of harvested cells were washed and resuspended in 1.4 mL of PBS buffer. This suspension was used for the flow cytometry and diluted so the number of cells passing through the flow cell did not exceed 20 000 $cells \cdot s^{-1}$ during the cell sorting. We utilized 1.0 ND filter and 70 µm nozzle for sorting of the *E. coli* cells.

17424658, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1111/febs.17376 by Cochrane Czech Republic, Wiley Online Library on [02/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

During the sorting procedure we sorted 50 000 cells and plated them onto solid agar LB medium containing carbenicillin antibiotics (50 $\mu g \cdot mL^{-1}$). For control samples, we used the FACS for sorting the same amount (50 000) of singlet cells. This way our control samples underwent the same procedure, as our experiment samples and we were able to match the amount of *E. coli* cells plated for each sample.

Gating strategy

Since the size of *E. coli* cell is close to the detection limit of the BD FACSAria™ Fusion Flow Cytometer, which we used for this experiment, we designed a P1 gate, shown in the Fig. 3A, to make sure that the smallest detected particles were living cells not just cell debris. Since we were discarding the majority of the cells running through the sort, we decided to design stricter gating and discard the smallest particles (gate P1) (see Fig. 3A), since we expected this gate to contain several contaminants such as bacterial debris. This was in concordance with our previous finding, which showed that if the gating of live cells was less strict, the P2 gated cells that we wanted to collect from the experiment had shown higher contamination.

Regarding the gating of singlet cells in case of *E. coli* cells, gating the singlet cells is usually not accurate, since bacteria are close to the detection limit of BD FACSAria™ Flow Cytometer. Similarly, as discussed above, we cut out the most extreme hits to minimize the contamination in the final cell suspension.

The gates designed to sort cells emitting the highest fluorescence (gates P2, P3 etc.) follow the limitations regarding sample cleavage discussed in the results section and shown in Fig. 8, which revealed, that if the population emitting the highest fluorescence was not uniform, the GFP may have been cleaved from the desired construct and there should be additional gate designed aiming for the top of population with the second highest fluorescence.

Acknowledgements

The authors acknowledge Imaging Methods Core Facility at BIOCEV, namely Mgr. Ondřej Honc and Michaela Kolařík Zázvorková for their support & assistance with the flow cytometry experiments. This research was supported by project nr. LX22NPO5102 (MEYS): Financed by European Union – Next Generation EU and the institutional grant to the Institute of Biotechnology of the Czech Academy of Sciences RVO 86652036.

Conflict of interest

The authors declare no conflicts of interest associated with this work.

Author contributions

SH, PM and BS conceived the project; SH, YP and PM planned experiments; SH, JS, LS and MH performed experiments and analyzed data; SH, BS, LS and YP wrote the paper; JS created the graphics for the publication.

Peer review

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/febs.17376.

Data availability statement

The data that support the findings of this study are openly available in the Zenodo server at https://zenodo.org/records/12772716.

References

1 Rosano GL & Ceccarelli EA (2014) Recombinant protein expression in *Escherichia coli*: advances and challenges. *Front Microbiol* **5**, 172.
2 Overton TW (2014) Recombinant protein production in bacterial hosts. *Drug Discov Today* **19**, 590–601.
3 Gopal GJ & Kumar A (2013) Strategies for the production of recombinant protein in *Escherichia coli*. *Protein J* **32**, 419–425.
4 Jia B & Jeon CO (2016) High-throughput recombinant protein expression in *Escherichia coli*: current status and future perspectives. *Open Biol* **6**, 160196.
5 Makrides SC (1996) Strategies for achieving high-level expression of genes in *Escherichia coli*. *Microbiol Rev* **60**, 512–538.
6 Bivona L, Zou ZC, Stutzman N & Sun PD (2010) Influence of the second amino acid on recombinant protein expression. *Protein Expr Purif* **74**, 248–256.
7 Goodman DB, Church GM & Kosuri S (2013) Causes and effects of N-terminal codon bias in bacterial genes. *Science* **342**, 475–479.
8 Gutierrez G, Marquez L & Marin A (1996) Preference for guanosine at first codon position in highly expressed *Escherichia coli* genes. A relationship with translational efficiency. *Nucleic Acids Res* **24**, 2525–2527.
9 Ojima-Kato T, Nagai S & Nakano H (2017) N-terminal SKIK peptide tag markedly improves expression of difficult-to-express proteins in *Escherichia coli* and *Saccharomyces cerevisiae*. *J Biosci Bioeng* **123**, 540–546.
10 Stenstrom CM, Jin HN, Major LL, Tate WP & Isaksson LA (2001) Codon bias at the 3′-side of the initiation codon is correlated with translation initiation efficiency in *Escherichia coli*. *Gene* **263**, 273–284.

The FEBS Journal (2024) © 2024 The Author(s). *The FEBS Journal* published by John Wiley & Sons Ltd on behalf of Federation of European Biochemical Societies.

17424658, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1111/febs.17376 by Cochrane Czech Republic, Wiley Online Library on [02/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

11 Thangadurai C, Suthakaran P, Barfal P, Anandaraj B, Pradhan SN, Boneya HK, Ramalingam S & Murugan V (2008) Rare codon priority and its position specificity at the 5′ of the gene modulates heterologous protein expression in *Escherichia coli*. *Biochem Biophys Res Commun* **376**, 647–652.

12 Verma M, Choi J, Cottrell KA, Lavagnino Z, Thomas EN, Pavlovic-Djuranovic S, Szczesny P, Piston DW, Zaher HS, Puglisi JD *et al.* (2019) A short translational ramp determines the efficiency of protein synthesis. *Nature Comm* **10**, 5774.

13 Stenstrom CM & Isaksson LA (2002) Influences on translation initiation and early elongation by the messenger RNA region flanking the initiation codon at the 3′ side. *Gene* **288**, 1–8.

14 Allert M, Cox JC & Hellinga HW (2010) Multifactorial determinants of protein expression in prokaryotic open Reading frames. *J Mol Biol* **402**, 905–918.

15 Yu ZB & Jin JP (2007) Removing the regulatory N-terminal domain of cardiac troponin I diminishes incompatibility during bacterial expression. *Arch Biochem Biophys* **461**, 138–145.

16 Charneski CA & Hurst LD (2013) Positively charged residues are the major determinants of ribosomal velocity. *PLoS Biol* **11**, e1001508.

17 Looman AC, Bodlaender J, Comstock LJ, Eaton D, Jhurani P, Deboer HA & Vanknippenberg PH (1987) Influence of the codon following the Aug initiation codon on the expression of a modified Lacz gene in *Escherichia coli*. *EMBO J* **6**, 2489–2492.

18 Ojima-Kato T, Nishikawa Y, Furukawa Y, Kojima T & Nakano H (2023) Nascent MSKIK peptide cancels ribosomal stalling by arrest peptides in *Escherichia coli*. *J Biol Chem* **299**, 104676.

19 Kozak M (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene* **361**, 13–37.

20 Tuller T, Waldman YY, Kupiec M & Ruppin E (2010) Translation efficiency is determined by both codon bias and folding energy. *Proc Natl Acad Sci USA* **107**, 3645–3650.

21 Del Campo C, Bartholomaus A, Fedyunin I & Ignatova Z (2015) Secondary structure across the bacterial transcriptome reveals versatile roles in mRNA regulation and function. *PLoS Genet* **11**, e1005613.

22 Keller TE, Mis SD, Jia KE & Wilke CO (2012) Reduced mRNA secondary-structure stability near the start codon indicates functional genes in prokaryotes. *Genome Biol Evol* **4**, 80–88.

23 Katz L & Burge CB (2003) Widespread selection for local RNA secondary structure in coding regions of bacterial genes. *Genome Res* **13**, 2042–2051.

24 Kudla G, Murray AW, Tollervey D & Plotkin JB (2009) Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* **324**, 255–258.

25 de Smit MH & van Duin J (1990) Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis. *Proc Natl Acad Sci USA* **87**, 7668–7672.

26 Bhandari BK, Lim CS & Gardner PP (2021) TISIGNER.Com: web services for improving recombinant protein production. *Nucleic Acids Res* **49** (W1), W654–W661.

27 Gu W, Zhou T & Wilke CO (2010) A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Comput Biol* **6**, e1000664.

28 Zhou T & Wilke CO (2011) Reduced stability of mRNA secondary structure near the translation-initiation site in dsDNA viruses. *BMC Evol Biol* **11**, 59.

29 dos Reis M, Savva R & Wernisch L (2004) Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res* **32**, 5036–5044.

30 Pop C, Rouskin S, Ingolia NT, Han L, Phizicky EM, Weissman JS & Koller D (2014) Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. *Mol Syst Biol* **10**, 770.

31 Chen GFT & Inouye M (1990) Suppression of the negative effect of minor arginine codons on gene-expression - preferential usage of minor codons within the 1st 25 codons of the Escherichia-Coli genes. *Nucleic Acids Res* **18**, 1465–1473.

32 Ikemura T (1985) Codon usage and transfer-Rna content in unicellular and multicellular organisms. *Mol Biol Evol* **2**, 13–34.

33 Ikemura T (1981) Correlation between the abundance of *Escherichia coli* transfer-Rnas and the occurrence of the respective codons in its protein genes – a proposal for a synonymous codon choice that is optimal for the *Escherichia coli* translational system. *J Mol Biol* **151**, 389–409.

34 Zamora-Romo E, Cruz-Vera LR, Vivanco-Dominguez S, Magos-Castro MA & Guarneros G (2007) Efficient expression of gene variants that harbour AGA codons next to the initiation codon. *Nucleic Acids Res* **35**, 5966–5974.

35 de Valdivia EIG & Isaksson LA (2004) A codon window in mRNA downstream of the initiation codon where NGG codons give strongly reduced gene expression in *Escherichia coli*. *Nucleic Acids Res* **32**, 5198–5205.

36 Puri N, Appa Rao KB, Menon S, Panda AK, Tiwari G, Garg LC & Totey SM (1999) Effect of the codon following the ATG start site on the expression of ovine growth hormone in *Escherichia coli*. *Protein Expr Purif* **17**, 215–223.

37 Cambray G, Guimaraes JC & Arkin AP (2018) Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in *Escherichia coli*. *Nat Biotechnol* **36**, 1005-+.

17424658, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1111/febs.17376 by Cochrane Czech Republic, Wiley Online Library on [02/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

38 Pham PN, Huliciak M, Biedermannova L, Cerny J, Charnavets T, Fuertes G, Herynek S, Kolarova L, Kolenko P, Pavlicek J *et al.* (2021) Protein binder (ProBi) as a new class of structurally robust non-antibody protein scaffold for directed evolution. *Viruses* **13**, 190.
39 Huliciak M, Biedermanová L, Berdár D, Herynek S, Kolárová L, Tomala J, Mikulecky P & Schneider B (2023) Combined in vitro and cell-based selection display method producing specific binders against IL-9 receptor in high yields. *FEBS J* **290**, 2993–3005.
40 Crameri A, Whitehorn EA, Tate E & Stemmer WPC (1996) Improved green fluorescence protein by molecular evolution using DNA shuffling. *Nat Biotechnol* **14**, 5–319.
41 Fukuda H, Arai M & Kuwajima K (2000) Folding of green fluorescent protein and the cycle3 mutant. *Biochemistry* **39**, 12025–12032.
42 Rasband WS (1997-2018) ImageJ. U.S. National Institutes of Health, Bethesda, MA, USA. https://imagej.nih.gov/ij/
43 Unger T, Jacobovitch Y, Dantes A, Bernheim R & Peleg Y (2010) Applications of the restriction free (RF) cloning procedure for molecular manipulations and protein expression. *J Struct Biol* **172**, 34–44.
44 Takara Bio Inc. USA (2024) In-Fusion® Snap Assembly User Manual. https://www.takarabio.com/documents/User%20Manual/In/In-Fusion%20Snap%20Assembly%20User%20Manual.pdf
45 New England Biolabs (2024) KLD Enzyme Mix Reaction Protocol (M0554). https://www.neb.com/en/protocols/2017/03/31/kld-enzyme-mix-reaction-protocol-m0554

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.
Data S1. Sequences of plasmids.
Data S2. Primers used.

The FEBS Journal (2024) © 2024 The Author(s). *The FEBS Journal* published by John Wiley & Sons Ltd on behalf of Federation of European Biochemical Societies.

17424658, 0, Downloaded from https://febs.onlinelibrary.wiley.com/doi/10.1111/febs.17376 by Cochrane Czech Republic, Wiley Online Library on [02/01/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License